This is to certify that the

dissertation entitled

The Relationship Between Item Format and Cognitive Processes in Wide-scale Assessment of Mathematics

presented by

Diane R. Garavaglia

has been accepted towards fulfillment
of the requirements for

Ph.D. degree in Counseling, Educational Psychology, and Special Education

Irwin J. Lehmann
Major professor

Date 5-10-01

DATE DUE	DATE DUE	DATE DUE

THE RELATIONSHIP BETWEEN ITEM FORMAT AND COGNITIVE PROCESSES IN WIDE-SCALE ASSESSMENT OF MATHEMATICS

By

Diane R. Garavaglia

A DISSERTATION

Submitted to

Michigan State University

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Department of Counseling, Educational Psychology, and Special Education

2001

ABSTRACT

THE RELATIONSHIP BETWEEN ITEM FORMAT AND COGNITIVE PROCESSES IN WIDE-SCALE ASSESSMENT OF MATHEMATICS

By

Diane R. Garavaglia

The purpose of this study was to determine whether test takers used different cognitive processes when they solved multiple-choice versus constructed-response items. I conducted this study in an era when school accountability and high-stakes, large-scale assessments were seemingly as important as student learning itself. The high-stakes nature of testing created an environment in which both testing advocates and challengers scrutinized tests even more than normal. In particular, some challengers asserted that multiple-choice items were ill-suited for assessing certain types of cognitive processes or for providing useful information about student achievement.

Using information as the central idea, I approached the question from an information-value (or value-added) perspective and attempted to determine whether there was a difference in the type of cognitive processes elicited by each item format. The question was narrowly contextualized in one area of mathematics, namely 8th grade, algebraic pattern items. I selected 34 students who were enrolled in 8th grade mathematics courses in the spring of 1998. I examined the question using a think-aloud procedure — an analysis tool seldom used in the field of measurement. The overall

results suggested that students used similar cognitive processes to solve both multiple-choice and constructed-response pattern items. However, the results were likely related to the characteristics of the items — that is, many constructed-response items allowed for one solution path. I referred to these items as "multiple-choice items in disguise." Recommendations were offered to test users, developers, and other researchers.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

Introduction

Testing has always been scrutinized, and it has come under even more scrutiny in recent years by everyone involved in it — test takers, teachers, parents, school administrators, community members, legislators, and measurement professionals. With the high stakes and large costs often associated with tests, it is no wonder they are scrutinized. Many critics of testing argue that there is too much testing in our schools, that the time spent on testing is time taken away from instruction and learning. Others argue that tests emotionally harm some students by putting too much stress on them, especially when younger students are the test takers. Still others say that tests do not tap important cognitive processes, such as higher-order thinking skills. And, within the last ten years, both testing advocates and challengers have asserted that particular item formats are ill-suited for assessing certain types of cognitive processes or for providing useful information about student achievement. Of all the critiques, this last one may be the most important; in my opinion, the most important reason for testing is to provide information to test users.

Information has "value" in that it helps users make decisions or draw conclusions about questions that matter to them (Pearson and Garavaglia, 1997). In the field of measurement and within the arena of large-scale testing, a few questions that matter stem from our interest in the interplay between assessment and curriculum. Test scores are just one type of useful information to answer questions such as,

- Are my students reading at grade level?
- How well are my students performing in reading comprehension?
- What do the students know about mathematical problem-solving?

Mehrens and Lehmann (1987) implicated both the qualitative and quantitative aspects of information-value in their statement that it is necessary to have as much of the relevant information as possible to make an informed decision: "The more, and more accurate, the information on which a decision is based, the better that decision is likely to be" (p.10). The question of interest from an information-value perspective is whether an additional datum of information would help test users better answer the question(s) of interest. The question of interest specific to this study, in a broad sense, was to explore the information-value of constructed-response items when they are mixed with multiple-choice items on large-scale assessments. More narrowly, the purpose of this study was to determine whether constructed-response and multiple-choice items require students to use similar cognitive processes.

This question is relevant given the developments witnessed over the last dozen years or so in test development, which is the creation of a test that measures a single content area by using multiple item formats on the test. Many state level tests and the National Assessment of Educational Progress (NAEP) use a combination of multiple-choice and constructed-response items to assess a content area. But, strong advocates of the constructed-response item format say that certain cognitive processes cannot be tapped by multiple-choice items: namely, higher-order thinking skills (HOTS). Thus, when using only multiple-choice items on a test, higher levels of knowledge are not being tested (Snow, 1993). Conversely, there are others who question how it can be said with

any certainty that multiple-choice items are not tapping HOTS or that they are not providing meaningful information about the content being assessed (Haladyna, 1994; Stiggins, 1994; Martinez, 1993; Mehrens, 1992).

Rather than discounting the multiple-choice item format in favor of another format, Snow (1993) offers a balanced perspective and suggests that further research be conducted to add meaningful information about the relationship between cognitive processing and item formats. Haladyna (1994) writes,

> We must learn quite a bit more about the effects of item format on cognitive learning before we can make confident statements about the effectiveness of any format. Research is needed that shows the optimal formats for measuring newly defined abilities and various forms of higher level achievement. (p.183)

The general goal of conducting this study echoes Haladyna's statement, that is, to add meaningful and practical empirical evidence to the literature on the interplay between cognitive processes and item format. The question is interesting both theoretically and practically. Theoretically, the question of ensuring measures that tap higher-order thinking is an essential feature of the validity of such a test. Practically, the question is one of cost-effectiveness and curricular information. Some people want to measure higher-order thinking, but how can we find a format that is simultaneously effective, informative, and precise and places the least burden on schools, teachers, and students in terms of both money and time?

I approached the issue from a value-added perspective and attempted to determine whether there was a difference in the type of cognitive processes elicited by each type of

item format. The question was contextualized in mathematics. I examined the question by using an analysis tool seldom used in the field of measurement, namely, a think-aloud procedure that provided verbal evidence about the cognitive processes utilized by students as they answered items. The items used were released and non-released 8th grade algebra items from the 1992 and 1996 National Assessment of Educational Progress (NAEP) and one 8th grade algebra item from the Balanced Assessment project (Balanced Assessment Package, 1997).

CHAPTER II

Review of Related Literature

The relevant literature to answer the research question can be classified in two ways: (a) do constructed-response items provide us with more information (information-value) about what students are capable of doing than we get from multiple-choice items alone, and (b) are the cognitive processes needed to answer constructed-response items unique to the constructed-response item format? Much of the relevant literature comes from studies that analyzed data from several of the College Boards Advanced Placement (AP) tests. Perhaps this is because the AP tests have been using both multiple-choice and constructed-response item formats for several years. The literature review also included studies that used achievement instruments other than the AP tests.

Besides the two main classifications above, three additional topics were reviewed to learn what information already existed about the issues relevant to this study. The topics included reviews of the think-aloud methodology, comparability of items written in multiple formats, and mathematical cognitive processes. Each topic is presented separately in this chapter. I end the chapter by identifying how this study will contribute to the literature.

Information-value Studies

Several of the AP studies looked at the amount of new-information gained when mixed item formats were used on a single test. Although the authors did not provide operational definitions for information-value (or value-added, they are seemingly used interchangeably), I interpreted it as how much additional or new information about a construct is gained when constructed-response items are added to multiple-choice items

on a test. (See Pearson and Garavaglia, 1997, for a description of how information-value can be conceptualized in large-scale assessment programs.)

Lukhele, Thissen, and Wainer (1994) used item response theory (IRT) models to examine the amount of information obtained from different item formats when presented on the same test. They analyzed the multiple-choice data using the 3-parameter IRT model and analyzed the constructed-response items using a graded response model. Using data from the 1989 AP Chemistry and 1988 AP US History tests, they found for both tests that adding constructed-response items on the tests provided little information beyond what the multiple-choice items yielded. The authors also examined the amount of time test takers used to respond to multiple-choice items in comparison to constructed-response items and the cost to score the two item types. They found that test takers could answer 16 multiple-choice items in the same amount of time that was needed to answer one constructed-response item, and that the 16 multiple-choice items cost much less to score compared to the one constructed-response item. Most important, the information yield of the 16 multiple-choice items (on the chemistry exam) was double that of the one constructed-response item. They also showed that multiple-choice items were more cost effective compared to constructed-response items. In conclusion, they found that constructed-response items yielded less information, required more testing time, and incurred larger costs compared to multiple-choice items.

Information value studies have also been conducted in the areas of science, chemistry, and computer science (Thissen, Wainer, & Wang, 1994; Wainer & Thissen, 1993; Wainer, Wang, & Thissen, 1991; Wang, Wainer, & Thissen, 1993). The findings from all of the studies suggested that when response data from constructed-response

items were combined with response data from multiple-choice items, little new information was obtained about any of the areas.

By contrast, research findings from content areas other than those discussed thus far suggested that differences (e.g., traits) existed across different item formats. For instance, in the area of writing, Werts et al. (1980) worked with first-year college students and attempted to determine whether different item formats would detect different writing traits. The design was a variation of a multitrait-multimethod design. Three administrations of the Test of Standard Written English (TSWE) and three short (20 minute) essay prompts were used to collect response data. The three essay prompts were considered to be three separate and therefore independent tests. (The authors did not provide information about the essay prompts.) All of the tests were given within the same year and over several test occasions. The nonzero covariation for the essay residuals showed that the essays measured a common trait that was different from whatever traits the essays and TSWE shared. So, even though all of the assessments were measuring writing, the essays seemingly measured something unique.

Bennett et al. (1990) conducted two studies using the same test, AP Computer Science Examination, but different measurement models to examine differences between item formats. They used a confirmatory factor analysis model in the 1990 study and a model hypothesizing separate format factors in the 1991 study. In the 1990 study they first treated each constructed-response item as a separate variable. They then grouped ten or more of the multiple-choice items; each group of multiple-choice items represented a separate variable. A one-factor covariance structure model was used to analyze the data. The results indicated that both item formats measured the same characteristics. They

concluded that adding constructed-response items to multiple-choice items did not add additional information about computer science. However, in 1991 Bennett et al. discovered that the disattenuated correlation coefficients from the model hypothesizing separate format factors were significantly different from unity. In the 1991 study, the researchers found differences between the item formats, but, like the 1990 study, a limitation of the finding was the lack of information about the source of the differences.

Using factor analysis Thissen et al. (1994) found evidence that the constructed-response items on AP Computer Science and Chemistry tests measured something unique from the multiple-choice items because the factors were significantly different for the constructed-response items compared to the general factor. The evidence also indicated that the constructed-response and multiple-choice questions both measured the same thing because the loadings for the constructed-response items were larger on the multiple-choice factor than they were on the constructed-response factor. Further, although the constructed-response items measured something different from the multiple-choice items, they did not measure that different thing very well. The researchers based this conclusion on the observation that the factor loadings for the constructed-response items were small on the factors specified for the constructed-response items.

In short, the findings about item format differences and, in particular, the value added of constructed-response formats when combined with multiple-choice items on a test, appear to be mixed. In some content areas, constructed-response items seem to add little or no information; while in others, they seem to add unique information to the process of making decisions on the basis of test scores.

Think-aloud Methodology

The think-aloud methodology is an interview between the researcher and the respondent (i.e., student). Generally, the researcher and student sit together at a table as the student performs the specified task. As the student performs the task, he or she talks aloud and tells the researcher what he or she is thinking. The researcher prompts the student for clarity or elaboration when necessary. The setting is informal and collegial.

There are two general approaches when conducting a think-aloud. Ericsson and Simon (1993) categorized them into two families, concurrent and retrospective interviews. I chose a concurrent approach for two reasons. First, the accuracy, and therefore utility, of retrospective verbal reports has been questioned by some (Mueller, 1911; Nisbett and Wilson, 1977). Mueller (1911) noted that subjects sometimes confused other retrievable information with information related to the processes used to solve the tasks. Hamilton et al. (1997) reported similar findings in a more recent study. Hamilton et al. found that the time lapse between responding to the item and participating in the interview could result in forgetting, interference, and other memory lapses that compromise the accuracy of the verbal reports. The findings from the studies provided convincing evidence that the use of retrospective verbal reports would likely introduce measurement error in the data.

Second, the concurrent approach also allowed me to observe the moment by moment sequential thinking of the student as he or she responded to the items, without altering the cognitive processes they used to solve the items (Ericsson and Simon, 1993). Because the purpose of this study was to determine whether similar cognitive process were needed to solve multiple-choice items and constructed-response items, the sequence

of the cognitive processes had to be maintained and not interrupted during data collection. Therefore, because the students' cognitive processes were of primary interest, the concurrent think-aloud procedure was well suited for the purpose of this study.

Ericsson and Simon (1993) examined the myriad of ways to conduct a think-aloud interview and the appropriate method to use for a particular purpose. Rather than reporting the myriad approaches here, I instead reviewed studies that used think-aloud procedures and principles that mirrored those planned for this study regardless of the content area.

Montague and Applegate (1993) used a think-aloud to compare the problem-solving behaviors used by learning disabled, average, and gifted groups of middle school students. They were particularly interested in learning whether the group of students identified as learning disabled used different cognitive processes when solving word problems compared to the other two groups of students. To test their hypothesis, they asked the three groups of students to think-aloud as they answered one-step, two-step, and three-step word problems.

The researchers identified the cognitive process categories *a priori*, based on an information processing theoretical framework, and then used the students' think-aloud data to count the number of verbalizations students made within a cognitive process category. By counting the number of times students, within each of the three groups, used a particular category, Montague and Applegate (1993) found that students identified as learning disabled used different approaches to problem-solving than the other two student groups. The researchers confirmed their hypothesis that students with disabilities approached problem solving in less effective ways than students without a disability.

Hamilton et al. (1997) used the think-aloud procedure to examine how useful the verbal data would be for supporting the findings from a statistical analysis (full-information item factor analysis). Specifically, the researchers wanted to learn whether combining quantitative and qualitative data would be an effective way to examine the validity of science items written in several different formats. They used a concurrent think-aloud procedure with high school students.

To examine their research question, they first analyzed the students' item responses with a factor analytical procedure. Three science dimensions emerged from the results of the factor analysis. Often this is where factor analysis studies end. But, these researchers, using the factor analysis results, then selected 16 multiple-choice items and three constructed-response items to represent the science knowledge assessed by the three dimensions. They tried to select items that varied in difficulty. To compare cognitive processes associated with the different item formats, they matched two items based on the content assessed by the items. They then used the interview data from these items to clarify the meaning of the three science dimensions. After the study, they concluded that "the most important benefit [of think-alouds] is in identifying knowledge and skills that test items require or permit but that are ignored in test interpretation" (p.196).

Research in other content areas used the think-aloud procedure as the main tool for collecting data, rather than combining it with a statistical procedure as described in the previous study. Reading comprehension was the content area most frequently studied. For example, a reading comprehension study by Farr et al. (1990) provided insight about the kind of information obtained from think-alouds. The researchers

examined only multiple-choice items to learn whether the items assessed the reading comprehension processes intended by the test developer. To make this determination, the researchers had 26 college students take a standardized reading comprehension test. They asked them to think-aloud as they read the passages that involved a set of context-dependent items. The most common of four strategies identified from the verbal reports showed that the students read the passage, then read the items, and then returned to the passage to find the correct answer, rather than reading for in-depth understanding the first time they read the text.

In addition to the type of reading comprehension utilized by test takers, Far et al. (1990) also concluded that the development of items determines the type(s) of cognitive processes that can be used by the respondent. Hamilton et al. (1997) support their conclusion. While this conclusion seems plausible, it begs the fundamental item development question: How do you develop items that encourage the type of thinking you intend to measure? To take this idea one step further, after the test items are written, it seems reasonable to assume that think-alouds will have to be conducted in parallel with field testing to validate the cognitive process intended by the item writers. Whether or not achievement test developers will use such a (costly) validation method is unknown. And, although Farr et al. (1990) did not draw this conclusion from their work, the quality of the item would seem to play an important part when determining whether items elicit the intended cognitive process(es). Hence, item quality became a major focus for this study, as is seen later.

Haladyna (1994) suggested that think-alouds could be used as an item review procedure employed during field testing. In his assessment of the procedure, he compared the think-aloud method to field testing items by saying,

> In formal testing programs, the *think-aloud* procedure has been used to study the thought processes of students during a test. The *developmental field test* is also designed to accomplish a similar end, to analyze student behavior during a test to determine if an item is working as intended. The procedures for the think-aloud and the developmental field test are essentially the same (p.138).

I agree that field testing and think-alouds are two ways to analyze student behavior on a test. However, I think the *kind* of information obtained from the two approaches is different. The think-aloud method provides qualitative information about the cognitive maneuvers made by the test-taker. The verbal data illuminate the students' behaviors *as* they respond to the items. Furthermore, the think-aloud method provides the researcher with the opportunity to probe the student to further explain his or her thoughts. By probing, the researcher can ask the student to clarify or elaborate his or her responses. And, when the researcher uses a one-to-one interview, he or she has the benefit of seeing (in real-time) how the student moves through a test booklet, interrelates information on the test, and how the student retrieves particular information from an item stem (or reading passage) to answer the item.

In contrast, field testing does not allow for probing so the researcher is limited by the information obtained from students' written or bubbled responses. It also does not provide an environment for closely observing students' test taking behaviors. However,

field testing allows for the collection of several data points on every item because of the minimal amount of direct interaction needed between the test taker and the test administrator. Thus, larger numbers of items and test data are collected in an efficient way from field testing than from think-alouds.

Providing the comparisons between think-alouds and field testing is not an argument against conducting field tests. However, the different kinds of information obtained from a think-aloud and from a field test could be used in complementary ways to validate items on achievement tests. The study by Hamilton et al. (1997) provided empirical evidence to support this assertion.

A final thought about the think-aloud method comes from Norris (1990) who said,

> Verbal reports of thinking would be useful in the validation of multiple-choice critical thinking tests, if they could provide evidence to judge whether good thinking was associated with choosing keyed answers and poor thinking was associated with unkeyed answers (p. 55).

Norris' comment about the accessibility of poor thinking associated with unkeyed answers is an important idea when the users of the information are teachers. Teachers often want to know why a student answered an item incorrectly. The following questions often come to mind,

- Did the student simply misread the test item?
- Was the item defective?
- Did the student have a misconception about the particular concept that prohibited him or her from responding correctly?

The teacher has a difficult time answering any of the questions without specific information obtained directly from the student. The think-aloud procedure meets this need.

Comparability of Items Written in Multiple Item Formats

Research that examines the relationship between cognitive processes and item formats is vulnerable to how items are selected for inclusion in the study. One cannot assume that an item written in the constructed-response format better assesses cognitive processes than an item written in the multiple-choice format or vice versa. Chaucey and Dobbin (1963) said, "multiple-choice questions can be written so as to require substantial thought." Hamilton et al. (1997) stated that there are some multiple-choice items that assess more than factual knowledge; typically, this happens when the items require students to generate answers that have not been previously memorized.

Hamilton et al. also say that performance items may assess factual or simplistic knowledge when written to assess those kinds of knowledge. For example, some constructed-response items require examinees to provide a short list of facts that are easily recalled directly from instruction. A question that requires students to list the steps in the water cycle is an example of this type of low-level item. With adequate item writing training, experience, and skill, multiple-choice and constructed-response items can be written at levels above recall. Haladyna (1994) provides extensive information about the technique of writing items to assess a range of cognitive processes and difficulty levels.

When researchers study the effects of item format on cognitive processes, they typically match pairs of existing items — one multiple-choice and one constructed-

response — using content as a means to match them. Matching two items based on content minimizes error introduced into the equation by only allowing the variable of interest, in this case cognitive processes, to vary. The following studies show how the researchers matched items when examining whether item format and cognitive processes interacted.

Campbell (1995) used NAEP reading items to look for an item format and cognitive process interaction. He used existing items and attempted to create item pairs, one multiple-choice and one constructed-response, so that the two items were as "similar" in content as possible. Three criteria were utilized to select and match the items: (a) NAEP reading stance classification (initial understanding, developing and interpretation, personal reflection, or critical stance), (b) national percent correct, and (c) type of reading text or situation (literary experience, informational, or perform a task).

Even by Campbell's admission, there was no guarantee, even after matching the items as best he could, that the content and comprehension aspects were similar between the paired items. Also, items that appear similar in content might vary in terms of quality. For example, in a set of matched items the multiple-choice item may have been a better item in terms of the depth of knowledge needed by the respondent to select the best answer, whereas the constructed-response version may have invited a vague or surface level response. A lesson from Campbell's (1995) work was that matching existing items did not necessarily guarantee that the matched-items were assessing similar content and/or cognitive processes. Thus, when examining cognitive processes associated with different item formats the quality and content for the paired items must be comparable.

Martinez (1991) used the stem equivalent approach when comparing item level statistical characteristics of figural items (items that require students to construct a response and use figural information, such as illustrations or graphs, as the response medium) and multiple-choice items in the area of science. Martinez wrote 25 figural science items to match the NAEP science specifications. He then matched the figural items with 25 existing NAEP multiple-choice items. The 50 items were administered on parallel test forms. He did not draw conclusions about the comparability of the stem equivalent items, but he did report item statistics for the matched items, which was the intended purpose of the study.

His finding suggested that the figural items were comparable to or better than their multiple-choice counterparts in terms of item difficulty and discrimination. The finding is useful for showing how different item formats compare in terms of item statistics. The researcher did not examine the cognitive processes associated with the two item formats or the degree of content or cognitive process similarity between the item formats. In fact, Martinez proposed that additional research was needed to determine the extent that item formats draw upon unique abilities.

Martinez (1991) did not report whether he conducted a content item review of the new figural items to ensure that they mapped back to the science framework. It also was unclear if he purposefully used the item specifications to write the figural items or whether he randomly wrote 25 figural items and then determined which of the 25 multiple-choice items most closely matched the figural items. If he did not first select the multiple-choice items, identify the item specifications that mapped to the items, and then develop the figural items using the identified specifications, the content match across the

item formats may have been different from the beginning. Furthermore, the statistical differences he found could have been confounded with substantive differences in the items themselves. The same argument holds for matching cognitive processes between the existing multiple-choice items and the new figural items.

Martinez's (1991) work established the basis for writing items for this study but I included an extra step to the process to address the comparability of content across item formats, as discussed above. To do this, I borrowed ideas from Frederiksen's (1984) research on test bias and Haladyna's extensive work in item development in an attempt to write content comparable items. Details about the item writing process are reported in chapter 3.

Mathematical Cognitive Processes

Demby (1997) used a retrospective interview approach to determine which procedures students used to perform algebraic operations on classroom level tests. A cohort of 108 students were first tested in the 7^{th} grade and re-examined in the 8^{th} grade. The study was conducted in two phases. First, the students were administered an algebra test in their regular classroom. The researcher then analyzed the students' written work, classified the observed errors, and selected 51 students to participate in a follow-up interview.

In the second phase of the study, Demby returned each student's original test. The students were instructed to correct any mistakes they made during their original solution strategies and re-work the items. After the students corrected the mistakes, the researcher interviewed 51 students and asked them to explain how they obtained the answer to each item. Seven common solution strategies emerged as the researcher

analyzed the interview data: automatization, formulas, guessing-substituting, preparatory modification of the expression, concretization, rules, and quasi-rules. Demby also noticed that some of the strategies were used independent of the others; other times combinations of one or more of the seven strategies were noted, e.g., PM +R+C and R+GS. The combinations seemed to represent consecutive steps of an algebraic transformation. The researcher noted that the seven common strategies occurred regardless of a right or wrong answer.

Demby also observed qualitative differences in the types of errors made and solution strategies used from grade 7 to grade 8. Most notably, she observed that 7th graders often used wrong rules in the beginning of the school year and improved their application of rules as the school year progressed. She also observed that students used heuristics to solve the items, rather than formal rules that were taught in class or presented in textbooks. Students used formulas infrequently. Her overall conclusion was that incorrect application of rules seemed to be a normal developmental stage of learning algebra.

Gerace and Mestre (1982) examined the cognitive processes employed by 9th graders enrolled in Algebra I classes, and more specifically the errors students made when solving algebra problems. Data were collected using a think-aloud procedure. The results indicated that students had difficulty differentiating between labels and variables. For example, the students were presented with the following question, Use S and P to represent that there are 6 times as many students as professors at this university. Thirty-five percent of the 14 students wrote 6S=P. The interviewers concluded that the students used S and P as labels rather than treating them as variables. Students made the same

error in three more "label versus variable" items like the one presented above. In fact, the researchers concluded that the first noun the student read in the problem statement triggered the students to treat the variable as a label.

The researchers also concluded that many of the students approached algebra as rule-based rather than concept-based. But, they observed that students often misapplied the use of algebraic rules. This finding was similar to Demby's (1997) finding.

The last study reviewed was conducted in the early 1980's and therefore was solidly grounded in information-processing theory. Leino (1981) investigated the relationship between cognitive processes and mathematical achievement (i.e., course grades), among other things. To examine the types of processes students used, Leino collected think-aloud data on 21 7^{th} grade students when solving mathematics items. He described the mathematics items as a collection of problems or tasks that assessed arithmetic, algebraic, and geometric problems. The items were included in an appendix; all of the items were presented in the constructed-response format.

Of particular interest was the list of cognitive processing and strategies Leino listened for while coding student think-alouds on mathematics items. The processes were grouped into three general categories:

1. Obtaining information
    - Perceiving the given information (facts, figures, etc.)
    - Perceiving geometric information in embedding context
    - Finding out the relations between information given
    - Grasping the formal structure of a problem

2. Processing information
    - Using trial-and-error method
    - Using appropriate notations and combining them to the initial information
    - Getting the expression of the solution
    - Operating with numerals and other symbols
    - Drawing inferences
    - Generalizing objects, relations, and operations
    - Changing the direction of reasoning (forward-backward)
    - Curtailing the reasoning process or using some curtailing model
    - Making helpful drawings, figures, or graphs
    - Processing fast
3. Retaining and recalling information
    - Recalling terminology, formulas, or concepts
    - Recalling generalizations
    - Recalling problem type

These processes represented one perspective about the development of cognitive processes in mathematics; they also served as a basis for subsequently comparing the cognitive processes developed for the current study.

Summary of Literature Review

A literature review was conducted on four primary aspects of this study 1) the amount of new information gained when mixed item formats are used on one test, 2) the think-aloud methodology as a research tool, 3) issues related to item characteristics and

item quality, and 4) what is already known about mathematical cognitive processes. Each aspect is briefly summarized below.

The findings from the first section of the literature review were mixed. In some studies, the researchers found significant differences between item formats; findings from other studies indicated no differences. But, I did observe that these findings varied by content area. That is, in some areas constructed-response items seemed to add little or no information, but in others they seemed to add unique information about the process of making decisions on the basis of test scores.

The second section of the literature review cited the various ways researchers have used think-alouds. For instance, researchers used think-alouds to (a) determine whether test items actually measure the cognitive processes intended by the item writers, (b) examine group differences, and (c) examine how useful the verbal data would be for supporting the findings from a statistical analysis. One researcher advocated the use of think-alouds during test development as another way to assess item performance. Although none of the studies employed think-alouds to specifically examine and identify cognitive processes used by test takers, they do confirm that the qualitative methodology would be an effective method for answering this type of question.

Findings from the third section of the literature review indicated that it was difficult to create item pairs from already existing items. I concluded that the limited number of items in a test's item bank and the unbalanced number of multiple-choice and constructed-response items available in an item bank compound the difficulty of matching two items.

The last section of the literature review pointed to the variety of procedures researchers have used to investigate the cognitive processes students used when answering mathematics items. Some of the researchers used prominent learning theories to create *a priori* categories, which were then used to code interview data. One researcher allowed the categories to emerge from retrospective interview data. Most of the categories differed across the studies but two researchers concluded that middle school students often misapplied algebraic rules when solving items.

Contribution of this Study

This study contributes to the literature in a few unique ways. First, all of the researchers who have examined the value-added of combining multiple-choice and constructed-response items on a test used analytical models. Although my question examines the issue from a value-added lens as well, I also employed a qualitative approach that I believe better assesses the question in general, and my question in particular. For example, if the goal of examining value-added is to ascertain the amount of added technical information (i.e., an IRT information perspective) gained by combining item formats on a test, then analytical models are the most appropriate means for that examination.

But I took a different perspective on the value-added question and focused on whether we gain information about the content area by looking at the cognitive processes students used when solving the two item formats. The think-aloud procedure better assesses this question rather than an analytical model. A secondary, but no less important, purpose of the study was to encourage practitioners and psychometricians to consider the benefit of the think-aloud methodology to inform classroom instruction and

curriculum and an item's contribution to the content area being assessed. Think-alouds can illuminate both the similar and different ways that students solve algebra test items, which could result in a change in the way teachers instruct or test developers write items. An analytical model would not have provided as useful information for these types of purposes.

Third, the results could conceivably contribute to the art of item writing. The item writing technique used in this study could be used to generate multiple items quickly, in both multiple-choice and constructed-response item formats. All of the items would presumably measure the same content but they would perhaps elicit different cognitive processes and thereby contribute the depth of assessing the content area.

Last, the results of this study provide information about what cognitive processes 8th grade students use when solving algebra items regardless of format. Researchers could compare and contrast these processes with their own research experiences or with other research available in the literature. Other researchers could single-out the methodology and duplicate the study using another content area.

Regardless which parts of the study are excerpted, the overall contribution of this study is two-fold. One, I hope to encourage measurement professionals to think about item information in an alternative way than it is traditionally considered. Two, I want to encourage measurement professionals to use a non-traditional measurement tool as they continue to examine how students interact with test items.

CHAPTER III

Study Design and Procedures

Research Question

A grounded theory model was used to examine the verbal protocols and answer the following research question,

> Are different cognitive processes used by test takers when responding to multiple-choice and constructed-response mathematics items?

Sample

The sample was drawn from two school districts in the Lansing, Michigan region. Two schools participated in the study, one school from each school district. One school was in an urban school district setting, with an ethnically diverse student population, and a range of low to middle socioeconomic status. The other school was in a suburban setting with a less ethnically diverse student population, comprised primarily of students from the middle socioeconomic status.

All of the students were enrolled in the 8^{th} grade. No students were intentionally omitted from participating in the study, but, as will be explained later, not all participated. Gender and ethnic information were used to provide details about the composition of the sample, rather than used as independent variables. See Table 1 for the demographic composition of the sample.

Table 1. Demographic Information.

	Urban School	Suburban School	Total Students
Female	8	10	18
Male	9	7	16
Total	17	17	34
African-American	2	1	3
Asian	1	0	1
Hispanic	3	0	3
White	7	16	23
Multi-racial	1	0	1
Other	1	0	1
Blank	2	0	2
Total	17	17	34

As seen, the number of boys and girls comprising the sample is similar. Most of the students were white (68%), with a small number of students represented by the other ethnic categories.

Procedure for Sample Selection

The selection of students was nonrandom because the teachers had the option of withholding their classes from participation. Teachers at the urban school selected students in two out of four 8th grade mathematics classrooms and teachers at the suburban school allowed students from four out of six 8th grade mathematics classrooms to participate. The four suburban classrooms represented four different levels (tracks) of mathematics instruction: transitional, regular, pre-algebra, and algebra. Classrooms in the urban school were not tracked, or at least not identified by the teachers as being tracked.

Parent permission was obtained prior to data collection. To facilitate this process, the classroom teachers distributed, and collected, parent permission letters to every 8th grade student in the selected classrooms. Passive parental permission (if parent did not

return the letter, then his/her child could participate in the study) was used in both schools. Over 400 students received permission and became part of the sampling pool. For unknown reasons, ten suburban parents and one urban parent denied permission.

As parent permission was obtained, I kept a master list of student names that was subsequently used to select the sample. (Students and teachers knew from the beginning that not every student would be asked to participate in the think-aloud study, as only a small number of students were needed.) A sample of 24 students was originally planned but I over sampled to account for attrition and other unforeseen problems that would reduce the final sample size. I selected 34 students — 17 students from each of the two schools — using a version of systematic sampling with a random start.

This type of sampling method can result in a biased sample if the list is ordered (i.e., alphabetical or rank ordered according to a criterion measure) (Fraenkel & Wallen, 1993). The list was not ordered in any particular way. Nonetheless, as another precaution against bias, I showed the 34 student names to their respective teachers to verify that a range of mathematics achievement was represented. The teachers verified the sample's range of achievement.

Because of the initial limitation imposed by the teachers, it was impossible to attain a true random sample. But, the sample was randomly selected within the school sampling constraints. Perhaps even more important, the sample size for this study was very small. Thus, combination of the constrained random sample and small sample preclude generalizing the results beyond the 8th grade or generalizing the results beyond the schools where this study occurred.

Instruments

Three instruments were used to collect the data: the test booklet, the protocol guide, and a short demographic survey. The interviewers also used the protocol guide to record notes about students' responses during the think-alouds. The instruments were all pre-coded with a unique number that was matched to each student's name.

The test booklet. Two test booklets comprised of items originally appearing on the National Assessment of Educational Progress (NAEP) program and one item from the Balanced Assessment Package (1997) project were used to collect the data. A total of 17 algebra items appeared on each booklet. In each booklet, eight of the items were multiple-choice items and nine items were constructed-response. The two item formats were dispersed throughout each booklet. The non-secure items are presented in Appendix B.

The protocol guide. The protocol guide mirrored the test booklet, with the addition of item specific prompts and space for the interviewers to record notes. The interviewers used the protocol guide as a script, which ultimately served to help standardize the think-aloud procedure.

The surveys. The students responded to a short demographic survey. They provided information about their gender, age, frequency of doing math and reading homework, and school name. The students recorded the information themselves.

Algebra Strand

Typically, a content framework defines the content and cognitive processes measured on a test. Framework developers often write a framework to represent a particular mathematics program. Although several different mathematics programs are in

use, the intent of this study was not to evaluate a particular program or to compare two or more programs. Instead, an effort was made to select a neutral mathematics framework — a framework that reportedly did not promote a specific mathematical program. The National Assessment of Educational Progress (NAEP) was a testing program that meets this criterion.

Because I used NAEP items as the medium for data collection, the five mathematical content strands assessed on NAEP limited my choices of content areas. The NAEP mathematical construct is defined as five content strands: number sense, properties and operations; measurement; geometry and spatial sense; data analysis, statistics, and probability; and algebra and functions. Of the five content strands, the selection of algebra, from which to select items, was not an arbitrary decision for many reasons. First, the NCTM *Standards* targeted algebra instruction at the eighth grade (Silver, 1997). Second, it is commonly the students' first class in mathematics where they are introduced to abstract concepts compared to the more concrete mathematical operations and number manipulations taught in early grades. And third, understanding algebraic concepts provides the foundation needed to be successful in more advanced mathematics courses.

Furthermore, the cognitive processes available to be studied were limited to the cognitive processes defined on NAEP. As described in the NAEP mathematics framework (1996) items are written to assess one of three mathematical abilities: conceptual understanding, procedural knowledge, and problem solving. The mathematical abilities describe the characteristics of the knowledge or process needed by the respondent to successfully manage the task presented in the item. Thus, when we are

in the parlance of NAEP, mathematical ability represents the cognitive process an item is supposed to elicit.

The NAEP framework provides detailed descriptions of the three processes assessed on the test. According to the NAEP framework (1996), conceptual knowledge is defined as a class of objects that share a common set of characteristics. Procedural knowledge is defined as a series of related actions connected with an object or result. And, problem solving is defined as a combination of conceptual and procedural knowledge. The NAEP designers provided more complete and descriptive definitions for each of the mathematical abilities.

Conceptual Understanding

- recognize, label, and generate examples and nonexamples of concepts;
- create, interpret, and relate models, diagrams, graphs, and varied representations of concepts;
- identify and apply mathematical principles;
- make valid statements that generalize relationships among concepts in conditional forms;
- understand the meaning of facts and definitions;
- compare, contrast, and integrate related concepts and principles;
- recognize, interpret, and use the signs, symbols, and terms used to represent concepts; or
- interpret the assumptions and relations involving concepts in mathematical settings.

Procedural Knowledge

- select and apply appropriate procedures correctly;
- analyze the efficiency of different procedures;
- verify or justify the correctness of a procedure using concrete models of symbolic methods;
- apply important formulas; or
- extend or modify procedures to address factors inherent in problem settings.

Problem solving

- use accumulated knowledge of mathematics in new situations;
- recognize and formulate problems;
- understand assumptions made with respect to given information;
- use strategies, data, models, and relevant mathematics;
- generate, extend, and modify procedures;
- use reasoning in new settings (i.e., inductive, deductive, algorithmic, or algebraic); or
- judge the reasonableness and correctness of solutions.

Students were not likely to use all of the components within each definition as they respond to any one item, but the complete definition was given for the reader's benefit. And, because any one item cannot capture every component of the definition, the items selected for the study limited which part(s) of the definition(s) were used by the students as they responded to an item.

Design

The study was designed to determine whether students use different cognitive processes when responding to multiple-choice and constructed-response items. The presentation of the design is organized into several sections. The first section is a review of a pilot study conducted prior to this study, because some of the decisions for this study were based on what I learned from the pilot study. The second section is a description of how the items for this study were developed. The third section explains how the items from each item family were assigned to the test booklets and to the students. The fourth section describes the testing procedures used to collect the data.

Section 1: Pilot Study

I conducted a pilot study during the summer of 1997 for another research project (Pearson and Garavaglia, 1997). Eighth grade children participating in an after school program participated in the study. Many of the students were from the urban school used in the current study. Three questions were examined,

1. How many items could 8th grade students answer during an hour think-aloud session?
2. Could 8th graders sustain thinking aloud for an hour?
3. Which mathematical strand, either algebra or measurement, worked best during a think-aloud?

Findings from the pilot study indicated that students could easily answer up to14 items (7 multiple-choice and 7 short constructed-response) in a 50-minute to one-hour think-aloud session, without experiencing fatigue.

Findings related to the third question indicated that measurement items provided almost no evidence about the cognitive processes used by respondents when solving the items. First, the items were so basic and not engaging that the evidence gathered was not very telling. This was seen for all of the measurement items piloted. Second, many of the measurement items required the students to use a ruler or a protractor to measure a diagonal or an angle. The students talked about how they *used the tool* rather than how they *solved the problem.* Some mathematics educators may say that describing how respondents use a tool is evidence about how one solves an item, but the items themselves did not allow for variation in responses, because they were very easy. Based on the findings from the pilot study, measurement items were not included in the current study.

Item Selection for the Pilot Study

This section briefly describes the number and types of items selected for the pilot study. The entire pool of 1992 and1996 released algebra and measurement NAEP items were available to select the pilot items. A total of 14 items were selected for the pilot study; seven of the items assessed algebra and seven assessed measurement. The item selection process was limited by the number of constructed-response algebra items in the set of released NAEP items. There were only three constructed-response algebra items in the entire set of the released algebra items, therefore all three of the constructed-response items were included in the pilot study. Four constructed-response items from the measurement area were then selected. In order to have seven items in each area, four algebra items were multiple-choice items and three measurement items were multiple-choice. See Table 2 for the distribution of measurement and algebra items by item

format. As displayed, a total of 14 items were used in the pilot study, seven from each mathematical strand and seven from each of two item formats.

Table 2: Distribution of Item Format and Mathematical Strand

Math Area	MC Format	CR Format	Total
Measurement	3	4	7
Algebra	4	3	7

In addition to the number of items represented in each item format, three other item-related criteria were used to select the items. Two of the criteria were statistical in nature and the third criterion was based on the cognitive process (procedural knowledge, conceptual knowledge, or problem solving) associated with the items. Item difficulty and discrimination statistics (the IRT parameters were obtained from operational administration of the NAEP items) were used to obtain a range of items, in terms of their statistical properties. Although an attempt was made to select items from each of the three cognitive processes, many of the items came from the problem-solving dimension.

Items within each strand, measurement and algebra, were matched across item format by using the above three item related criteria: item difficulty, item discrimination, and cognitive process. A match was defined as two algebra items, for example, from different item formats, with similar item difficulty and discrimination statistics, and the same cognitive process. However, by the end of the pilot study I learned, as did Campbell (1995), that matching on these criteria did not necessarily result in "perfectly" matched item pairs. Furthermore, obtaining close matches using the three criteria was often difficult when using already existing items — the items were not intentionally developed for the purpose of this type of study. Based on what was learned from the

pilot study and from Campbell's (1995) experience, it became apparent that an alternate method for matching items was needed for the full-scale study, namely a method that did not exclusively rely on existing items.

An unanticipated finding from the pilot study also informed the current study. I developed a protocol guide that included common prompts across items and item-specific prompts, to standardize the think-aloud sessions. The protocol guide booklet worked well. The protocol was user-friendly and the prompts were easily understood by the students. The protocol guide was very helpful during the think-aloud sessions because it standardized the think-aloud sessions across students (Ericsson and Simon, 1993). I retained the protocol guide for this study.

The information obtained from the pilot study was valuable in many ways. The three questions examined by Pearson and Garavaglia (1997) provided information about how to design a think-aloud study. And, I learned how to conduct a research study in a school setting (e.g., negotiating a space, getting students out of class, accurately projecting the length of time students would spend in a think-aloud session). Everything that was learned during the pilot study was carried forward to this study in an attempt to improve upon what should or should not be done.

Section 2: Item Development

The goal of writing items for this study was to develop multiple-choice and constructed-response algebra items that were as comparable in content as possible. It was important to hold content constant across formats so that any differences in cognitive engagement that might be observed could be attributed to format. In other words, item format was the only item related factor allowed to vary in this study.

There is no one tried and true method of developing comparable or parallel items mentioned in the literature. To maximize the chance of developing genuine comparability across item formats, I relied on the experiences of others (c.f., Campbell, 1995; Frederiksen, 1984) as well as on existing item development procedures (Haladyna, 1994). One way to write similar items in two different item formats is to use already developed multiple-choice and constructed-response items and transform them into the corresponding item format (Frederiksen, 1984). Fredericksen suggests that this approach maximizes the likelihood of obtaining construct equivalence. On the surface, removing or adding response options to existing items seems to be a good suggestion. However, Frederiksen also suggests that existing items should be used when making the conversion. But, taking into account what was learned from the experiences of other researchers, I did not think that Frederiksen's suggestion was sufficient, in and of itself, for the purpose of this study. So, I buttressed Frederick's item conversion suggestion with Haladyna's (1994) item-shell method of writing items. The term "model-shell" references the item writing method used for this study.

Model-shell Item Selection Criteria

I started the item development process by choosing eight NAEP items and used them as models for developing the other items. Because the model item became one of the items studied, the model had to meet certain item selection criteria. The four criteria were:

- Items had to measure algebraic patterns.
- Items represented a range of item difficulties (to ensure variability in the data).

- Respondents had to use different algebraic equations to solve the items.
- Four of the eight model items had to be multiple-choice items and four had to be constructed-response items.

The first three item selection criteria were met; however, the last criterion was not met. After sorting all of the 1992 and 1996 released and secure NAEP items that measured algebraic patterns, I discovered that few of these items were written in the constructed-response format. Achieving a perfect balance in the number of multiple-choice and constructed-response model items was not possible. The final selection of model items was five multiple-choice and three constructed-response items.

One of the multiple-choice model items is used here to simultaneously illustrate the item writing procedure and to introduce the notion of a "family" of items — two multiple-choice and two constructed-response, each with similar content. I followed the same item writing process whether the original item was a multiple-choice item or a constructed-response item. Table 3 displays the composition of items within an item family.

Table 3. Composition of an Item Family

Item	Format
1	Original –original content and format (either MC or CR)
2	Converted –Change the format: MC to CR or CR to MC
3	Transformed – This is a "clone" of the original in the same format as the original. For example, different numbers or different stimuli (stars versus dots) might be used
4	Converted Transformed – Change the format of the transformed (clone) item.

Item Writing Procedure and an Example

The first item in the family was always the original NAEP item. To write the second item in the family, the response options were removed to convert a multiple-choice item into a constructed-response item (response options were added to convert a constructed-response item into a multiple-choice item). The conversion left the item stem in intact. An example is provided.

Original NAEP multiple-choice item:

Puppy's Age	Puppy's Weight
1 month	5 lbs.
2 months	12 lbs.
3 months	17 lbs.
4 months	20 lbs.
5 months	?

1. Jim records the weight of his puppy every month in a chart like the one shown above. If the pattern of the puppy's weight gain continues, how many pounds will the puppy weigh at 5 months?

A. 30
B. 25
C. 23
D. 21

Converted constructed-response item:

Puppy's Age	Puppy's Weight
1 month	5 lbs.
2 months	12 lbs.
3 months	17 lbs.
4 months	20 lbs.
5 months	?

2. Jim records the weight of his puppy every month in a chart like the one shown above. If the pattern of the puppy's weight gain continues, how many pounds will the puppy weigh at 5 months?

Answer:_____

At this point, two of the four items in the item family were written. As seen in the example, the same algebraic equation can be used to answer both items, regardless of the item format.

The original multiple-choice item served as a model-shell to write the third and fourth items of the item family (see Table 4). To hold the content constant in the item family, the algebraic concept assessed in the original item was changed by slightly altering some feature of the item, such as the algebraic equation needed to solve the problem. For example, in the original and revised sample items, the pattern for the puppy's weight gain is decreasing by a difference of two pounds each month. For the transformed items, the pattern for the puppy's weight gain decreases by a difference of one pound each month. The transformed constructed-response (multiple-choice) item was converted to the rewritten transformed multiple-choice (constructed-response)

format. The two examples below exemplify the development of the third and fourth items in an item family.

Transformed constructed-response item:

Puppy's Age	Puppy's Weight
1 month	10 lbs.
2 months	15 lbs.
3 months	19 lbs.
4 months	22 lbs.
5 months	?

3. John records the weight of his puppy every month in a chart like the one shown above. If the pattern of the puppy's weight gain continues, how many pounds will the puppy weigh at 5 months?

Answer: _____

Rewritten transformed multiple-choice item:

Puppy's Age	Puppy's Weight
1 month	10 lbs.
2 months	15 lbs.
3 months	19 lbs.
4 months	22 lbs.
5 months	?

4. John records the weight of his puppy every month in a chart like the one shown above. If the pattern of the puppy's weight gain continues, how many pounds will the puppy weigh at 5 months?

A. 30
B. 27
C. 25
D. 24

The four items represent a "family" of comparable items, two written in the multiple-choice format and two written in the constructed-response format. In total, there were eight families of four comparable items developed, resulting in 32 items. (See Appendix B. To maintain the integrity of the secured items, only the public released NAEP items appear in Appendix B.) As seen, the NAEP constructed-response items consist of short answer (one or two sentences), fill-in the blank, or extended constructed-response item-types.

The final step in the item development process consisted of a content review to ensure that all of the items measured algebraic patterns. A mathematics educator reviewed the 32 items for content validity. She also reviewed the four items within an item family to review their content comparability.

In addition to the NAEP items, a performance item from the Balanced Assessment Package (1997) was included in the item set. Items from the Balanced Assessment Package were intentionally developed to be integrated with classroom instruction and to assess mathematical concepts common to middle school curricular goals.

The Balanced Assessment item selected for this study measured an algebraic pattern, consisted of multiple, scaffolded steps, and required several minutes to solve. I purposefully added this item to the assortment of NAEP items so that students could respond to an item that had been intentionally developed to appear on a performance assessment. So, if I found no differences between the NAEP constructed-response and multiple-choice items but I found some differences in the cognitive processes elicited by the Balanced Assessment item, I then would be able to attribute the absence of between

item format differences to the idea that the constructed-response items did not tap the sorts of cognitive processes that were tapped by the performance item. To this end, the constructed-response items could be thought of as multiple-choice items in disguise.

In summary, the purpose for writing new items rather than only using existing items, was to obtain a set of comparable items whose content would be similar across two item formats. The item writing process was purposefully developed because, as indicated in previous research (Campbell, 1995; Haladyna, 1994), matching items by using item difficulty and discrimination statistics is not a guarantee that the matched items will have equivalent content.

Section 3: Assignment of Items and Students to Forms

Recall that four items defined a family of items. Placing items with the same pattern but a different item format in a test form would likely introduce item dependency and a practice effect. To address these issues, I assigned items with the same pattern to two different forms. Conversely, items with the altered patterns were assigned to the same form. Table 4 represents a sample assignment of items to different forms. This distribution resulted in the assembly of the two test forms.

Table 4. Assignment of Items to Test Forms

Form Designation	Item Assignment
A	Original content and format (multiple-choice)
A	Transformed constructed-response: slightly changed content-different format than original
B	Converted constructed-response: same original content-different format than original version
B	Rewritten transformed multiple-choice: same slightly changed content-different format than transformed version

The two forms were randomly assigned to the 34 students. Random assignment would control for pre-existing achievement level differences within the non-random sample (Stanley and Campbell, 1963). And, random assignment of forms would control for curricular and instructional differences between the classes.

By randomly assigning the forms, half of the students responded to two members of an item family (say 1 and 4) while the other half responded to the other two members of a family (2 and 3). One member of a given family was randomly assigned to a serial position within the first half of a form; the other member was assigned a comparable position within the second half of that form. The last item in each form was the performance item from the Balanced Assessment package.

Section 4: Testing Procedures

An interviewer escorted each student from his or her regular classroom to a quiet room where the think-aloud took place. Prior to the start of an interview, the interviewer told every student what to expect during the think-aloud session, to eliminate or reduce

any feelings of nervousness or apprehension. The explanation included the interviewer's role and the student's role throughout the session. Furthermore, the interviewer assured each student that his or her answers to the items would not count towards classroom grades. The interviewer also explained that the intent of the think-aloud was to obtain verbal accounts of what the student was thinking as he or she solved the items, rather than whether or not the student provided a correct answer. Finally, the interviewer used a protocol guide during every think-aloud session, to standardize each of the 34 sessions.

Seventeen algebra pattern items presented in multiple-choice (8 items) and constructed-response (9 items, one was a performance task) formats were administered during a single think-aloud session. Each session lasted about an hour and was audiotaped. The reason for taping the think-alouds was to facilitate the transcription of the qualitative data. Hand-written notes also were taken during the think-alouds, however copious notes were not recorded to ensure that the interviewers would not miss something a student said or miss an opportunity to probe a student's verbal account.

The following steps were followed for every think-aloud:

1. Introductions between the interviewer and the students.
2. The interviewer explained what a think-aloud was and shared with the student exactly what would happen during the session.
3. The demographics survey was completed by the student.
4. The interviewer began the session with a warm-up think-aloud question. The interviewer answered the question first to demonstrate how to think-aloud. The interviewer then presented the same warm-up question to the student. The student answered the question while thinking-aloud. (The warm-up

question asked was "how many times have you talked on the phone over the last 3 days?").

5. If the student did not have questions, the think-aloud began.
6. The students were instructed to read every question aloud, and then verbally express what they were thinking while they solved each item.
7. When necessary, the interviewer reminded students to "think-aloud" if they became quiet, or introspective, while answering an item.
8. The students continued through all 17 items at their own pace.
9. The interviewer administered the "think-aloud method perception survey".
10. The interviewer asked the students whether they had any questions about the session.
11. The interviewer thanked the students for their participation.

Threat of Confounding

Potential threats to the outcome of the study needed to be realized, and if possible controlled for, prior to its implementation. One potential threat may come from some students feeling inhibited to express themselves verbally because of the audiotapes. To address this threat, the students were assured their comments would be kept confidential and anonymous.

Lack of student motivation may be one of the largest threats to obtaining accurate and complete information in situations like this one. That is, the students knew that no stakes were attached to their performance on the items and therefore they may not have exerted much effort to solve them. This phenomenon often is found when pilot testing new items. To counter this likely problem, the interviewer encouraged the students (and

teachers) to take this study seriously and to do their best when answering the items. The interviewers also told the students that the purpose of the study was to determine how they solved the items rather than on the number of items they solved correctly.

Data Collection

Data collection occurred during the spring of 1998. All of the data were collected within two weeks. Tape recorders were used to facilitate data collection rather than relying on interviewer notes alone. One benefit of recording the sessions was to decrease data recording errors that would likely occur with hand written accounts. The interviewers were also free to concentrate on probing the students.

Conducting the data collection in two schools introduced a few logistical issues. First, a quiet location with an electrical outlet for the tape recorder was needed for the think-aloud sessions. And second, I had to work within the schedule provided to me by the teachers. As it turned out, neither of the logistical issues was difficult to solve. Adequate space was provided at both schools and the teachers were very flexible with their classroom schedules.

Four trained interviewers and I conducted the think-alouds. Two interviewers were involved in the pilot study and were therefore already familiar with the protocols. I trained two additional interviewers to use the interviewer protocol guides and to conduct a think-aloud session. And, prior to conducting an interview, both of the interviewers observed one of the three experienced interviewers conduct a think-aloud, to further familiarize them with the process. Because of the limited time in which to collect all of the think-aloud data, the two novice interviewers did not conduct an initial, supervised think-aloud interview. Instead, I monitored their interviews by sitting in on some think-

aloud sessions to ensure that they followed the protocol guides and that they did not ask leading questions. I was accessible to the interviewers throughout data collection.

Data Analyses

The purpose of this section is to present the analyses used to examine the research questions. The data analyses consisted of five steps. These were:

1. Use descriptive statistics to examine each item's difficulty, standard deviation, and frequency distribution of score points.
2. Use the grounded research approach to identify and validate emerging categories (in the tradition of the constant comparative analysis) in students' verbal protocols.
3. Complete a full-scale analysis using the identified themes.
4. Compare cognitive process similarities and/or differences between item formats within item-pairs. Create an index from the original themes that represented depth of cognitive processing engagement.
5. Conduct a post hoc evaluation of Steps 2 and 3 using external evaluators who have content expertise and curriculum and instruction knowledge.

Descriptive Item Level Statistics

Descriptive statistics were calculated separately by form. All non-responses (e.g., skipped items) were considered wrong answers and subsequently re-coded as zeroes. I first calculated frequencies, maximum, and minimum statistics for all variables (e.g., student id, test number, form, school, item1 through item16) to verify that the data was keyed in correctly. I then calculated traditional classical test theory item means and standard deviations to get an initial examination of each item's distribution. Finally, I

calculated the mean score on the eight multiple-choice items and the mean score on the eight constructed-response items.

Identify and Validate Cognitive Processes

The verbal data were used to identify which cognitive processes the students used when answering the algebra questions. To that end, a grounded theory approach was used to examine whether students used different cognitive processes when responding to multiple-choice and to constructed-response items. The first step in analyzing the data was the establishment and validation of the cognitive processes used by the students as they answered the algebra items.

The steps for identifying and validating the cognitive processes are presented here rather than in the methodology section for two reasons: (a) they are integral parts of the protocol analysis phase, and (b) grounded theory blurs methodology and analysis. To facilitate the initial development of categories that exemplified the cognitive process, six interviews were transcribed (almost verbatim) so that the cognitive moves were easily identifiable. Two graduate students and I began the analysis by examining several responses to one item and recording the cognitive processes used to answer the item. We then broadened our analysis by carrying the cognitive processes forward to other items and different students, revising, adding, and deleting cognitive processes as necessary (i.e., open coding).

We developed plausible categories that accounted for most of the verbal data. We then tested the categories with another tape to build our confidence that the categories accounted for most of the responses (category saturation). This constant comparative nature of grounded theory gave the emerging concepts specificity because we

continuously asked questions of ourselves while we established the categories (Strauss and Corbin, 1990).

During the initial phase of identifying the cognitive processes described above, we listened to six tapes (two tapes per person) and independently recorded the cognitive processes verbalized by the students. To ensure that we were on a similar analysis path, we met after listening to eight items and discussed the cognitive processes identified. We identified similar processes and were able to justify the ones that differed. We compiled a larger list of cognitive processes by combining the processes each researcher independently identified. We each then finished recording the cognitive processes associated with the remaining eight items.

After coding two tapes, we met again to compare notes. Twenty-eight categories exemplified the cognitive processes used by the eighth graders (see Figure 1 below).

	Category	Definition	Example
A	Overall pattern recognition	Indicates an overall grasp of the item	Understands pattern represented in item (from beginning to end of item) "pattern repeats itself."
B	Pattern not recognized	Indicates lack of understanding	Test taker indicates, "I can't figure out what the pattern is" "I know there's a pattern, but I don't see it."
C	No information used from item	Indicates lack of organizing information	"I've seen an item like this before and the answer was —."
D	Partial information used from question	Indicates concern for organizing or fully understanding information	Student knew 28x2=56 but then did not add last two tacks. Or, uses information in beginning of pattern and ignores information in middle and end of pattern.
E	All information used from question	Indicates thoroughness in organizing information	Determines pattern by using the information given in item, e.g., uses all numbers listed in a column, not just first few numbers and then skips the rest.
F	Visual representation (e.g., draws picture)	Indicates importance of transforming information into a manageable framework	Draws chart, picture, or table to solve item. No indication student understands an algebraic equation can also be used. "I have to draw a picture to solve this." I have to make a chart to figure out the pattern."
G	Applicable equation used	Indicates a connection between problem and learned mathematical knowledge	Uses an equation to solve item. "The equation is 28x2+2" "I solved item by using picture+1= number of pictures."
H	Informed guess (uses some data given in item, e.g., information in the multiple-choice options)	Indicates concern for understanding problem	Uses mc options as a guide to solve item. "I looked at the answers and used B to solve the pattern." Knows answer is wrong because it isn't listed as an option.
I	Guess without use of information given (blind guess)	Indicates lack of understanding problem	Student admits to guessing. "I picked an answer that looks the best." "I don't know, I just guessed."

J	Calculation error	Indicates lack of concern or attention	Subtracts rather than adds. Adds numbers wrong.
K	Calculation error, but adjusts answer to fit choices provided	Indicates ability to recognize error and connect it to problem	Provides a wrong answer then recognizes answer is wrong. Re-works item.
L	Non-applicable equation used	Indicates inability to connect problem and prior knowledge	An equation is used that doesn't fit the pattern. "28/2=14-2=12"
M	Test "wiseness"	Indicates some ability to connect problem with prior knowledge	Uses something in item to help solve it. "That choice was weird b/c item said that she didn't want to draw all to the dots." "D is too big. C is too low and 220 is kinda low. So, 420 is the answer." "My answer doesn't make any sense."
N	Information from previous situation recalled	Indicates carry-over from one situation to another	Student recalls how he/she solved item in different situation. "That's how I solved the item before."
O	Information from previous, comparable item recalled (carry-over effect)	Indicates carry-over from one item to another item	Student recalls how he/she solved comparable item on test. "This item looks like the other one."
P	Student returns to question and changes answer	Indicates concern for understanding problem	Student returns to problem after solving the comparable item and changes answer. "I think I did the other one wrong. I'm going to go back and check."
Q	Uses estimation	Indicates ability to connect response with likely answer	Solves problem to certain step and then sees answer is higher than 2 of the mc options and thinks another is too high/low. Or, picks an answer from mc options that is close to answer student computed. "because 420 is the closest to my answer."
R	Mental math - work not shown	Indicates organizational method	Student doesn't have to solve item by writing pictures or equations on paper.
S	Student checks work	Indicates thoroughness in overall approach	Checks solution by using other information in item. "I came up with an equation and checked whether it was correct by seeing if it worked for steps 2 and 3."

T	Misinterprets question asked/answers question other than that being asked	Indicates inability to connect problem with prior knowledge	Student thinks question is asking him/her to solve for something it really isn't. "I think they mean to solve for the area."
U	Partial pattern recognized	Indicates a grasp of the item	Thinks pattern stops at some point and a different one is used.
V	Complex pattern extension	Indicates understanding of item and ability to generalize process	Student identifies the pattern and then extends it several 'steps' beyond that which is given in the item. e.g., information provided for the first few steps and student has to solve for step 20.
W	Simple pattern extension	Indicates understanding of problem	Sequential steps in solving problem are provided. "It's a continuous pattern. The next arrow would be left."
X	No control of math vocabulary/says or writes operation but does not use that operation	Indicates concern for mathematical understanding	Student says add but then multiplies. Uses nonmathematical terms to express computation. "Numbers go up 5, down 2."
Y	Relationship between numbers given in question recognized	Indicates some concern for organizing information	When two sets of numbers are given, student sees a pattern exists between them..
Z	Relationship between numbers given in question not recognized	Indicates concern for ability to organize information	When two sets of numbers are given, student sees information given in each column as being independent. "I don't need to use the numbers in this column to figure out the pattern in this column."
AA	Grapples with information to try to solve question	Indicates a tendency to consider multiple data sources or possibilities	Tries multiple computational strategies to solve item. "That's not working so I have to try something else." Persists to solve item.
BB	Vocabulary in question not understood	Indicates concern for mathematical understanding	Doesn't understand mathematical terms. "I don't know what that word is." (infinity)

Figure 1. Cognitive Process Categories

To validate the categories — before starting the full-scale analysis — we each listened to the same two tapes and independently analyzed them using the 28 categories. We regrouped to determine whether additional cognitive processes were identified, to further explain and discuss our interpretation of the 28 categories, and to examine the degree of agreement in identifying the cognitive processes, for each item.

Because we all analyzed the same two tapes (see Figure 2 for a graphical representation) agreement was determined by comparing how each of us categorized each item.

Initial Round		Validation Round	
Tape	Researcher	Tape	Researcher
1/A	A	7/A	A
2/A	A	8/B	A
3/A	B	7/A	B
4/B	B	8/B	B
5/B	C	7/A	C
6/B	C	8/B	C

Figure 2. Tape Distribution Between Researchers

For example, researchers A, B, and C's categorizations were compared to each other. Agreement was defined as the percentage of matches across all items on a form. Agreement ranged from .91 to .98. This level of agreement indicated that we had internalized similar meanings of the 28 cognitive processes and that we were able to reliably identify the cognitive processes verbalized by the students.

The Categories. Refer to Figure 1 for the 28 cognitive processes that emerged from the analysis. The categories were not hierarchically arranged. As the analysis progressed, it became evident that the categories appeared in different frequencies within

and across items. In fact, some categories were used infrequently but frequency of appearance did not result in the deletion of a category.

Full-scale Analysis

The two graduate students and I independently analyzed the protocol data. Because of equipment failures and/or inaudible tapes, a total of 28 protocols (about 14 protocols from each of Form A and B) were included in the analysis. Responses to some items on a usable protocol were inaudible resulting in a different number of usable responses across the items.

As we listened to a tape, we recorded the category that represented the cognitive processes verbalized by the student in the same sequence as the students verbalized them. Besides using the cognitive process categorizations, the researchers took notes that explained/justified the identified cognitive processes. About mid-way through analyzing the protocols, another check of rater-agreement was conducted to ensure the reliable categorizations of the protocols. To calculate an agreement index, each researcher independently coded the same two protocols. Agreement here meant that all three researchers selected the same cognitive processes for each item. Agreement across the three researchers was high (.90).

To organize the qualitative data and facilitate analysis, the categories, notes, student information (e.g., student id, school), and item information (e.g., item format, right/wrong answer) were entered into a database. One record was established for each student (that is, each student represented an individual record). Queries were used to facilitate analyzing the data in the following ways,

- Cognitive processes associated with one item (i.e., frequency of categories), and
- Cognitive processes between item formats within item-pairs.

Cognitive Process Similarities and Differences

The next phase of the analyses involved examining the cognitive processes used to answer each item and then compare the similarities and/or differences of cognitive processes between the two item formats within an item-pair. For the individual item analysis, the cognitive processes were grouped for each item. The number of unique categories per item were examined to get a sense of the type of processes used to answer each item, regardless of item format. The between item format analysis appeared more informative and useful for answering the research question. Here, the cognitive processes between the item formats for item-pairs were analyzed using a meta-analysis-like approach. This analysis was done for all eight item-pairs on each form. The one Balanced Assessment item was compared to the other items, as it did not have a comparable multiple-choice item.

One of the graduate students and I independently examined the cognitive processes for the two comparable items on a form (see Table 5) and the Balanced Assessment item. I started the analysis by comparing the cognitive processes associated with each item in the item pair. Frequencies of cognitive processes were computed for each item and presented in tables. Each item's mean was presented as well. A narrative account was prepared for each set of items.

Closer examination of the 28 cognitive process categories revealed that some represented deeper engagement of cognitive processes than did others. Thus, to focus the

analysis, the 28 categories were examined to determine which categories represented key elements associated with deeper cognitive engagement in algebra. To aid in identifying the key elements, I examined the cognitive processes used by some low and high performing mathematics students.

Nine categories were selected as indicators of deeper cognitive engagement (see Figure 3). The categories were not listed in the table in any kind of hierarchy. The following rationale led to their selection. The first two categories were prerequisites for understanding the area of algebraic patterns. They also would provide evidence about the degree of understanding the students had about patterns. The next two categories (E and D) would capture whether students had the capacity to identify and use information that was necessary to solve the item. To elicit these processes, the students would have to be mentally engaged with the item to even begin to think about how to solve it. Categories F and G indicated engagement because the student would have to be thoughtful about the way he or she decided to manipulate the information. The student would also have to re-present the information given in the question using one of these two processes. The observance of category Y would occur if the student understood one of the fundamental concepts in algebra and therefore would facilitate the student solving the item. Furthermore, the processing and manipulation of variables would occur only if the student was being thoughtful as he or she made the cognitive moves. I thought category S indicated engagement because it represented a thoughtful and deliberate action on the part of the student. And finally, category AA would occur when the student had a difficult time finding a solution for an item and therefore would have to change strategies or continue to ineffectively fumble with information. Neither process would occur if the

student were disengaged from the situation — in this case, the item. The other categories indicated some thoughtfulness as the student solved an item (e.g., H, K, P, Q, and R), a description of the item itself (V and W), or a description of the events as the student solved the item (e.g., J, N, O, P, and X). None of these categories represented depth of engagement when compared to the nine selected categories.

Code	Cognitive Process Category	Reason for Selection
A	Overall pattern recognition	Indicates an overall grasp of the item
U	Partial pattern recognition	Indicates a grasp of the item
E	All information used from question	Indicates thoroughness in organizing information
D	Partial information used from question	Indicates concern for organizing information
F	Visual representation of information (e.g., makes chart or table)	Indicates importance of transforming information into a manageable framework
Y	Relationship between numbers recognized	Indicates some concern for organizing information and the recognition of variables
G	Applicable equation used	Indicates a connection between problem and prior knowledge
S	Student checks work	Indicates thoroughness in overall approach
AA	Grapples with information to try to solve question	Indicates a tendency to consider multiple data sources or possibilities

Figure 3. Depth of Cognitive Engagement Categories

The final step in the analysis was to use the nine categories to report a "depth of cognitive process index". However, after further inspection of the data, it became evident that additional analyses would be helpful. Thus, in addition to reporting the depth of

process index, one additional index common to all items (overall pattern recognition) and two to three family specific indices were reported because it became evident that not every category had equal relevance to every family.

The item family was the focus of the analysis. To examine the trends across and within the item formats, the four items within an item family were grouped in a table. The depth of cognitive engagement global index was the arithmetic mean of the proportions of each of the nine categories. The global index for each item was obtained by summing the frequency of the nine categories (see Table 7) and dividing by the number of students who responded to each item times nine (the number of categories). The denominator varied because of the different number of students (between 13 and 17 students) who responded to an item.

For the general and family specific indices the frequency as well as the relative frequency (proportion) of responses were reported. The proportion for each item was obtained by dividing the frequency of the category (see Table 7) by the number of students who responded to the item. The value in the denominator varied because different numbers of students responded to each item. The mean for each item was reported to provide an index of difficulty.

Post hoc Evaluation of Steps 2 and 3

A small panel of math educators and 8th grade teachers were convened to review the category system and the process used to develop the system. Specifically, the panel was asked to listen to and code two students' think-aloud interviews using the 28 cognitive processes. Almost simultaneously, I replicated the think-aloud procedure with one of the 8th grade mathematics teacher from the suburban school. These activities

served several purposes: (a) validate the category system and the coding process, (b) learn whether the teacher used similar cognitive processes as the students, and (c) determine whether additional cognitive processes emerged when an expert solved the pattern items.

The panel met for a one-day training session. The purpose of the training was to familiarize the panel with the development of the category system and how it was subsequently used to analyze the verbal data. During the session, the panelists were shown copies of the original protocol guides, the student test booklets, and the list of 28 cognitive process categories. During the session, the panel had time to review the training materials and reflect on the activities presented to them. However, most of active review work occurred outside of the training session.

The four panel members were divided into two review groups consisting of one teacher and one mathematics educator. Each of the two subgroups were given one think-aloud tape, a transcription of the think-aloud interview, and the associated test booklet. The reviewers were instructed to review the transcript and identify the cognitive processes expressed by the student. They were given the audiotape to supplement their understanding of the transcription. The reviewers were told to attend to whether the list of 28 categories captured the student's processes and whether they heard any processes that were missing from the list. The reviewers listened to their assigned tape individually.

During the same time frame as the reviewers were conducting these activities, I duplicated the think-aloud session with an 8^{th} grade teacher. Working within the teacher's schedule, I arranged for a one hour meeting to conduct the think-aloud. I

instructed the teacher to first solve each item using a strategy that she thought the students would probably use and then to look for alternative strategies that the students may have missed but that she saw because of her content knowledge. The latter instruction was used as a vehicle to generate alternative cognitive processes, especially if the alternative strategy required the use of different cognitive processes. The teacher's think-aloud was audiotaped and subsequently analyzed using the same procedures I used to analyze the students' think-alouds. The results of this and the other four primary data analysis steps are presented in the next chapter.

Summary of Design and Procedures

The study was designed to determine whether students use different cognitive processes when responding to multiple-choice and constructed-response items. To examine the research question, thirty-four 8th grade students from two schools in two different school districts were selected to participate in think-aloud interviews. I selected the students using a version of systematic sampling with a random start. To build the master list of participant candidates, parent permission was obtained from over 400 students. Teachers, parents, and students knew that not every student on the list would be selected to participate in the study.

Three instruments were used to collect the data: the test booklet, the protocol guide, and a short demographic survey. Each student was administered the three instruments during their think-aloud interview. The main instrument was the test booklet because it contained the items that were used to elicit the students' cognitive processes. The items that appeared in the test booklet were written specifically for this study. The goal of writing items was to develop multiple-choice and constructed-response algebra

items that were as comparable in content as possible. It was important to hold content constant across formats so that any differences in cognitive processes that might be observed could be attributed to format. A total of 33 items were used for this study; 16 unique NAEP items and one Balanced Assessment item, which appeared on both test booklets. Seventeen items appeared on each test booklet. The test booklets were randomly assigned to the students.

Think-alouds were used to collect the data. I first used the procedure in a pilot study to iron-out some of the details (e.g., number of items 8th graders could comfortably answer in an hour) and to practice using the technique. For this study I applied what I learned from the pilot study and from the published think-aloud research (c.f., Ericsson and Simon, 1993). Specifically, I employed the concurrent think-aloud procedure, standardized the think-alouds by using protocol guides, wrote prompts for each item that were asked of every student, and trained and monitored the interviewers who helped me collect the data. I also audiotaped every interview to facilitate analyzing the data, rather than relying on the interviewers' notes.

I identified six main steps to analyze the think-aloud data. The steps consisted of both quantitative and qualitative analysis approaches. The quantitative tools consisted of item level statistics to gauge an item's level of difficulty, frequency counts of the category codes, and arithmetic means and proportions of a subset of categories for each item family. The qualitative tools consisted of identifying the categories that represented the cognitive processes and narrative accounts of the students use of the categories and the post hoc reviewers findings.

To review the data collection and analysis procedures, I convened four experts in the field of mathematics — two mathematics educators and two 8th grade mathematics teachers. I asked them to listen to a think-aloud interview, review the cognitive processes to determine whether they adequately represented the cognitive processes they heard during the interview and identify whether cognitive processes were missing from the list. The last part of the post hoc review consisted of duplicating the think-aloud and analysis procedures used on the students by interviewing one of the teachers and analyzing her think-aloud verbalizations. The main purposes of these activities were to (a) validate the category system and the coding process, (b) learn whether the teacher used similar cognitive processes as the students, and (c) determine whether additional cognitive processes emerged when an expert solved the items.

CHAPTER IV

Results

Descriptive Statistics

Form A. Descriptive statistics were calculated for the 17 students who responded to Form A during the think-aloud interview. Whether or not the students answered the items correctly was not the primary intention of the interviews, but the information was useful to review as it provided a general idea about student performance. The item means were reported in Table 5. The means were consistently higher for the multiple-choice items than for the constructed-response items, except for one item-pair. And, in two item-pairs the means for the two item formats where the same. Out of 24 possible points, the mean test score was 18.69 with a standard deviation of 4.21.

Table 5. Item Statistics - Form A

Item/Format	Mean	*Std. Dev.*	**Item/Format**	Mean	*Std. Dev.*
1/mc	1.0	*.00*	**9/cr**	1.0	*.00*
2/cr	.62	*.51*	**10/mc**	.70	*.48*
3/cr	.85	*.30*	**11/mc**	.92	*.19*
4/mc	.85	*.38*	**12/cr**	.69	*.48*
5/cr	.62	*.51*	**13/mc**	.85	*.38*
6/mc	.92	*.28*	**14/cr**	.92	*.28*
7/mc	.92	*.28*	**15/cr**	1.0	*.00*
8/cr	.58	*.40*	**16mc**	.67	*.40*

Form B. The same statistics calculated for Form A were calculated for Form B (Table 6). The item means were larger for multiple-choice items except for one item-pair, where the mean for the constructed-response item was larger. One set of item-pairs had identical means. The mean test score on the 24 point test was 16.61 with a standard deviation of 4.23.

Table 6. Item Statistics - Form B

Item/Format	Mean	*Std. Dev.*	**Item/Format**	Mean	*Std. Dev.*
1/mc	.70	*.48*	**13/cr**	.54	*.52*
2/cr	.62	*.51*	**14/mc**	.92	*.28*
3/cr	.92	*.28*	**15/mc**	.92	*.28*
4/mc	.54	*.42*	**16/cr**	.49	*.33*
5/cr	.92	*.28*	**9/cr**	1.00	*.00*
6/mc	.85	*.38*	**10/cr**	.69	*.48*
7/mc	.77	*.39*	**11/cr**	.81	*.38*
8/cr	.85	*.38*	**12/mc**	.92	*.28*

Summary of Descriptive Statistics

Consistent findings were seen across both Forms A and B. On each form, the students generally found the multiple-choice items easier than the constructed-response items, with a few exceptions. The overall means for the multiple-choice items were greater than the means for the constructed-response items. However, the mean score for the multiple-choice items on Form B was slightly larger than the mean score for the multiple-choice items on Form A, .84 and .81 respectively. The converse finding was seen for the constructed-response items on Forms A and B, .69 and .66 respectively.

Receiving higher mean scores on the multiple-choice items was not surprising. First, students could check their answer against the options. If their answer was not listed as an option, they knew that they must have made a mistake in their solution strategy. The same checking procedure was not available with constructed-response items. Second, multiple-choice items facilitated educated guesses because students could eliminate options that they knew, with some level of certainty, were not the correct answers. The constructed-response items did not lend themselves to this type of guessing. Last, students could solve some multiple-choice items using the options. That is, they could use each of the options to inform and guide their solution strategy, because

they knew one option had to be the correct answer. The interviewers saw all of the possible explanations used during the think-alouds.

Cognitive Process Categories

Only one primary result was expected from the second and third steps of the analysis plan, namely, the discovery of the categories that exemplified the cognitive processes and the full-scale analysis of the protocol data using the categories. After using data analysis strategies from the grounded theory tradition, 28 categories emerged from the verbal data (see Figure 1). Detailed information about the development of the categories can be found in Chapter III, *Identify and Validate Cognitive Processes.*

The full-scale analysis also is described in Chapter III. Grounded theory blurs the line between methodology and analysis, which are often discernible with experimental design traditions. When analyzing the verbal data it became evident that the line between analysis and results also became indistinguishable. The constant comparative nature of grounded theory forced the blurring of the methodology, analysis, and results components of research. Because the results of the analysis emerged during the analysis itself, I avoided re-reporting the results here and refer the reader to Chapter III for details.

Cognitive Process Comparisons

The means reported above provided an index of performance on each item, across and within item-pairs. However, the index was not useful for examining cognitive processes. To get a general idea about the types of cognitive processes used by the test takers, I examined the frequency distributions of the cognitive process categories for each item in the eight item-pairs, per form. As it turned out, not all of the 28 categories were used often.

To get the most information from the descriptive analysis, I reduced the number of categories by excluding the ones that did not offer much information (i.e., the categories used infrequently). I first reviewed the original list of 28 categories. I then selected the categories that best referenced a significant cognitive move or that were needed for the students to have an understanding of algebraic patterns. This activity resulted in nine categories: A, D, E, F, G, S, U, Y, and AA (refer to Figure 1 for details). The categories also had to appear with some frequency to be analytically useful, which they did. I then determined the frequency at which each category appeared for all 33 items. The frequencies in Table 7 were grouped by the eight item-families. The grouping facilitated within-item format and across-item format comparisons, for each item family.

Table 7. Frequency Distribution of Abridged Category List

FAMILY 1: ARROWS & U-SHAPED

ARROWS: MULTIPLE CHOICE

Item #	A	D	E	F	G	S	U	Y	AA
1	6	1	4	0	0	1	0	0	0
9	5	3	2	2	0	0	1	0	0
FREQ.	11	4	6	2	0	1	1	0	0

ARROWS: CONSTRUCTED RESPONSE

Item #	A	D	E	F	G	S	U	Y	AA
1	6	3	3	0	0	0	1	0	0
9	7	2	4	0	0	0	0	0	0
FREQ.	13	5	7	0	0	0	1	0	0

U-SHAPED: MULTIPLE CHOICE

Item #	A	D	E	F	G	S	U	Y	AA
1	7	2	5	0	0	0	0	0	0
9	8	5	1	0	0	0	0	0	0
FREQ.	15	7	6	0	0	0	0	0	0

U-SHAPED: CONSTRUCTED RESPONSE

Item #	A	D	E	F	G	S	U	Y	AA
5	5	2	3	1	0	0	1	0	8
13	5	4	2	0	0	0	1	0	0
FREQ.	10	6	5	1	0	0	2	0	8

FAMILY 2: TACKS

TACKS TOP ONLY: MULTIPLE CHOICE

Item #	A	D	E	F	G	S	U	Y	AA
2	5	2	6	3	0	0	0	0	0
10	2	2	4	1	3	2	0	0	0
FREQ.	7	4	10	4	3	2	0	0	0

TACKS TOP ONLY: CONSTRUCTED RESPONSE

Item #	A	D	E	F	G	S	U	Y	AA
2	2	2	4	3	0	0	0	1	0
10	2	0	2	1	0	0	1	1	0
FREQ.	4	2	6	4	0	0	1	2	0

TACKS TOP & BOTTOM: MULTIPLE CHOICE

Item #	A	D	E	F	G	S	U	Y	AA
6	3	2	3	1	0	0	1	1	0
14	3	1	5	1	1	1	0	2	0
FREQ.	6	3	8	2	1	1	1	3	0

TACKS TOP & BOTTOM: CONSTRUCTED RESPONSE

Item #	A	D	E	F	G	S	U	Y	AA
2	1	4	3	1	2	0	2	1	0
10	5	1	5	3	1	0	0	0	0
FREQ.	6	5	8	4	3	0	2	1	0

FAMILY 3: EXTEND PATTERN OF NUMBERS

EXTEND NUMBER PATTERN (1,6,4,9): MULTIPLE CHOICE

Item #	A	D	E	F	G	S	U	Y	AA
7	5	1	4	2	0	0	1	3	0
15	5	3	3	0	0	1	1	0	0
FREQ.	10	4	7	2	0	1	2	3	0

EXTEND NUMBER PATTERN (1,6,4,9): CONSTRUCTED RESPONSE

Item #	A	D	E	F	G	S	U	Y	AA
7	5	2	3	1	0	2	1	1	0
11	6	2	5	1	0	0	2	0	0
FREQ.	11	4	8	2	0	2	3	1	0

EXTEND NUMBER PATTERN (4,3,7,6): MULTIPLE CHOICE

Item #	A	D	E	F	G	S	U	Y	AA
3	6	3	3	0	1	0	0	0	0
11	7	3	3	1	0	0	0	3	0
FREQ.	13	6	6	1	1	0	0	3	0

Table 7 (cont'd).

EXTEND NUMBER PATTERN (4,3,7,6):CONSTRUCTED RESPONSE

Item #	A	D	E	F	G	S	U	Y	AA
3	5	2	1	0	0	0	1	0	0
11	5	0	3	3	0	0	1	1	0
FREQ.	10	2	4	3	0	0	2	1	0

FAMILY 4: VERTEX

VERTEX DIAGONAL: MULTIPLE CHOICE

Item #	A	D	E	F	G	S	U	Y	AA
4	3	4	1	0	0	0	3	3	0
12	4	4	2	4	0	0	0	0	0
FREQ.	7	8	3	4	0	0	3	3	0

VERTEX DIAGONAL: CONSTRUCTED RESPONSE

Item #	A	D	E	F	G	S	U	Y	AA
4	5	4	2	2	1	1	0	2	0
12	5	4	1	0	0	0	0	4	0
FREQ.	10	8	3	2	1	1	0	6	0

VERTEX TRIANGLE: MULTIPLE CHOICE

Item #	A	D	E	F	G	S	U	Y	AA
4	5	2	5	1	1	1	1	6	0
12	7	1	5	2	2	0	0	5	0
FREQ.	12	3	10	3	3	1	1	11	0

VERTEX TRIANGLE: CONSTRUCTED RESPONSE

Item #	A	D	E	F	G	S	U	Y	AA
8	5	2	2	3	1	2	0	4	0
16	4	2	3	0	0	0	2	4	0
FREQ.	9	4	5	3	1	2	2	8	0

FAMILY 5: COLUMNS

COLUMNS 13: MULTIPLE CHOICE

Item #	A	D	E	F	G	S	U	Y	AA
5	4	2	3	1	1	0	2	2	1
13	4	3	3	1	0	0	2	5	0
FREQ.	8	5	6	2	1	0	4	7	1

COLUMNS 13:CONSTRUCTED RESPONSE

Item #	A	D	E	F	G	S	U	Y	AA
5	5	1	4	1	1	2	1	3	0
13	2	2	2	2	0	0	4	2	0
FREQ.	7	3	6	3	1	2	5	5	0

COLUMNS 14: MULTIPLE CHOICE

Item #	A	D	E	F	G	S	U	Y	AA
1	3	1	3	3	0	0	3	2	0
9	5	2	4	0	1	0	1	2	0
FREQ.	8	3	7	3	1	0	4	4	0

COLUMNS 14: CONSTRUCTED RESPONSE

Item #	A	D	E	F	G	S	U	Y	AA
5	3	4	2	1	0	0	2	2	0
13	6	1	7	3	1	1	3	3	1
FREQ.	9	5	9	4	1	1	5	5	1

FAMILY 6: PUPPY'S WEIGHT

PUPPY'S WEIGHT 21LBS: MULTIPLE CHOICE

Item #	A	D	E	F	G	S	U	Y	AA
6	5	1	4	1	0	1	1	3	0
14	7	0	8	1	0	0	0	3	0
FREQ.	12	1	12	2	0	1	1	6	0

PUPPY'S WEIGHT 21LBS: CONSTRUCTED RESPONSE

Item #	A	D	E	F	G	S	U	Y	AA
2	2	2	3	2	0	0	0	1	1
10	5	1	5	1	1	0	1	2	0
FREQ.	7	3	8	3	1	0	1	3	1

Table 7 (cont'd).

PUPPY'S WEIGHT 24LBS: MULTIPLE CHOICE

Item #	A	D	E	F	G	S	U	Y	AA
6	5	2	4	2	0	0	0	2	0
14	3	3	2	2	0	0	2	0	0
FREQ.	8	5	6	4	0	0	2	2	0

PUPPY'S WEIGHT 24LBS: CONSTRUCTED RESPONSE

Item #	A	D	E	F	G	S	U	Y	AA
6	7	0	6	1	1	0	0	3	0
14	5	4	3	1	0	0	0	3	0
FREQ.	12	4	9	2	1	0	0	6	0

FAMILY 7: PATTERN OF LETTERS

As & Bs: MULTIPLE CHOICE

Item #	A	D	E	F	G	S	U	Y	AA
7	6	4	2	0	0	0	0	0	0
15	6	1	4	2	0	1	1	0	1
FREQ.	12	5	6	2	0	1	1	0	1

As & Bs: CONSTRUCTED RESPONSE

Item #	A	D	E	F	G	S	U	Y	AA
3	4	1	4	1	0	0	1	0	0
11	4	2	3	2	0	0	1	0	0
FREQ.	8	3	7	3	0	0	2	0	0

Cs & Ds: MULTIPLE CHOICE

Item #	A	D	E	F	G	S	U	Y	AA
7	5	3	2	2	0	0	0	0	0
15	6	1	4	1	0	1	0	0	0
FREQ.	11	4	6	3	0	1	0	0	0

Cs & Ds: CONSTRUCTED RESPONSE

Item #	A	D	E	F	G	S	U	Y	AA
7	7	2	5	2	0	1	1	0	1
15	7	3	4	1	0	1	0	0	0
FREQ.	14	5	9	3	0	2	1	0	1

FAMILY 8: DOTS & STARS

DOTS: MULTIPLE CHOICE

Item #	A	D	E	F	G	S	U	Y	AA
8	1	3	1	1	2	1	2	0	0
16	5	1	4	1	2	2	0	1	0
FREQ.	6	4	5	2	4	3	2	1	0

DOTS: CONSTRUCTED RESPONSE

Item #	A	D	E	F	G	S	U	Y	AA
8	3	2	4	2	1	0	1	1	2
16	1	1	3	2	0	0	2	1	0
FREQ.	4	3	7	4	1	0	3	2	2

STARS: MULTIPLE CHOICE

Item #	A	D	E	F	G	S	U	Y	AA
4	3	2	1	3	0	0	0	2	2
12	4	1	5	3	0	1	1	0	1
FREQ.	7	3	6	6	0	1	1	2	3

STARS: CONSTRUCTED RESPONSE

Item #	A	D	E	F	G	S	U	Y	AA
8	2	0	3	2	0	0	3	1	0
16	3	1	5	0	2	0	2	0	1
FREQ.	5	1	8	2	2	0	5	1	1

DIAGONAL RECTANGLE

CONSTRUCTED RESPONSE

Item #	A	D	E	F	G	S	U	Y	AA
17	2	7	11	14	3	6	2	2	7

The results indicated that categories A ("overall pattern recognition") and E ("all information used from the question") appeared most often for all 33 items. Category Y appeared in the "polygon," "column," and "puppy weight" questions. These three item families present numbers in a columnar format. To solve the items, the students had to recognize the pattern of the independent and dependent variables (that is, the pattern within each column of numbers). For example several students solved the multiple-choice "column" item by noticing that the "B's are going up 4 and A's are going up 2." This item required students to solve the pattern represented by each variable so that they could continue the pattern despite the missing cells in the A and B columns. But, to solve the item, the students had to add on 4 three times in the B column. That is, they did not have to know how many numbers were missing in column A to know how many steps to extend the pattern in column B. In fact, for all three of the item-families, the students only had to recognize the pattern of the dependent variable to complete the pattern.

A similar type of approach was used to answer the "polygon" items. Here, the pattern extended several steps beyond the initial part of the pattern presented to the students in the item stem. The students used one of two approaches to solve the polygon items. Both solutions required students to understand that the pattern ended at 20. That is, they had to use the information given about the polygon and the number of triangles or diagonals to know where to stop extending the pattern. The students first recognized that the polygon variable consistently increased by one. They then saw that the triangle/diagonal variable also consistently increased by one. The solution path varied at this point. The students either continued to extend the variables by 1's or subtracted 2 from the polygon variable to get the number of diagonals represented in the other

variable. The following student quote exemplifies the latter solution strategy; "the pattern is going down by 3's because the diagonals are 3 less then the sides." Or the students continued the pattern by listing the polygon numbers to 20 and then they continued to list the diagonal/triangle pattern until that pattern ended at the same place as the polygon pattern. The students had to generalize the pattern beyond one step to successfully answer the polygon and column items.

Next, a more detailed analysis was conducted to examine the trends in the cognitive processes. In addition to applying the scheme across all items, I selected the categories that best represented the cognitive moves made by the students within an item family. The categories differed by families because the trends indicated that certain categories represented what might be called family-appropriate processes. The indices were labeled "family specific index" to indicate that the index should be interpreted in relation to the particular items within a family. To this point, most of the previous analyses were reported at the item-pair level within a form. For this analysis the indices were reported by item format within the eight item families to facilitate comparing the two item formats across the four items in an item family.

In addition to the family-specific indices, two additional indices were reported for every item-family: (a) a global cognitive process index that represented the mean proportion of the 9 categories, and (b) a common index that represented the most frequently used cognitive category seen across all items. The items' means were reported to serve as another comparison index. The aforementioned indices are illustrated in Tables 8 through 15. A sample interpretation of the indices are provided for the first item-family.

Table 8: Family 1: Arrows and U-shapes

			Global Index	**Common Index**	**Family Specific Index**	**Family Specific Index**
Family	Format	Mean	Overall cognitive process	Overall pattern recognition N (p)	Visual representation N (p)	Used relevant information in item * N (p)
Arrows	mc	100	.20	11 (.85)	2 (.14)	10 (.71)
U-shapes	mc	100	.18	15 (.88)	0	12 (.76)
MC mean		*100*	*.19*	*26 (.87)*	*2 (.06)*	*22 (.74)*
Arrows	cr	100	.20	13 (.77)	0	15 (.88)
U-shapes	cr	.92	.27	10 (.77)	1 (.08)	11 (.85)
CR mean		*.96*	*.24*	*23 (.77)*	*1 (.08)*	*26 (.87)*

* For all item-families, categories D and E were added together.

The items associated with Family 1 were the easiest of all 33 items. A high value for the mean index indicated that the item was easy (several students correctly answered the item). As seen, the items' means were very high, and therefore very easy. The overall cognitive process global index was slightly larger for the items presented in the constructed-response format, although the mean index was very similar. The u-shaped constructed-response item captured more of the overall cognitive process compared to the other three items. This means that of the students who responded to the item, they used more of the nine categories compared to the other items in the family. Very few students used a visual representation, as indicated by the very low value of the family specific index. And, a similar number of students across item formats recognized the pattern illustrated in the items. That is, the high index values (.77-.88) indicated that many students all used this cognitive process when solving the items. In general, the indices did not indicate variations in cognitive processes across the four items.

The protocol data showed that students most often saw that the figures rotated to the right or to the left, depending on the item. They then continued the rotation pattern to fill-in the missing figure. They often used the words "continuous pattern" during the interview.

The next family was referred to as "tacks". An illustration of overlapping pictures with either tacks on the top of them or tacks on the top and bottom of the pictures accompanied the item. The students were asked how many tacks it would take to hang 29 pictures. These items were more involved in terms of the cognitive load the students had to maintain because they not only had to find the pattern but then they had to extend it several steps (25 steps) beyond the illustration.

Table 9. Family 2: Tacks-Top only/Top-bottom

			Global Index	**General index**	**Family specific index**	**Family specific index**	**Family specific index**
Family	Format	Mean	Overall cognitive process	Overall pattern recognition N (p)	Visual representation N (p)	Applicable equation N (p)	Used relevant information N (p)
Top	mc	.70	.20	7 (.41)	4 (.24)	3 (.17)	14 (.82)
Top-bottom	mc	.92	.25	6 (.46)	2 (.15)	5 (.38)	11 (.85)
MC mean		*.81*	*.22*	*13 (.43)*	*6 (.20)*	*8 (.27)*	*25 (83)*
Top	cr	.62	.21	4 (.31)	4 (.31)	0	8 (.62)
Top-bottom	cr	.62	.19	6 (.35)	4 (.24)	3 (.17)	13 (.77)
CR mean		*.62*	*.20*	*10 (.33)*	*8 (.28)*	*3 (.10)*	*21 (.71)*

As with the first family, these data were more remarkable for similarities than differences in the global index; however, there was some differentiation among family specific indices for the four items. The multiple-choice items appeared to elicit better use of relevant information and appropriate equations, as indicated by the larger index values compared to the constructed-response items. The constructed-response versions

prompted students to construct visual representations a little more often than the multiple-choice variations. However, the qualitative analysis showed that the visual representations did not help the students.

The items in Family 3 could be described as completing a pattern by filling-in the last two numbers in a series of numbers. As part of the answer, the students had to provide the rule they used to solve the pattern. Their responses exhibited almost identical indices for both item difficulty and overall cognitive processing responses. The constructed-response variations led students to rely more often on visuals representations (i.e., write pattern of how numbers increased and decreased across the pattern) and the multiple-choice versions apparently led students to use information provided in the items to solve them. Perhaps a reliance on visuals dampened the need to attend to all of the information offered in the item.

Examination of the protocol data showed a slight difference in the way the students solved the items, although the cognitive process appeared to be similar, as exemplified in the general index. Two patterns were inherent in these items. The students found the alternate rule (every other number increased by 3) equally often in both item format presentations, even though the multiple-choice options did not reflect the alternate pattern.[1] Essentially, the qualitative analysis supported the general index

---

[1] The original item was a 2-point constructed-response item. The counter part multiple-choice versions were written to reflect the 2-point characteristic of the original item (according to the scoring rubric). To accomplish this, the part of the item that addressed the rule mirrored the answer on the rubric. The disconnection between the multiple-choice options and the alternate pattern perplexed some students to the point of not being able to see the rule in the correct answer choice. This happened rarely as indicated both by the protocol data and the percentage correct index.

because the same cognitive process was used to solve the items even though students were able to solve them using one of two valid solution paths.

Table 10: Family 3. Extend Pattern of Numbers

			Global Index	**General index**	**Family specific index**	**Family specific index**
Family	Format	Mean	Overall cognitive process	Overall pattern recognition N (p)	Visual representation N (p)	Used relevant information N (p)
Ptrn 1	mc	.77	.23	10 (.71)	2 (.14)	11 (.39)
Ptrn 2	mc	.92	.20	13 (.81)	1 (.06)	12 (.38)
MC mean		*.85*	*.21*	*23 (.76)*	*3 (.10)*	*23 (.38)*
Ptrn 1	cr	.85	.20	11 (.65)	2 (.12)	12 (.35)
Ptrn 2	cr	.81	.19	10 (.77)	3 (.23)	6 (.23)
CR mean		*.83*	*.20*	*21 (.70)*	*5 (.17)*	*18 (.32)*

The results of the analysis for the Family 4 items (Vertex Diagonal and Vertex Triangle) mirrored the results of Family 3. All of the cognitive process indices were remarkably similar, each differing by only a few percentage points.

Table 11. Family 4. Vertex-Diagonal, Vertex-Triangle

			Global Index	**General Index**	**Family specific index**	**Family specific index**	**Family specific index**
Family	Format	Mean	Overall cognitive Process	Overall pattern recognition N (p)	Visual representation N (p)	Relationship between numbers recognized N (p)	Used relevant information N (p)
Diagonal	mc	.92	.11	7 (.61)	4 (.19)	3 (.21)	3 (.10)
Triangle	mc	.85	.29	12 (.71)	3 (.18)	11 (.65)	10 (.26)
MC mean		*.89*	*.21*	*19 (.67)*	*7 (.19)*	*14 (.56)*	*13 (.19)*
Diagonal	cr	.69	.12	10 (.70)	2 (.13)	6 (.55)	3 (.09)
Triangle	cr	.85	.29	9 (.69)	3 (.23)	8 (.62)	5 (.12)
CR mean		*.77*	*.20*	*19 (.69)*	*5 (.17)*	*14 (.59)*	*8 (.11)*

The mean index associated with Family 5 suggested that the constructed-response format depressed performance (.58 versus .78), which may have signaled significant differential cognitive processes. But comparison of the other indices indicated relatively similar patterns of cognitive processes at the global and family specific levels. Slight differences were observed between the visual representation and the overall pattern recognition indices. This suggested that more students used a visual approach (students wrote the pattern they observed in each column) to solve the constructed-response items compared to the multiple-choice items. The students may have used the visual approach because they were unable to recognize the overall pattern or because the visual approach negated the need to recognize the overall pattern.

The results for Families 6 (puppy's weight) and 7 (pattern of letters), were notably flat in many of the indices, especially the global and general indices, reflecting the consistent pattern of results found for earlier families, most notably 1, 2, and 3. Slight differences were observed in the visual representation index again.

Table 12. Family 5. Columns

			Global Index	**General Index**	**Family specific index**	**Family specific index**	**Family specific index**
Family	Format	Mean	Overall cognitive process	Overall pattern recognition N (p)	Visual representation N (p)	Relationship between #'s recognized N (p)	Used relevant information N (p)
Columns,13	mc	.85	.24	8 (.50)	2 (.13)	7 (.44)	11 (.34)
Columns,14	mc	.70	.27	8 (.64)	3 (.23)	4 (.29)	11 (.39)
MC mean		*.78*	*.26*	*16 (.57)*	*5 (.19)*	*11 (.37)*	*22 (.37)*
Columns,13	cr	.54	.25	7 (.50)	3 (.21)	5 (.36)	9 (.32)
Columns,14	cr	.62	.26	9 (.53)	4 (.24)	5 (.29)	14 (.41)
CR mean		*.58*	*.26*	*16 (.52)*	*7 (.23)*	*10 (.33)*	*23 (.40)*

Table 13. Family 6. Puppy's Weight

			Global Index	**General Index**	**Family specific index**	**Family specific index**	**Family specific index**
Family	Format	Mean	Overall cognitive process	Overall pattern recognition N (p)	Visual representation N (p)	Relationship between #'s recognized N (p)	Used relevant information N (p)
Weight 21	mc	.92	.23	12 (.7)	2 (.12)	6 (.35)	13 (38)
Weight 24	mc	.92	.21	8 (.57)	4 (.29)	2 (.14)	11 (.39)
MC mean		*.92*	*.22*	*20 (.64)*	*6 (.20)*	*8 (.26)*	*24 (.38)*
Weight 21	cr	.62	.14	7 (.55)	3 (.15)	3 (.15)	8 (.43)
Weight 24	cr	.92	.22	12 (.71)	2 (.12)	6 (.35)	13 (.38)
CR mean		*.77*	*.19*	*19 (.64)*	*5 (.14)*	*9 (.26)*	*21 (.42)*

Table 14. Family 7. Pattern of Letters

			Global Index	**General Index**	**Family specific index**	**Family specific index**
Family	Format	Mean	Overall cognitive process	Overall pattern recognition N (p)	Visual representation N (p)	Used relevant information N (p)
As & Bs	mc	.92	.20	12 (.71)	2 (.12)	11 (.32)
Cs & Ds	mc	.92	.20	11 (.79)	3 (.21)	10 (.36)
MC mean		*.92*	*.20*	*23 (.75)*	*5 (.16)*	*21(.34)*
As & Bs	cr	.92	.20	8 (.62)	3 (.23)	10 (.38)
Cs & Ds	cr	1.00	.23	14 (.82)	3 (.18)	14 (.41)
CR mean		*.96*	*.22*	*22 (.73)*	*6 (.20)*	*24 (.40)*

A different sequence of patterns became apparent in Family 8. First, the mean index stands apart from the previous indices because the four items in this family were much more difficult than the other items, but the two formats within the family continue to be canonical.

Second, the constructed-response version, despite its similar mean difficulty, prompted a somewhat different global cognitive process (.21 versus .15). The difference appeared to come primarily from the "used relevant information" index, which represented a difference in the number of students using relevant information provided in the item. The general index complemented the family specific index because students would find it difficult to recognize the overall pattern if they were unable to use the relevant information provided in the item, and vice versa.

This item family was the most challenging of the entire set of eight item families, from writing comparable items to the students' apparent difficulty at answering the items. Of the eight item families, it most closely represented a performance item, in that the items had multiple solution paths and required the students to not only show their mathematical work but offer a written explanation about their solution strategy. The items required either the facile use of an equation (which only seven students managed to use, and then not always correctly) or some careful logical reasoning paired with sequential problem representation.

Table 15. Family 8. Dots and Stars

			Global Index	**General Index**	**Family specific index**	**Family specific index**	**Family specific index**
Family	Format	Mean	Overall cognitive process	Overall pattern recognition N (p)	Visual representation N (p)	Applicable equation N (p)	Used relevant information N (p)
Dots	mc	.58	.23	6 (.46)	2 (.15)	4 (.31)	9 (.35)
Stars	mc	.54	.08	7 (.48)	6 (.45)	0 (0)	6 (.28)
MC mean		*.56*	*.15*	*13 (.48)*	*8 (.31)*	*4 (.13)*	*14 (.33)*
Dots	cr	.49	.19	4 (.27)	4 (.27)	1 (.07)	10 (.33)
Stars	cr	.67	.23	5 (.42)	2 (.17)	2 (.17)	9 (.38)
CR mean		*.58*	*.21*	*9 (.34)*	*6 (.23)*	*3 (.11)*	*19 (.35)*

The data from all eight families were summarized (see Table 16) for the global and general indices, overall cognitive process and overall pattern recognition, respectively. Again, the data were more remarkable for their similarities rather than their differences. Recognizing that some variation existed among item families, the overall trends, both within and between families, pointed to a conclusion that similar cognitive processes were used by students as they solved both item formats.

Table 16. Overall Patterns for Families 1 through 8

		Global Index	**General Index**
Format	Mean	Overall cognitive process	Overall pattern recognition
MC mean	*.78*	*.21*	*.60*
CR mean	*.75*	*.21*	*.56*

To examine the robustness of the conclusion, the data associated with the Balanced Assessment item were summarized (see Table 17) and subsequently compared to the previous results. This item required students to extend the given pattern, which

was contextualized in a diagonal rectangle problem. There were five separate scaffold pattern extensions.

Table 17: Diagonal Rectangle Item

			Global Index	**General Index**	**Family specific index**	**Family specific index**	**Family specific index**
Family	Format	Mean	Overall cognitive process	Overall pattern recognition N (p)	Visual representation N (p)	Grapples with information in item N (p)	Used relevant information N (p)
Diagonal rectangle	cr	.23	.26	2 (.09)	14 (.61)	7 (.30)	18 (.39)

The first index that stood apart from the previous ones was the global index, but the finding was not overly striking when compared to the other 32 items. For instance, the global index was only slightly larger than the summarized index in Table 16. It was comparatively larger across each individual item family, except for item Family 5 (columns). As indicated by the low index, very few students recognized the overall pattern, a finding that complements the mean index, despite their attempts at representing the pattern visually. The applicable equation was difficult for the students to figure out, so they apparently relied on using visual approaches to find the pattern.

A new family specific index emerged to examine trends, which was used so infrequently with the other items that it did not emerge as a family specific index. As seen, seven students struggled with the item, manipulating the given information every possible way, resolute to successfully extend the pattern. The fact that 30% of the students who attempted to answer the item were observed using the strategy primarily

with this item suggested that it may have elicited some different cognitive processes, even though, as the qualitative data suggested, most of them were perplexed by the item.

A close look at the qualitative records for all the students who attempted this item revealed the majority of them were highly engaged cognitively. One possibility, supported in the analysis, was that they could not rely on "easier" methods such as visually representing the diagonal rectangle and/or counting the dots (a cognitive process often used in the stars-dots items). Most of the students gave an answer. They did not guess blindly or give up entirely. This observation was significant because this item came at the end of the think-aloud protocol, a point where students were more likely to be fatigued. Further, it seemed that many of the students had never encountered a problem like this one; they could have easily become frustrated with the item and became resolute with their failed attempt.

The protocol evidence supported the assertion that for the performance item, the students used different cognitive processes for the multiple-choice and constructed-response items.

Two trends emerge from the interviews with students completing the Balanced Assessment item. First, the number of students using formulas, even if they were wildly incorrect, increased significantly from part A to part D. In part A, only 1 student used any formula, but eight students were using formulas by part D. Though unsuccessful, students were actively trying to solve the problem by looking for a pattern and creating an equation that accurately represented it. Most of the students used the area equation to solve the problem, but other students tried to look for equations from the patterns that they saw in the earlier parts of the item. Charlie's response illustrates this:

"I am trying to find a pattern...in a 4x5 rectangle you make each number go up by 1, and it was 18 higher in the 5x6, so then the 10x11 would be...well I could do this: 10 (in the 10x11)-5(in the 5x6)=5 so then the pattern is going up by 5. So 18x5=90, and 90+50 (number of dots in 5x6)= 140."

Many students used this approach to find a pattern and an equation.

Second, students attempted several different mathematical processes to solve the problem. In general, students drew upon the strategies they knew from previous experiences with solving mathematical items and tried many of them before deciding upon an approach. This was particularly true of students who had used visual representations of the rectangles to solve earlier problems but found that they could not continue to use this strategy. Allison's transcript illustrates this point.

Allison drew the 5x4 and the 5x6 rectangle and counted the dots inside. On the 10x11 problem, she commented, "I need a way to get it so I don't have to count." She attempted to draw the 10x11 rectangle on the worksheet, but realized that it will not fit. She said she did not know how to do the item, but continued to struggle with various solution strategies to find a way to solve it. She looked at the 5x6 rectangle and said, "Well this 10 by 11 is double 5x6" and began to double the answer for the 5x6 to solve the problem. Then she realized that 6 is not double 11, and looked at the problem again. She then commented, "This is really hard" and began to look for patterns of numbers in the 4x5 and the 5x6 items to get the answer (i.e. "5x6 is 4x5+1x1). She commented, "That's not working, so I need to try something else." After a few other attempts, she finally decided to calculate the area for her final answer (10x11=110).

Students like Allison explicitly demonstrated the cognitive complexity of the item and the various cognitive processes they used in their attempt to solve the item. They commented that the problem was difficult, they did not know what they were doing, and they thought their answers were wrong. Despite these challenges, they continued to work to solve the problem. I called this "playing with the numbers" or "grappling" (category AA) because students went in with a pretty wide lens of possibility solution strategies. They generally tried different mathematical computational strategies or looked for patterns inside of the given numbers. For example, Allison revisited the 5x4 and 5x6 rectangles to look for patterns within the numbers by 'playing with them.

This kind of "playing" was not observed to the same extent in the multiple-choice or constructed-response items. Generally, the students either knew the answers to the items or they solved the items quickly, without having to search their mathematical toolbox for multiple solution strategies or ways of reasoning or making sense of the information. Consequently, students were less likely to "play" or grapple with the information provided in the multiple-choice and constructed-response items, or to think meta-cognitively about what they were doing as they solved the items.

Summary of Cognitive Process Comparisons

The findings from the descriptive analyses and the two analyses that used the subset of cognitive process categories did not, in general, reveal significantly different cognitive processing between the multiple-choice and constructed-response item formats, within or across the eight item families. The global index, which represented the nine cognitive process categories that represented the significant cognitive moves on the part of the students, did not elicit substantial findings when looked within the item families

and in the aggregate. The most interesting findings were found with the family specific indices, most likely because the aggregate indices masked the cognitive processes differences, albeit small, that were elicited by the items.

Many of the family specific indices were comparable between item formats within an item family but a few inconsistent findings were observed. Most notably, the students used visual representation more often to solve the constructed-response items (five of the eight item families plus the Balanced Assessment item) than the multiple-choice items. This was evident in the higher indices associated with the constructed-response items compared to the indices associated with the multiple-choice items within an item family.

In some item families, especially for those prone to visual solution paths, the visual index interacted with the overall pattern recognition index. Apparently, students had to recognize the entire pattern to employ a visual solution strategy. But, it should be noted that although a visual solution strategy worked with the items in this study, it may not work as well with pattern items that are more complex and cognitively challenging.

Although the protocol data suggested some differences in the cognitive processing used to answer the multiple-choice and constructed-response items, the differences were small in terms of the relative differences in proportions observed within an index in any one-item family. For example, the largest difference in proportions in the visual representation index was .08, which represented two additional students using the cognitive process (see item families 2 and 8).

The most remarkable difference was seen in the Balanced Assessment performance item. A unique cognitive process (category AA), not seen in the other

items, was elicited when the students solved the performance item. The item challenged the students to grapple with the information in a way that was not seen in the other 32 items. The items that came closest to a similar cognitive struggle were the stars-dots items. But, the particular cognitive process was used less frequently compared to the other eight categories so category AA was not reported as a family specific index (see Table 7, to examine the frequency distribution of the nine categories).

External Post hoc Evaluation

The purposes of the expert panel review were to evaluate whether I had accurately and sufficiently identified the cognitive processes expressed by the students and to evaluate whether the analysis procedures were appropriately conducted. The review occurred after the student think-alouds and data analyses were completed. Nonetheless, evaluating the cognitive processes that emerged and the data analysis procedures from which they emerged, even in a post hoc review, would strengthen the results of the study. The results of the evaluation were generally positive. The next section details the germane findings from the evaluation — both the post hoc review and the think-aloud interview with the 8th grade teacher.

Results of the Evaluation

After the panel of four mathematics experts met for a one-day information sharing and training session, they individually evaluated the validity of the 28 cognitive process categories and the coding system by listening to two think-aloud audiotapes. I divided the panel into two groups and each group listened to the same think-aloud interview. One 8th grade teacher and one mathematics educator were assigned to each group. The panelists reconvened after their review and each member had the opportunity to share her

impressions. The panelists' comments fell within three general categories. Specific examples follow each general comment. As seen, their suggestions would clarify the meaning of the original cognitive process categories.

Change the wording of some categories to better convey their meaning.

- Reword Category G: Uses applicable formula (where formula means a symbolic representation, e.g., length x width), rather than generating a heuristic.
- Reword Category F: Re-presentation of information in question, e.g., tables, lists.
- Reword Category R: Work not shown.
- Reword Category Y: Relationship between independent (single) and dependent (multiple) variables recognized.
- Reword Category Z: Relationship between independent (single) and dependent (multiple) variables not recognized.
- Reword Category AA: Persists, even when uncertain of solution path.
- Reword Category M: Get right answer for wrong reason; uses some irrelevant information to get answer right (e.g., selected longest answer).
- Reword Category X: Mis-states math terms; mismatch between verbal explanation and actual use of math terms.

Create new categories to account for some unidentified cognitive processes.

- Uses existing knowledge from previous experiences (e.g., classroom instruction).
- Monitors understanding of item.

- Uses real world knowledge or practical reasoning to solve item.
- Applies appropriate arithmetic operations (+, -, x, \).
- Applies inappropriate arithmetic operations (+, -, x, \).
- Uses deep cognitive processing or engagement: Student shifts strategy; has a repertoire of tools available to solve pattern.

Retain all original 28 categories (but reword as indicated above).

After the panelists shared their evaluation of the cognitive process categories, they were asked to comment on their experience of coding the students' think-alouds. (Time constraints prevented one of the reviewers from completing this exercise.) The purpose of the activity was to expose potential problems with the original coding procedure. I was not interested in assessing inter-rater reliability. The panel's general observations follow.

1. There seemed to be an inverse relationship between students mathematical ability and their depth of engagement when solving the items.
2. The students' low reading level seemed to prevent them from understanding the items. The panelists were uncertain about whether the students misunderstood the meaning of certain words, whether they accidentally misread certain words, or whether they could not read an item at all.
3. There were qualitative and quantitative differences between what students verbally reported and what they wrote as answers to the constructed-response items.
4. The coding procedures were appropriate and easy to apply.

5. The audiotapes were essential to code the cognitive processes; that is, the transcripts and test booklets alone were insufficient information to accurately code the processes.

As seen, most of their comments represented observations about the students' behaviors rather than problems uncovered with the coding procedures. In fact, the last two observations indicate that the panelists found the procedures sound.

Because the panel listened to two of the 34 student think-aloud interviews, I assessed the panel's observations by discussing them with the other think-aloud interviewers — I would have been remiss to assume that their observations, from two tapes, reflected all of the qualitative data collected for this study. We concluded that the first point noted above was not universally observed. For example, some of the students enrolled in an algebra course and/or who had strong mathematical ability found the diagonal rectangle item difficult, yet they were deeply engaged when solving the item. Their engagement was evident in their persistence and struggle to solve the item. Conversely, some of the students enrolled in non-algebra classes and/or who appeared to find most of the items challenging solved some items with little engagement.

Scant evidence existed to confidently refute or support the second point, except for the last notion, which questioned the students' literacy. The interviewers asked all of the students to read every item aloud. None of the interviewers encountered illiterate students.

The think-aloud interviewers also observed the third point listed above, even though it was not formally documented during the think-aloud interviews. I did not

disregard this observation when I encountered it during the think-alouds because of lack of interest but rather because it was not the focus of the analysis.

As mentioned early, because of limited resources, the panel only reviewed two tapes when they evaluated the categories and the coding system. Because of this limitation, one must consider the conclusions from the post hoc review with some caution. For instance, the two students did not use all 28 cognitive processes when solving the items; therefore, the panelists were unable to conduct a full review of the categories based on verbal evidence.

However, in the absence of complete verbal data, the panelists relied on their knowledge of 8th grade students and on 8th grade mathematics curriculum and instruction. Their experiences allowed them to evaluate the categories for which no direct think-aloud evidence existed on the two tapes available to them. Based on the direct verbal evidence and on their experience, the panel ultimately concluded that the original categories seemed appropriate and represented the cognitive processes students used when solving the pattern items.

Results of the Teacher Think-aloud Interview

The think-aloud procedure used with the students was duplicated with an 8th grade mathematics teacher. She responded to all 18 items on Form A. The teacher was first asked to solve the items from an 8th grader's perspective and then to determine whether there were alternate solution strategies that may have been foreign to the students. Three results emerged from this activity and suggested that: (a) the teacher used similar cognitive processes when solving most of multiple-choice and constructed-response items, (b) the teacher and the students often used the same verbal explanations and

cognitive moves when solving many of the items, and (c) the students saw alternative ways to solve a few of the items but the teacher did not.

The information in Table 18 represents the cognitive moves made by the teacher for each of the 17 items. It was used to base the first and second conclusions. Evidence to support the third conclusion was presented in narrative form using excerpts from the teacher's verbal protocol when appropriate.

Table 18. Teacher Response Distribution for Abridged Category List

FAMILY 1: ARROWS & U-SHAPED

ARROWS: CONSTRUCTED RESPONSE

Item #	A	D	E	F	G	S	U	Y	AA
9	1		1						

U-SHAPED: MULTIPLE CHOICE

Item #	A	D	E	F	G	S	U	Y	AA
1	1		1						

FAMILY 2: TACKS

TACKS TOP ONLY: MULTIPLE CHOICE

Item #	A	D	E	F	G	S	U	Y	AA
10	1		1		1				

TACKS TOP & BOTTOM: CONSTRUCTED RESPONSE

Item #	A	D	E	F	G	S	U	Y	AA
2	1		1		1				

FAMILY 3: EXTEND PATTERN OF NUMBERS

EXTEND NUMBER PATTERN (1,6,4,9): CONSTRUCTED RESPONSE

Item #	A	D	E	F	G	S	U	Y	AA
3	1		1	1				1	

EXTEND NUMBER PATTERN (4,3,7,6): MULTIPLE CHOICE

Item #	A	D	E	F	G	S	U	Y	AA
11	1		1	1				1	

FAMILY 4: VERTEX

VERTEX DIAGONAL: CONSTRUCTED RESPONSE

Item #	A	D	E	F	G	S	U	Y	AA
12	1		1		1			1	

VERTEX TRIANGLE: MULTIPLE CHOICE

Item #	A	D	E	F	G	S	U	Y	AA
4	1		1		1			1	

FAMILY 5: COLUMNS

COLUMNS 13: MULTIPLE CHOICE

Item #	A	D	E	F	G	S	U	Y	AA
13	1		1		1			1	

COLUMNS 14: CONSTRUCTED RESPONSE

Item #	A	D	E	F	G	S	U	Y	AA
5	1		1		1	1		1	

FAMILY 6: PUPPY'S WEIGHT

PUPPY'S WEIGHT 21LBS: MULTIPLE CHOICE

Item #	A	D	E	F	G	S	U	Y	AA
6	1		1	1				1	

PUPPY'S WEIGHT 24LBS: CONSTRUCTED RESPONSE

Item #	A	D	E	F	G	S	U	Y	AA
14	1		1	1				1	

FAMILY 7: PATTERN OF LETTERS

As & Bs: MULTIPLE CHOICE

Item #	A	D	E	F	G	S	U	Y	AA
7	1		1	1		1			

Cs & Ds: CONSTRUCTED RESPONSE

Item #	A	D	E	F	G	S	U	Y	AA
15	1		1	1		1			

FAMILY 8: DOTS & STARS

DOTS: MULTIPLE CHOICE

Item #	A	D	E	F	G	S	U	Y	AA
16	1		1	1				1	

STARS: CONSTRUCTED RESPONSE

Item #	A	D	E	F	G	S	U	Y	AA
8	1		1	1	1	1		1	1

DIAGONAL RECTANGLE

CONSTRUCTED RESPONSE

Item #	A	D	E	F	G	S	U	Y	AA
17	1		1	1	1	1			1

The first result was based on an examination of the multiple-choice and constructed-response items within a family. As seen in Table 18, the teacher used the same categories to solve the multiple-choice and constructed-response items across six of the eight families (excluding the diagonal rectangle item because it did not have a multiple-choice counterpart). For Family 5 and 8 she used additional cognitive processes to solve the constructed-response items compared to the multiple-choice items. Specifically, in Family 5 she checked her work (category S) while solving the constructed-response item but she did not check her work while solving the multiple-choice item. Three additional cognitive processes emerged while the teacher solved the constructed-response item in Family 8, namely applicable equation (category G), checks work (category S), and grapples with information (category AA).

I am not sure why she checked her work when solving the constructed-response item in Family 5 except to hypothesize that she did not have the response options available to confirm her own solution, as she did for the multiple-choice item within the family. The same reason could be offered for category S emerging for the constructed-response item in Family 8. The other two categories may have emerged because of the ordering of the items in the booklet and/or because of bias I may have introduced during the interview.

This conjecture is based on the following facts. In Form A, the constructed-response, stars item appeared before the multiple-choice, dots item. Compared to the other items, the teacher struggled to find a solution (category AA) for the stars item as the item was more complex and was therefore more difficult. There were two general strategies for solving the item, either using a visual approach by noticing the relationship

between the number of stars in the columns and rows or by finding the algebraic equation that represented the pattern. The teacher relied solely on her knowledge that an algebraic equation (category G) was at the root of every pattern, and thereby applied this knowledge to the star item. She had difficulty finding the applicable equation but eventually solved the item.

Interviewer bias likely occurred when we discussed the various ways that the students could have solved the item. During the conversation I shared with her that some students used a visual strategy and described it to her. The teacher commented that the visual approach was easier than an equation approach. When she encountered the dots item she used the visual approach and easily solved the item – hence the omission of categories G, S, and AA for the multiple-choice version of the item. If the conjecture is correct, then the difference observed between the stars and dots items was spurious. If it is wrong, then the teacher likely learned from her first encounter with the stars item and applied her learning to solving the dots item.

The second result — the teacher and the students often used the same verbal explanations and cognitive moves when solving many of the items — was based on the data from the teacher's and students' (see Table 7) abridged category systems. I observed that the teacher and students often used the same cognitive moves when solving the 16 items. A few differences were identified in the data. The most obvious difference occurred in categories D (partial information used from question) and U (partial pattern recognition). For every item, at least one student, and often three or four students, received a D and for the majority of the items at least one student received a U. These two categories were not assigned to the teacher. Apparently, the teacher always

recognized the entire pattern (e.g., she never assumed that the pattern identified in the beginning of a series was the same pattern at the end of the series) and used all of the information available in the item to solve the item (e.g., she never jumped from the beginning of the item to immediately solving the item, or ignored information in the middle or at the end of the series of information in the pattern).

The teacher's cognitive moves also differed for categories F (visual representation) and G (applicable equation used). She often used the two strategies, moving back-and-forth between the two, whereas the students often relied on only one of the two cognitive processes when solving an item. The use of multiple cognitive moves within an item may be an indication of the teacher's knowledge that there are often several ways to solve an algebraic pattern item.

I also observed that the teacher relied on formulas much more often than did the students. She first attempted to find a formula that fit the observed pattern, regardless if a visual approach would have been easier. For example, while solving the stars item, she first saw that each step had one more row and column compared to the previous step. She then noticed that the difference between steps 1 and 2 was 5 and steps 2 and 3 was 7, and thereby noticed that the pattern between steps was an increase of two. She verified her observations by applying this knowledge to the fourth step. Up to this point, she and most of the students used similar cognitive moves. However, her knowledge of mathematical rules and relationships (e.g., she knew a quadratic formula was needed to represent the pattern) led her down a path to find a formula that fit her observed pattern, whereas most of the students used some type of visual approach (e.g., listed the numbers in the pattern all the way to the 16^{th} step). She maneuvered through several formulaic

strategies, changing strategies several times before completing the item. She perseverated on finding a formula to the point of her not seeing the easier visual approach. When I explained the "area" method (rows x columns) to her she responded "This is an easy way, wow. Students have an advantage if they see the item like this because it is easier to solve, rather then looking for a formula."

As mentioned earlier, some of the students used alternative solution strategies when solving some of the items that the teacher did not notice until I drew her attention to them. This happened for three item-pairs: columns, extend pattern of numbers, and dots & stars. The first two item-pairs had multiple patterns but the students used the same cognitive moves to solve the items. The last item-pair had multiple solution strategies, which required different cognitive moves.

The columns items required students to complete a pattern by identifying the pattern in each of two columns (independent and dependent variables), continuing the pattern across a gap in the pattern, and then completing the pattern for the dependent variable. Most of the students and the teacher solved the items the same way; they applied the cognitive moves just listed. However, a few students noticed the relationship *between* the two numbers in columns A and B, rather than the difference between consecutive numbers *within* a column. After the teacher solved the item I asked her if there were alternative ways to solve it. She said no. I then shared with her that a few students discovered an alternative pattern. Upon hearing this she re-examined the item but still could not see the second pattern. I explained that there was a relationship between the numbers across the two columns; but she had a difficult time seeing the

alternate pattern. She did not see the alternate pattern until I pointed to the numbers and said it aloud.

A similar conversation occurred between the teacher and I when she was solving the extend pattern of numbers item-pair. These items required students to fill-in the last two numbers of a series eight numbers that followed a certain pattern. Most of the students and the teacher solved the two items in the same way, that is, by recognizing the pattern between consecutive numbers. However, some students noticed that a difference of three existed between every other number. After the teacher completed the items, I told her that some students noticed a second pattern. The second pattern was not apparent to her until I said it aloud. She explained that some pattern items have more than one pattern within them, which makes working with pattern items difficult of students.

The dots and stars items had multiple solution strategies inherent in them. The teacher immediately looked for a formula to solve the items whereas most of the students used some kind of visual, or re-presentation of information approach. Most of the students relied on listing a series of numbers from the 4^{th} step to the 16^{th} (or 20^{th} step, depending on the item), using a "counting-on" approach to figure out the answer. A few students were able to see a very straightforward way to solve the item, which required the least number of cognitive moves compared to the other solution strategies. These students recognized that the number of dots (stars) in a row multiplied by the number of rows equaled the number of dots (stars) in the block of dots (stars). They also noted that the number of rows and the step number were the same number. They solved the item by using these two pieces of information. That is, they saw how the item grew

geometrically. This approach became known as the "area" method. The area method was not apparent to the teacher until I explained it aloud. She immediately recognized the simplicity of this approach and commented that students who can see things visually often have an easier time solving pattern items.

I do not have a good explanation for the above results except that they may have something to do with the richness and depth of the teacher's mathematical knowledge, which ultimately led her to find the underlying mathematical formula of the observed pattern. Conversely, the students limited exposure to patterns (in terms of their level in school and variations of patterns) may have hindered their understanding of what really composed a pattern and therefore resulted in them using visual or other familiar mathematical approaches (e.g., counting-on strategy) that they had learned up to the 8th grade. The teacher could not proffer additional insights except to explain that patterns are difficult items to work with because one really does not know the true pattern within an item. That is, a series of five numbers could be presented in an item with a pattern up plus one, minus two, plus one, minus two. This pattern may match the five presented numbers; however, the sixth and subsequent numbers, which are unknown to the respondent, could result in a second viable pattern for that series of numbers.

An unexpected and unintended result occurred during the teacher's think-aloud session. As mentioned earlier, after the teacher solved an item, I shared with her some of the alternate ways that the students solved the item. Because of the teacher's classroom experience, she was able to see the item and solution strategy through the lens of a student. She took notes during the think-aloud session that connected their solution strategies with her instructional approaches. She seemed quite engaged and excited about

her learning and commented about how useful think-alouds could be to inform instruction.

Summary of External Post hoc Evaluation

The main results of the post hoc evaluation showed

- That the teacher and students often used the same cognitive processes to solve the pattern items,
- The category system sufficiently captured the major cognitive processes used by the students when solving the items,
- The need to re-word some of the categories to make them more understandable to the mathematics community,
- That a teacher and students can see different patterns in an item and both patterns can be correct,
- That a teacher is unable to see a pattern that some students can see,
- The importance of including content experts at the beginning of this type of research, and
- The usefulness of the think-aloud procedure to inform instruction.

Overall Summary of Results

This study examined whether students used different cognitive processes to answer multiple-choice and constructed-response items. Every analysis led to the same conclusion; that is, regardless of item format, students tended to use the same cognitive processes when solving the algebraic pattern items. One exception arose when I compared the cognitive processes associated with the Balanced Assessment item with those of the other 32 items. In this one situation, students verbalized a few different

cognitive processes (e.g., category AA) and they appeared much more engaged (e.g., relentless pursuit of a solution) when solving the item. This finding suggested that if test developers want students to be deeply engaged with an item and foster their use of problem solving processes, then the items have to be intentionally written in a way that elicits such thinking. The Balanced Assessment item was developed to elicit these behaviors; however, it also required more time to solve and to score compared to the other items in this study.

The multiple-choice and constructed-response items in the dots and stars item-family came closest to eliciting the level of engagement and types of cognitive processes observed in the Balanced Assessment item. Each of these items had a few common characteristics, which perhaps contributed to the common results:

- They allowed for multiple solution paths,
- They required several minutes to solve,
- The problem space required students to extend the pattern several steps beyond the set-up pattern,
- They required an understanding of an algebraic formula to facilitate finding the solution, and
- They were difficult items (low *p*-values).

These complex combinations of item characteristics were absent in the other seven item-families. As anticipated from the beginning of the study, item characteristics and item quality apparently played integral roles in eliciting higher order thinking processes, in both multiple-choice and constructed-response item formats.

One unintended finding emerged from the panel's post hoc evaluation. I came to realize the importance of including mathematics experts when conducting this type of study. In hindsight, the experts should have been involved from the beginning of the category development, rather than enlisted as post hoc reviewers. It became clear that their expert knowledge would have aided in the initial development of the cognitive process categories and it would have facilitated in the identification of the cognitive processes, especially when the students' thinking was difficult to follow. Furthermore, the categories would have been based on a formal, content based language, which would have been familiar to mathematics teachers, compared to the informal, unrefined language invented by a non-mathematics expert.

CHAPTER V

Summary, Conclusions and Next Steps

Overview of Study

The purpose of this study was to determine whether test takers use different cognitive processes when they solve multiple-choice and constructed-response items. I conducted this study in an era when school accountability and high-stakes, large-scale assessments were seemingly as important as student learning itself. Thus, given the importance placed on accountability and testing by policy-makers, parents, school personnel, and other stake-holders, I sought to deepen our understanding about how students interact with test items – the foundation of a test score.

Measurement professionals use a myriad of statistical procedures to assess how students interact with test items. For example, psychometricians use either classical test theory or IRT models to examine the difficulty and discriminating ability of an item. The impact of an item's contribution to a student's final test score partly depends on how the total test score is calculated (i.e., pattern scoring, unweighted number correct, or weighted number correct). But, regardless of how the final test score is calculated, each item "possesses" a certain amount of information (in the parlance of IRT) that ultimately contributes to a student's test score.

The amount of item information varies by item and often by item format. Historically, multiple-choice items, on average, contribute more information (per unit of testing time and cost) to a test score compared to constructed-response items. This type of item information is important for many aspects of testing, for instance during forms assembly when a test developer has to select items that match a predetermined test

information function. Pearson and Garavaglia (1997) proposed additional ways to think about test and item information for large-scale assessments. One of their notions was to look at item information through the lens of information-value (or value-added). This general perspective was adopted for this study.

The sample for this study consisted of 8th grade students who were enrolled in mathematics courses that ranged from general mathematics to algebra. The teachers selected the classrooms from which they allowed students to be selected. Because all 8th graders were not accessible to me, the sample could not be considered a true random sample. However, to obtain a sample that was as representative as possible, while working within the constraints of the two schools, I generated a sample list from which I drew the sample. The list consisted of over 400 students from the two schools. Using a variation of systematic sampling with a random start, I selected 17 students from each of the two schools. The students knew that their verbal responses would remain anonymous, and that their participation would neither improve nor harm their mathematics course grades. There was no attrition in the sample.

Three instruments were used to collect the data: (a) a short demographic survey, (b) a test booklet, and (c) a protocol guide booklet. The test booklet was the main instrument as it contained the test items. The items were from the algebra strand and specifically assessed the area of algebraic patterns. Each of two test booklets was composed of 17 pattern items; eight of the items were multiple-choice, eight were constructed-response, and one was a performance item. The one performance item appeared in both test booklets. The 17 students within each school were randomly

assigned to each test booklet to control for any curricular and instructional differences between classrooms and schools.

I developed 32 (except the performance item) pattern items from the model-shell item writing procedure I created for this study. Generally, I started the item writing process with an original NAEP item that was written as a multiple-choice (or, in some instances the original item was written as a constructed-response) item. I then removed the response options (or created response options) and created a constructed-response version (or a multiple-choice version) of the original item. The third item was created by slightly altering the content of the original item and writing a multiple-choice (or constructed-response) item. The fourth item was created by removing the response options (or adding response options) from the third item to create a constructed-response (multiple-choice) version of the third item. These four items were called an item-family. The first and third and the second and fourth items were called comparable items because they represented the items with the slightly altered, yet comparable content, presented in the two different item formats.

I created eight item-families from eight original NAEP items. The purpose of writing the 32 items was to maintain content across the four items within an item-family so that differences in cognitive processes could be attributed to item format rather than item content. The performance item was added to the other 32 items for the following reason. If I found no differences between the constructed-response and multiple-choice items but I found some differences in the cognitive processes elicited by the Balanced Assessment item, I then would be able to attribute the absence of between item format

differences to the idea that the constructed-response items did not tap the kinds of cognitive processes that were tapped by the performance item.

A think-aloud procedure was used to collect the data for this study. I selected the procedure after I reviewed several analytical models (e.g., IRT, factor analysis, MANOVA, structural equation models) that were used in studies from the relevant published literature. However, these models do not capture the cognitive processes students used to solve the items. The think-aloud procedure does capture the needed data so I ultimately selected a methodology that was (and is) seldom used within the field of measurement, but that was most appropriate for this study.

The verbal data that emerged from the think-alouds were analyzed using a constant comparative approach to identify the categories that represented the cognitive processes. Both descriptive statistics and qualitative narratives were used to analyze the verbal data. After the data were analyzed, I convened a group of four mathematics experts to conduct a post hoc evaluation of the cognitive process categories and the analysis procedures. As part of the post hoc review, I duplicated the think-aloud procedure with one of the reviewers. She answered the items twice, first as a mathematics expert and second as an 8^{th} grade student. That is, she was instructed to solve each item using a strategy that she thought an 8^{th} grade student would use.

The overall result from both the qualitative and quantitative data analyses indicated that students employed similar cognitive processes to solve algebraic pattern items written in the multiple-choice and constructed-response formats, especially for the cloned NAEP items. A few unique cognitive processes emerged from the dots and stars and Balanced Assessment items that were not observed in the other items; however, the

cognitive processes were similar between the multiple-choice and constructed-response versions of these items.

Conclusions

I offer four conclusions based on the findings in this study. First, the addition of constructed-response items, at least as they have been operationalized in the current National Assessment of Educational Progress, does not appear, on the basis of the current analysis, to add additional information about algebraic patterns above and beyond that which could be gathered from multiple-choice items. This conclusion is based on the common cognitive processes students used to solve the multiple-choice and constructed-response items, both within each item-family and between them.

A plausible reason for the lack of between item format differences is that the constructed-response items available in this study do not function like genuine performance items. Even the stars and dots items, which come closest to resembling a performance item, do not elicit different cognitive processes between the item formats. Two distinctive features are observed between the items in the item families and the performance item. First, all the constructed-response items, except the stars and dots items, had, or at least elicited, only one solution path. The Balanced Assessment item, the only item in the corpus that had the look and feel of a genuine performance, elicited multiple solution paths (albeit largely unsuccessful for the students in my sample). Second, the students found most of the items easy to moderately easy to solve, except for the stars and dots item and the performance item. In fact, very few students successfully solved the latter item. Thus, perhaps one way to capture different processing between item formats is to develop constructed-response items that more closely resemble the

unique features in a performance item — an open solution paths and a high or moderate level of item difficulty.

Second, students appeared to use the most efficient strategy to solve any given item, regardless of item format. That is, if the presentation of an item leads students to the one most efficient way to solve an item, then that is the strategy the students will use, regardless of item format. Evidence of this conclusion was even seen in the Balanced Assessment item. On the surface, the most efficient way to solve the item is to count the dots within the diagonal rectangle. All of the students who attempted this item used the counting strategy to solve the first problem of the item. Most of the students continued to use, or at least attempted to use, this strategy as the problem became more difficult.

Third, item features — item difficulty, complex problem space, and multiple solution paths — seem to play an integral role in determining the types of cognitive processes elicited by an item. Comparing the dots and stars item-family and Balanced Assessment items to the other items provides the evidence to support this conclusion. For instance, most of the items used in this study are easy (indicated by item means ranging from .49 to 1.0), have one solution path, and are not intrinsically complex patterns. The dots and stars and Balanced Assessment items had the opposite features. As mentioned above, different cognitive processes became evident when I analyzed the verbal data for these five items.

And, even more telling is the evidence that emerges when comparing the Balanced Assessment item to the dots and stars items. Specifically, the students used a few different cognitive processes to solve these items, compared to the other items in the study, but the degree to which they use the cognitive processes differs qualitatively. That

is, although category AA (grapple with information) emerged for the dots and stars and Balanced Assessment items, the students struggled longer and harder on the Balanced Assessment item compared to the dots and stars items. The Balanced Assessment item was the most difficult and complex item in the study; yet, it allowed for the emergence of unique cognitive processes. If test developers want items on large-scale assessments to measure a range of cognitive processes, then perhaps the items need to be moderately difficult, have multiple solution strategies, and a complex problem space.

Fourth, I conclude that the use of think-alouds is a valuable tool in the field of measurement, especially when investigating research questions that deepen or broaden our understanding about the cognitive traits underlying items or adding new information about the construct of interest. As mentioned previously, psychometricians rarely employ the think-aloud methodology for either of these purposes. This study provides some evidence to promote its effectiveness.

Recommendations

Researchers, test developers, and test users would likely be interested in the results of this study. I propose recommendations for each group by suggesting improvements to this study's design or by suggesting applications of the methodology.

Researchers. There are a few design features that other researchers may consider altering if they attempt to replicate this study. First, I would recommend the involvement of content experts and teachers from the outset, rather than involving them at the end of the study in a post hoc fashion. Their content knowledge will help define the cognitive processes that emerge from the students' thinking. Their content knowledge also allows them to precisely describe the categories by using terminology familiar to the

mathematics community. They also have unique insight about how students approach certain mathematics concepts that comes from their daily interaction with the students, especially teachers with several years of teaching experience.

Second, I recommend that researchers practice using the think-aloud procedure in a pilot study or other practice event. Knowing when to remain quiet and when to probe comes with practice. In the pilot study, my first few think-alouds resembled tutoring sessions; a common occurrence with new users (Paulsen, 1999). I also had to practice maneuvering through the protocol guide and asking prompts in a way that would not interrupt or lead the students' thinking. Beyond practice, I suggest that novices work with an experienced user of think-alouds and be observed and critiqued by an experienced interviewer during practice interviews. Researchers could also consider using videotapes to monitor and evaluate their interview technique.

Third, researchers could examine the relationship between instruction and performance by determining whether students are more likely to use (and more successful when using) strategies that they have been explicitly taught as a part of their curriculum. If the students all use the same solution approach, regardless of instructional practice and item format, then the results would support the assertion that students use the most intuitively efficient and straightforward solution strategy rather than one they have been taught as a part of their curriculum.

Finally, when feasible, researchers should use think-alouds to supplement information obtained from IRT or factor analysis when examining test- or item-related research questions. For instance, the item information curve generated by an IRT model is a useful indicator of the variability of an item, but the reason(s) for the variability is not

detectable by the model, and therefore remains unknown. The same is true for the factors produced by factor analysis. The unknown can become more known if the information obtained from a think-aloud complements, deepens, broadens, or contradicts the analytical information.

Test Developers. I recommend that the think-aloud methodology be used to inform item development and research on the validity of test items. The results of one study (Paulsen, 1999) indicated that the think-aloud procedure can be successfully employed to assess the construct validity of items by asking students to state their understanding of the items and then compare their understandings with the intention of the item writer. Paulsen also suggested that think-alouds can be used to detect potential item problems — such as confusing or awkward sentence structure, multiple correct answers, typographical errors, or confusing graphics — that may confuse the test takers and thereby lead them to giving an erroneous answer. This use of think-alouds can bolster the validity of items.

The timing of think-alouds in the Paulsen study occurred when the items were still in the development stage and therefore were not fully screened or pilot tested by the time the study occurred. Perhaps the think-aloud procedure should be used after items have been fully screened as the think-aloud approach is too expensive and labor intensive to detect the sorts of item problems that experienced item reviewers can detect more efficiently. Paulsen (1999) agrees and adds the point that the elimination of obvious item problems before items are subjected to a think-aloud means that students will spend less time deciphering the errors, and more time thinking about how to answer the items.

Test Users. Teachers could use the think-aloud procedure to make curriculum and instruction decisions. For instance, teachers can gain first hand information about the various solution paths students use to solve the items, which can in turn inform them about how students apply (or misapply) the concepts taught by the teacher. Teachers also can detect whether subgroups of students — that is, students with some type of cognitive disability — have difficulty applying certain mathematical concepts by purposefully studying the cognitive processes used by the subgroup(s). Finally, teachers can assess their own item writing ability by conducting think-alouds on the items they write for their classroom tests. They may learn that the students' understanding of an item and their intention of what the item is supposed to measure are not aligned. Finally, there is some evidence in the reading education literature (c.f., Baumann and Seifert-Kessel, 1992) that the think aloud approach can actually help students develop new and transferable strategies.

Limitations and Next Steps

As I have implied throughout this report, this work has serious limitations, limitations which compromise its generalizability. First, the subjects in the study are by no means nationally representative, even though I did attempt to obtain diversity with respect to ethnicity and income. Second, the sample is small; a larger sample would have increased the power of the analysis, thus increasing capacity to detect subtle but significant differences. Third, the sample of items is even more limited; they represent one very narrow strand in the middle school mathematics curriculum. Fourth, a different and more elaborate categorization scheme for coding the think-alouds, the kind that may have emerged had I engaged the content experts earlier in the process, may have yielded

greater sensitivity to the cognitive depth that proved so elusive to uncover. Fifth, other items — other items from the algebra strand and items from other mathematical strands — might have yielded different cognitive processes.

There is need for further research in determining the cognitive processes that students use to solve test items. Specifically, certain student populations may use or apply cognitive processes differently, which in turn may affect how teachers decide to instruct the concepts to the various subgroups. This type of research could be implemented by purposefully sampling the subgroup(s) of interest and replicating the think-aloud methodology used for this study.

In addition, further research could be conducted to determine whether mixing items from more than one mathematical strand (e.g., number sense and geometry) affects how students cognitively process test items. In this study, I purposefully limited the area to algebraic patterns to obtain a "tight" study design. But, the design does not reflect the multiple mathematics concepts assessed on large-scale assessments. This design alteration could also determine how easily students switch their cognitive processing when alternating from one mathematical area to another. Such information could also affect test development. That is, a mathematics test may include items from geometry and algebra, and alternate the presentation of the two strands. If the results of the think-aloud study indicate that students are not effective or efficient at moving between the strands, then perhaps test developers would reconsider the ordering of items on a test. Finally, this study focused on a very narrow area of mathematics. Additional research is needed within mathematics and in other content areas, especially in the areas often tested on large-scale assessments, i.e., science and writing.

APPENDIX A

APPENDIX A

Student Demographic Survey

Student ID:_____

1. How old are you? _____

2. What school do you attend? _____

3. Are you a male or female? (Circle one)

   A. Female

   B. Male

4. How often do you do math homework? (Circle one)

   A. Few times a month
   B. Once a week
   C. Few times a week
   D. Every day

5. How often do you read for fun? (Circle one)

   A. Few times a month
   B. Once a week
   C. Few times a week
   D. Every day

APPENDIX B

APPENDIX B

Algebra Items

Form A

Family 1 - comparable items:

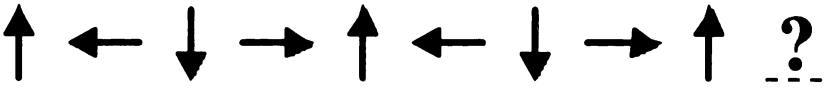
1. In the pattern above, which figure would be next?

A. 

B. 

C. 

D. 



2. In the pattern above, what figure would be next?

Answer:_____

Family 4 - comparable items:

1. From 1 vertex of a 4-sided polygon, 2 triangles can be drawn.
   From 1 vertex of a 5-sided polygon, 3 triangles can be drawn.
   From 1 vertex of a 6-sided polygon, 4 triangles can be drawn.
   From 1 vertex of a 7-sided polygon, 5 triangles can be drawn.

   How many triangles can be drawn from 1 vertex of a 20-sided polygon?

   A. 17
   B. 18
   C. 20
   D. Infinity

2. From any vertex of a 4-sided polygon, 1 diagonal can be drawn.
   From any vertex of a 5-sided polygon, 2 diagonals can be drawn.
   From any vertex of a 6-sided polygon, 3 diagonals can be drawn.
   From any vertex of a 7-sided polygon, 4 diagonals can be drawn.

   How many diagonals can be drawn from any vertex of a 20-sided polygon?

   Answer:_____

Family 5 - comparable items:

1. If the pattern shown in the table were continued, what number would appear in the box at the bottom of column B next to 14?

A	B
2	5
4	9
6	13
8	17
14	?

Answer: _____

2. If the pattern shown in the table were continued, what number would appear in the box at the bottom of column B next to 13?

A	B
1	4
3	8
5	12
7	16
13	?

A. 18

B. 26

C. 28

D. 32

Family 6 - comparable items:

Puppy's Age	Puppy's Weight
1 month	5 lbs.
2 months	12 lbs.
3 months	17 lbs.
4 months	20 lbs.
5 months	?

1. Jim records the weight of his puppy every month in a chart like the one shown above. If the pattern of the puppy's weight gain continues, how many pounds will the puppy weigh at 5 months?

A. 30

B. 25

C. 23

D. 21

Puppy's Age	Puppy's Weight
1 month	10 lbs.
2 months	15 lbs.
3 months	19 lbs.
4 months	22 lbs.
5 months	?

2. John records the weight of his puppy every month in a chart like the one shown above. If the pattern of the puppy's weight gain continues, how many pounds will the puppy weigh at 5 months?

Answer: _____

Family 8 - comparable items:

1. This question requires you to show your work and explain your reasoning. You may use drawings, words, and numbers in your explanation. Your answer should be clear enough so that another person could read it and understand your thinking. It is important that you show all your work.

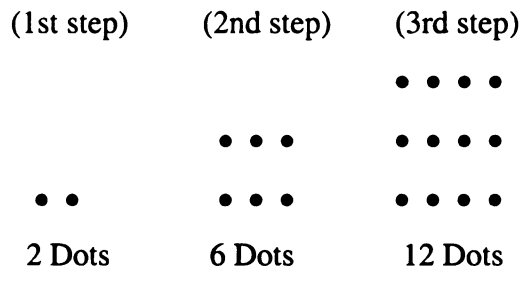
A pattern of stars is shown below. At each step, more stars are added to the pattern. The number of stars added at each step is more than the number added in the previous step. The pattern continues infinitely.

(1st step)	(2nd step)	(3rd step)
		* * * * *
	* * * *	* * * * *
* * *	* * * *	* * * * *
3 Stars	8 Stars	15 Stars

Joan has to determine the number of stars in the 16th step, but she does not want to draw all 16 pictures and then count the stars.

Explain or show how she could do this and give the answer that Joan should get for the number of stars.

2. A pattern of dots is shown below. At each step, more dots are added to the pattern. The number of dots added at each step is more than the number added in the previous step. The pattern continues infinitely.



Marcy has to determine the number of dots in the 20th step, but she does not want to draw all 20 pictures and then count the dots.

2a. How could Marcy figure out how many dots are in the 20th step?

A. She could figure out an algebraic formula that explains all 3 steps given and then apply the formula to the 20th step.

B. She could figure out the answer for the 4th step and then multiply the answer by 5.

C. She could draw the figures for steps 4 through 10 and then double the answer she got in the 10th step.

D. She could subtract the number of the step (20thstep) from the square of the step (20^2).

2b. What answer should Marcy get in the 20th step?

A. 100

B. 220

C. 420

D. 1,220

Form B

Family 1 - comparable items:

1. If the pattern shown in the table were continued, what number would appear in the box at the bottom of column B next to 14?

A	B
2	5
4	9
6	13
8	17
14	?

A. 19

B. 21

C. 23

D. 25

E. 29

2. If the pattern shown in the table were continued, what number would appear in the box at the bottom of column B next to 13?

A	B
1	4
3	8
5	12
7	16
13	?

Answer: _____

Family 2 - comparable items

Puppy's Age	Puppy's Weight
1 month	5 lbs.
2 months	12 lbs.
3 months	17 lbs.
4 months	20 lbs.
5 months	?

1. Jim records the weight of his puppy every month in a chart like the one shown above. If the pattern of the puppy's weight gain continues, how many pounds will the puppy weigh at 5 months?

Answer: _____

Puppy's Age	Puppy's Weight
1 month	10 lbs.
2 months	15 lbs.
3 months	19 lbs.
4 months	22 lbs.
5 months	?

2. John records the weight of his puppy every month in a chart like the one shown above. If the pattern of the puppy's weight gain continues, how many pounds will the puppy weigh at 5 months?

A. 30

B. 27

C. 25

D. 24

Family 4 - comparable items:

1. A pattern of stars is shown below. At each step, more stars re added to the pattern. The number of stars added at each step is more than the number added in the previous step. The pattern continues infinitely.

(1st step)	(2nd step)	(3rd step)
		* * * * *
	* * * *	* * * * *
* * *	* * * *	* * * * *
3 Stars	8 Stars	15 Stars

Joan has to determine the number of stars in the 16th step, but she does not want to draw all 16 pictures and then count the stars.

1a. How could Joan figure out how many stars are in the 16th step?

A. She could figure out an algebraic formula that explains all 3 steps given and then apply the formula to the 16th step.

B. She could figure out the answer for the 4th step and then multiply the answer by 5.

C. She could draw the figures for steps 4 through 8 and then double the answer she got in the 8th step.

D. She could subtract the number of the step (16thstep) from the square of the step (16^2).

1b. What answer should Joan get in the 16th step?

A. 100

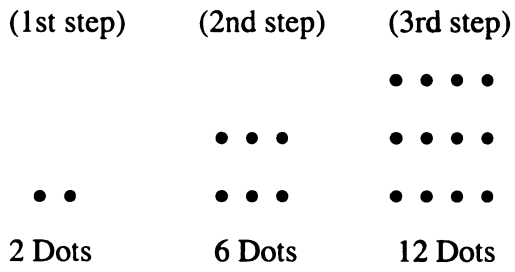
B. 220

C. 288

D. 1,118

2. This question requires you to show your work and explain your reasoning. You may use drawings, words, and numbers in your explanation. Your answer should be clear enough so that another person could read it and understand your thinking. It is important that you show all your work.

A pattern of dots is shown below. At each step, more dots are added to the pattern. The number of dots added at each step is more than the number added in the previous step. The pattern continues infinitely.



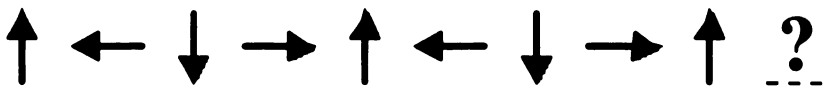
Marcy has to determine the number of dots in the 20th step, but she does not want to draw all 20 pictures and then count the dots.

Explain or show how she could do this and give the answer that Marcy should get for the number of dots.

Family 5 - comparable items:

1. In the pattern above, what figure would be next?

Answer:_____



2. In the pattern above, which figure would be next?

A.  B.  C.  D. 

Family 8 - comparable items

1. From 1 vertex of a 4-sided polygon, 2 triangles can be drawn.
From 1 vertex of a 5-sided polygon, 3 triangles can be drawn.
From 1 vertex of a 6-sided polygon, 4 triangles can be drawn.
From 1 vertex of a 7-sided polygon, 5 triangles can be drawn.

How many triangles can be drawn from 1 vertex of a 20-sided polygon?

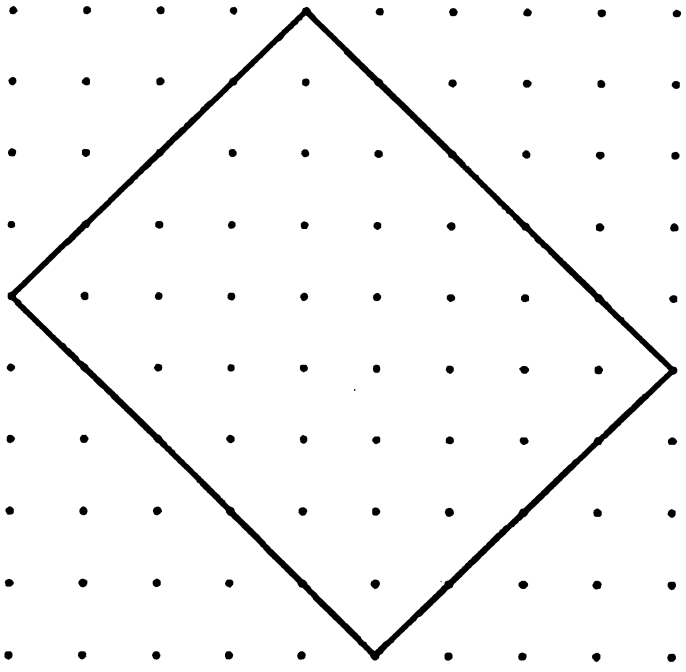
Answer:_____

2. From any vertex of a 4-sided polygon, 1 diagonal can be drawn.
From any vertex of a 5-sided polygon, 2 diagonals can be drawn.
From any vertex of a 6-sided polygon, 3 diagonals can be drawn.
From any vertex of a 7-sided polygon, 4 diagonals can be drawn.

How many diagonals can be drawn from any vertex of a 20-sided polygon?

A. 14
B. 17
C. 19
D. 20
E. Infinity

Diagonal Rectangle Problem



Let's call this shape a "4 by 5 diagonal rectangle." (One edge is 4 diagonal units long, and the other is 5 diagonal units long.)

1. How many dots are *inside* the 4 by 5 diagonal rectangle?

2. How many dots will lie inside a 5 by 6 diagonal rectangle?

3. How many dots will lie inside a 10 by 11 diagonal rectangle?

4. How many dots will lie inside a 100 by 101 diagonal rectangle? Explain how you got your answer.

5. How many dots will lie inside an n by (n+1) diagonal rectangle? Explain how you got your answer.

BIBLIOGRAPHY

BIBLIOGRAPHY

Balanced Assessment Package (1997). Balanced Assessment Package. Balanced assessment for the mathematics curriculum. Berkeley, CA: University of Ca.

Baumann, J. F., Seifert-Kessel, N., & Jones, L. A. (1992). Effect of think-aloud instruction on elementary students' comprehension monitoring abilities. Journal of Reading Behavior, 24, 143-167.

Bennett, R., Rock, D., Braun, H. Frye, D., Spohrer, J., and Soloway, E. (1990). The relationship of constrained free-response to multiple-choice and open-ended items. Applied Psychological Measurement, 14(2), 151-162.

Bennett, R., Rock, D., Wang, M. (1991). Equivalence of free-response and multiple-choice items. Journal of Educational Measurement, 28(1), 77-92.

Bennett, R. and Ward, W. (Eds) (1993). Constructing versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment. Lawrence Erlbaum Associates, Hillsdale, NJ.

Campbell, J. (1995) A comparison of thinking processes elicited by multiple-choice and constructed-response questions on an assessment of reading comprehension, Paper presentation at the National Reading Conference, 1995, New Orleans, LA.

Campbell, D. T. & Stanley, J. C. (1963). Experimental and quasi-experimental designs for research. Houghton Mifflin Company, Boston, MA.

Chauncey, H. And Dobbin, J. E. (1963). Testing: Its place in education today. New York:Harper and Row. In M. Martinez, (1993). Problem-solving correlates of new assessment forms in architecture. Applied measurement in Education, 6(3), 167-180.

Curriculum and evaluation standards for school mathematics. Reston, VA: National Council of Teachers of Mathematics, 1989.

Demby, A. (1997). Algebraic procedures used by 13-to-15 year-olds. *Educational Studies in Mathematics*, 33, 45-70.

Downing, S. and Haladyna, T. (1997). Test item development: Validity evidence from Quality assurance procedures, Applied Measurement in Education, 10(1), 61-82.

Ericsson , K. A. and Simon, H. A. (1993). Protocol analysis: Verbal reports as data (Rev. ed.). The MIT Press, Cambridge, Mass.

Farr, P., Pritchard, R., and Smitten, B. (1990). A description of what happens when an examinee takes a multiple-choice reading comprehension test. Journal of Educational Measurement, 27, 209-226.

Fraenkel, J. R. and Wallen, N. E. (1993). How to design and evaluate research in education, 2nd Ed. McGraw-Hill Inc., New York.

Frederiksen, N. (1984). The real test bias. American Psychologist, 39(3), 193-202.

Gerace, W. J., & Mestre, J. P (1982). The learning of algebra by 9th graders: Research finding relevant to teacher training & classroom practice. Paper prepared for the National Institutes of Education: Washington DC.

Goldstein, H. (1994). Recontextualizing mental measurement, Educational Measurement: Issues and Practice, 12(1), 16-19, 43.

Hair, J., Anderson, R., Tatham, R., & Black, W. (1992). Multivariate data analysis with readings, 3rd edition, Macmillan Publishing Company.

Haladyna, T. (1994). Developing and validating multiple-choice test items. Lawrence Erlbaum Assoc., Hillsdale, NJ.

Hambleton, R. K. and Swaminathan, H. (1987). Item response theory: Principles and applications. Kluwer-Nijhoff, Boston, MA.

Hamilton, L., Nussbaum, E., and Snow, R. (1997). Interview procedures for validating science assessments. Applied Measurement in Education, 10(2), 181-200.

Harris, D. (1993). Practical issues in equating. Paper presented at the Annual Meeting of the American Educational Research Association, Atlanta, GA.

Lukhele, R., Thissen, D., & Wainer, H. (1994). On the relative value of multiple-choice, constructed response, and examinee-selected items on two achievement test. Journal of Educational Measurement, 31, 234-250.

Martinez, M. (1991). A comparison of multiple-choice and constructed-response figural response items. Journal of Educational Measurement, 28,(2), 131-145.

Martinez, M. (1993). Problem-solving correlates of new assessment forms in architecture. Applied measurement in Education, 6(3), 167-180.

Mehrens, W. A. & Lehmann, I. J. (1987). Using standardized tests in education, fourth edition. Longman, Inc.: New York, NY.

Montague M. & Applegate B. (1993). Middle school students' mathematical problem solving: An analysis of think-aloud protocols. Learning Disability Quarterly, 16, 19-30.

Mueller, G. E. (1911). Zur analyse der gedachtnistatigkeit und des vorstellungsverlaufew: Teil I. Zeitschrift fur psychologie, 5. In Ericsson, K. and Simon, H. (1993). Protocol Analysis: Verbal reports as data. The MIT Press, Cambridge, Mass.

National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1996 Mathematics Assessment.

Nisbett, R. E. & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. Psychological Review, 84, 231-159.

Norris, S. P. (1990). Effects of eliciting verbal reports of thinking on critical thinking test performance. Journal of Educational Measurement, 27, 41-58.

Paulsen, C. (1999). An exploratory study of cognitive laboratories for development and construct validation of reading and mathematics achievement test items. Unpublished doctoral dissertation, University of Pennsylvania.

Pearson, P. D., Garavaglia, D. (1997) Improving the information value of performance items in large scale assessments. Washington, DC: Author.

Pearson, P. D., Garavaglia, D., Rodriguez, M., Danridge, J., & Montanez, M. (1997). Investigating cognitive engagement through think-alouds. Unpublished manuscript.

Silver, E. (1997). Algebra for all - Increasing students' access to algebraic ideas, not just algebra courses. Mathematics Teaching in the Middle School, 2(4), 204-207.

Snow, R. E. (1993). Construct validity and constructed-response tests. In R. E. Bennett and W.C. Ward (Eds.), Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment (pp. 45-60). Lawrence Erlbaum Associates, Hillsdale, NJ.

Thissen, D., Wainer, H., & Wang, X-B. (1994). Are tests comprising both multiple-choice and free-response items necessarily less unidimensional than multiple-choice test? An analysis of two tests. Journal of Educational Measurement, 31, 113-123.

Werts, C., Breland, H., Grandy, J., & Rock, D. (1980). Using longitudinal data to estimate reliability in the presence of correlated errors of measurement. Educational and Psychological Measurement, 40(1), 19-29