DOCUMENT RESUME

ED 390 924	TM 024 348

AUTHOR Suzuki, Kyoko; Harnisch, Delwyn L.

TITLE Measuring Cognitive Complexity: An Analysis of Performance-Based Assessment in Mathematics.

PUB DATE Apr 95

NOTE 35p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, April 18-22, 1995). The examples of students' responses may not reproduce well.

PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.

DESCRIPTORS Academic Achievement; Cognitive Processes; Communication Skills; Criteria; Educational Change; Grade 9; Grade 10; High Schools; *High School Students; *Mathematics Tests; *Measurement Techniques; *Scaling; Teaching Methods; *Test Theory

IDENTIFIERS *Cognitive Complexity; *Performance Based Evaluation

ABSTRACT

A new test theory for performance-based assessment is proposed. Criteria for "good" performance-based items, ways of measuring cognitive complexity, methods for determining maturity levels of understanding, and scaling systems are discussed. A performance-based task completed by 51 ninth and tenth graders in June 1993 was studied. Results demonstrated these criteria for tasks: (1) modeling real-world phenomena; (2) having multiple strategies; (3) having ordered categories for measuring maturity ıevels; (4) connecting several concepts to solve; (5) depicting the achievement levels by verbal explanations; (6) detecting the discrepancy between an intuitive solution and a mathematical solution; and (7) matching complexity of task with a scaling system. The analyses of students' responses suggest the importance of improving communication skills in classroom learning. The instability of cognitive shifts in students' solving strategies has implications for improving instructional strategies. Appendix A explains the scoring rubrics, and Appendixes B, C, and D present examples of performance responses. (Contains 11 tables and 19 references.) (Author/SLD)

Measuring Cognitive Complexity : An Analysis of Performance-based Assessment in Mathematics

Kyoko Suzuki

Delwyn L. Harnisch

University of Illinois at Urbana-Champaign

Department of Educational Psychology

210 Education Building
1310 S. Sixth Street
Champaign, IL 61820
Tel: (217) 333-4416
Fax: (217) 244-7620
email: suzuki1@uxa.cso.uiuc.edu

Running head: MEASURING COGNITIVE COMPLEXITY: AN ANALYSIS OF PERFORMANCE-BASED ASSESSMENT IN MATHEMATICS

Paper presented at the 1995 Annual Meeting of the American Educational Research Association, San Francisco, CA, April 18 - 22.

Abstract

The purpose of this study is to construct a new test theory for performance-based assessment. Criteria for "good" performance-based items, ways of measuring cognitive complexity, methods for determining maturity levels of understanding, and scaling systems are discussed. A performance-based task is investigated and analyzed from 51 ninth and tenth graders collected in June, 1993. The results demonstrated seven criteria for tasks: 1) modeling real-world phenomena, 2) having multiple strategies 3) having ordered categories for measuring maturity levels, 4) connecting several concepts to solve, 5) depicting the achievement levels by verbal explanations, 6) detecting the discrepancy between an intuitive solution and a mathematical solution, and 7) matching complexity of task with a scaling system. The analyses of students' responses suggest the importance of improving communication skills in classroom learning. The instability of cognitive shifts in students' solving strategies has implications for improving instructional strategies.

Measuring Cognitive Complexity :
An Analysis of Performance-based Assessment in Mathematics

Authentic assessment has moved to center stage as the focus of assessment has changed in the past decade. Since multiple-choice tests derive their value as educational indicators, the indicators are often confused with instructional goals, which has led to an overemphasis on indicators as an educational goal. The lack of correspondence between indicators and goals provides the motivation for "authentic" assessment which directly measures complex performance including more open-ended problems, essays, hands-on activities, etc. However, it's not enough to assume that alternative assessment for complex learning and processes are more valid than multiple-choice tests. Certain criteria need to be addressed for evaluating new assessments to consider a theoretical framework of validity (Harnisch, 1994a; Linn, Baker & Dunbar, 1991).

Moreover, the recent studies of cognitive psychology suggest the need for achievement test changes. The conventional achievement tests are based on the behaviorism of psychology so that the future tests should assess observable student behaviors that can be reliably recorded as either present or absent (Bloom, Hastings, and Madaus, 1971). However, the recent studies of cognitive psychology have changed the view of "learning". What is "learning"? The differences between a novice and an expert is not the amount of knowledge, but the ways of viewing phenomenon and of structuring problems. "Learning should be a qualitative change in a person's conception of a certain phenomenon or of a certain aspect of reality" (Johansson, *et al.*, 1985). Therefore, the purpose of assessment is not to establish the presence or absence of specific behaviors, but to infer the nature of students' understandings of a particular phenomenon (Masters and Mislevy, 1993; Mislevy, 1995).

Assessing performances should be clearly distinguished from assessing products. Assessing performances should be considered if task procedures have been explicitly taught and deviations from accepted practice can be detected, whereas assessing products should be considered if proper task procedures are diverse, indeterminate, or have not been explicitly taught (Fitzpatrick & Morrison, 1971; Magone, *et al.*, 1994). In performance assessment of competencies or other constructs, replicability and generalizability are important. Inferences from observed behavior should be made to construct the knowledge and skills underlying test behavior (Harnisch & Hanson, 1994; Messick, 1994). When we consider the criteria of performance-based items, it should link-up with the instructional goal. Also, the scoring system should reflect the view of inferences about the achievement and understandings, and not the matter of absence or presence of knowledge.

Considering a performance-based assessment in a large-scale achievement testing, a certain degree of task structure is required to ensure a valid assessment of students' proficiency. A goal of performance-based testing is to provide assessments which allow students to display their thinking, reasoning and strategic process. The other goal is to provide assessment tasks that allow all students to perform at their best and that is related to goals of the instructional programs in schools (Magone, *et al.*, 1993; Parke & Lane, 1993). To realize these goals, assessment instruments need to be developed to measure cognitive complexity such as variety of strategies in solving problems, reasoning skills, and communicating their thinking mathematically (Lane, 1993, Lane *et al.*, 1993).

This study examines performance-based assessment tasks in mathematics and student responses to these tasks revealing varieties of thinking and reasoning processes. A performance-based task in this paper means a paper-pencil test with constructed responses, and it doesn't include the broader meanings such as a portfolio assessment. A criterion of a performance-based task is the capability of providing students opportunity to display their thinking process at their best. A problem having varieties of strategies in students' responses was chosen in this study to analyze the task structure, so that the quality of performance-based items can be determined. The strategies taken in solving problems are analyzed precisely so that the cognitive complexities of student responses are classified based on the task structure. This research paper provides a step of constructing "orderedoutcome categories" for measuring cognitive achievement in mathematics, which will be a new measurement for "learning."

Purposes

As the current movement toward alternative assessment is in process, the need for constructing a new test theory for performance-based assessment is increasing. In performance-based assessment, the cognitive mechanisms that underlie both the learning and the assessment of mathematics need to be investigated. Two major objectives are addressed in this study; 1) how we can assess achievement and understandings in subject matter learning, and 2) how we can measure the student's thinking process and strategies they use in solving problems.

This study focuses on five general purposes:

1) Developing new achievement testing which can infer an understanding level, not measuring only the presence or absence of knowledge
2) Developing assessment tasks in which all students can display their understanding
3) Determining criteria for "good" performance-based items
4) Developing ordered categories for measuring achievement levels
5) Developing a new scaling system for measuring levels of thinking

To realize these purposes, five specific objectives are used for this study:

1) Analyzing logical structures of a performance-based assessment task
2) Analyzing cognitive structures of students' responses
3) Determining if the task could detect the cognitive level of a student's understanding
4) Defining ordered categories for assessing achievement levels based on the analysis of students' responses
5) Discussing the validity of the task structure and the scoring system

The logical analyses of performance-based assessment tasks in mathematics and the analyses of students' responses to these tasks reveal varieties of thinking and reasoning processes. Based on these detailed analyses, a system of ordered categories for levels of mathematical maturity can be proposed. The task structure and the validity of the scoring system were examined to determine if they could evoke and evaluate the various levels of cognitive processes for constructing a cognitive model for a new test theory.

Methods

An item including three sub-questions was investigated in this study. The item was originally developed for the Alternative Assessment Project for IGAP (Illinois Goal Assessment Program, Harnisch, 1994b). Among the many items developed for IGAP, this item was chosen for the study because it involved varieties of strategies for solving the problems, and the strategies could be ordered based on the achievement levels.

> Telephone area codes in the U.S. and Canada consist of 3 digits, in which the first is a digit from 2 through 9, the second is either 0 or 1, and the third can be any digit except 0.
>
> (1) According to these rules, how many different area codes can begin with 6?
>
> (2) How many different area codes can be an odd number?
>
> (3) What is the probability that an area code is a multiple of 3?

This item was administered to 9th and 10th graders in June, 1993, by teachers who participated in the Alternative Assessment Project conducted at the University of Illinois at Urbana-Champaign. The task was treated as either a group or an individual task for students with no time limitations, and the responses were written separately by each individual. Responses from 51 students were collected and examined for this study.

The responses were scorud using the general rubric of the QUASAR project[1] , which consists of three interrelated components: mathematical knowledge, strategic knowledgc, ⁊d communication. These components are specified for each of five score levels (0-4; see Appendix A) and the responses were graded from 0 to 4 with holistic perspectives considering three components. Cognitive strategies taken in each question were classified based on the student's responses.

Four raters scored and reviewed the students' responses. Two raters were familiar with performance-based assessment, while two were not. In this study, the scaling was based on the two skilled raters when there was a discrepancy the ratings.

The students' responses were analyzed in detail, both quantitatively and qualitatively.

Results and Discussions

The performance-based item evoked a variety of cognitive strategies. As a result, the students' responses triggered several topics to consider for performance item development and item scoring. Item structures and the students' responses were analyzed in detail, both qualitatively and quantitatively. The analyses provided opportunity to develop a model of ordered categories of achievement levels and scaling. The analyses focused on four aspects:

1. logical analyses of the task structure,
2. cognitive analyses of the ɔtudents' responses,
3. scoring the responses, and
4. classifications of the examples with strategies by scoring.

[1] QUASAR (Quantitative Understanding: Amplifying Student Achievement and Reasoning) is a national project that seeks to demonstrate that is feasible to i nplement instructional programs in the middle-school grades that promote the acquisition of thinking and reasoning skills in mathematics (Silver, 1991, Silver & Cai, 1993). The project is directed at students attending schools in economically disadvantaged communities.

1) Logical Analyses of the Task Structure

The task was designed for assessing the concepts of number sense[2] and probability with a real-world situation. In fact, there are some exceptions to area codes for this rule. For example, 210 is not legitimate, but it is the code for San Antonio, Texas, and 911 is legitimate, but it is used as a fire/emergency number in many local communities, and not as a telephone area code. However, the authors believe this problem can be a realistic task because it is modeling a real-world situation, and therefore it is effective as a performance-based item. A realistic task is not necessarily a real occasion in the world. Many times a real-world phenomenon contains several exceptions and complex situations. The real-world phenomenon may be simplified for a specific purpose in learning or completing a task. This is a critical distinction between assessment materials and real-world phenomena. It is important that a performance task reflects some aspect of a real-world setting, but it is not necessary to be an exact real-world problem. Thus, modeling a real-world phenomenon is an additional criterion for consideration in developing performance-based measurements.

The task requires students to find the number of possibilities which satisfy several conditions. Question 1 asks students to identify the rule for generating numbers given in the problem under a restricted condition. Question 2 requires students to find odd numbers in all possible area codes, and Question 3 requires students to organize several steps to solve the problem. These stages are:

(a) counting all possibilities of area codes,

(b) identifying a property of a multiple of 3,

(c) listing all numbers of a multiple of 3 under a given condition, and

(d) finding the probability.

The purpose of an item having multiple stages like Question 3 is to assess the ability to connect several concepts in solving a problem. It is notable that it is easy for students to solve a problem in a same content area they studied, but it is difficult to connect some

[2] There are various ways to solve the problem, and students could solve it without resorting to the principle of counting.

concepts which were taught in different content areas. This is a serious disadvantage of many math curricula at this time. Question 3 was used to detect the lack of connections among several related concepts taught in different conten. areas. The results from Question 3 will be analyzed closely to evaluate this dimension.

2) Cognitive Analyses of the Students' Responses

The different types of cognitive strategies for solving a problem were examined for each question of this task from a sample of 51 high school algebra I students in a midwestern community.

__Question 1__. The purpose of this problem is for students to understand the condition of the task. Most students could manage the setting of the task. Three strategies taken and represented were:

Strategy 1	listing all numbers
2	using combinations with a pictorial chart (such as □□□ or a tree diagram)
3	using the concept of combination without a visual aid or chart

__Question 2__. This task reveals a greater variety of the students' understandings. Five strategies taken and included were:

Strategy 1	listing all possible area codes and counting the total number of odd numbers
2	listing all odd numbers for the 600s (or for other hundreds like the 200s) and multiplying it by 8 because of the 8 cases where the first digit ranges from 2 to 9
3	knowing 5 cases for the third digit, and listing all possibilities for the first and second digit to obtain 5 x 16 = 80
4	using combinations with a pictorial chart (□□□, or a tree diagram) and finding the formula 8 x 2 x 5 = 80
5	using combinations without any chart and applying the formula 8 x 2 x 5 =80

__Question 3__. To solve the problem a student needs to go through three stages: 1) finding the total number of all possible area codes, 2) finding the total number of multiples of 3, and finally 3) finding the probability, $\frac{1}{3}$. Since there were several strategies for solving each stage, the strategies for solving the task varied because of the combinations of each stage. In fact, there were nine different patterns observed from the sample to obtain the correct

number of the probability; however, five of them have insufficient or incorrect mathematical reasoning. Moreover, the type of errors varied.

- Stage 1. Finding the number of all possible area codes

Strategy 1	listing all 144 possibilities
2	using combinations without a chart to find the formula 8 x 2 x 9 = 144
3	using combinations with a pictorial chart or diagram to find the foı mula 8 x 2 x 9 = 144
4	using the result of Question 1 to find 18 possibilities beginning with 6 for the first digit, and multiplying it by 8 because of 8 cases for the first digit from 2 to 9
5	listing 18 possible area codes beginning with a number for the first digit (such as 2 □□), and multiplying i. by 8 because of 8 cases for the first digit from 2 to 9

- Stage 2. Finding the total number of multiples of 3

Strategy 1	listing all 48 possibilities
2	listing all possibilities knowing a property of a multiple of 3 (the sum of each digit of a number equals to a multiple of 3)
3*	counting the number of multiples of 3 for the 600's to get 6, and multiplying 6 by 8 because of 8 cases for the first digit from 2 to 9
4*	counting the number of multiples of 3 for the 200's and the 300's to find 6 cases for both, and multiplying 6 by 8 to get 48 because of 8 cases for the first digit from 2 to 9
5*	dividing the number of all 144 possible area codes by 3 to get 48 because of the divisibility by 3

(Note: * indicates an insufficient or incorrect strategy.)

- Stage 3. Finding the probability, $\frac{1}{3}$

Strategy 1	following the definition of the probability, $48 \div 144 = \frac{1}{3}$
2*	finding the 3 cases of a multiple of 3 for 600's, and getting $6 \div 18 = \frac{1}{3}$
3*	finding $\frac{1}{3}$ because of the divisibility by 3 (intuitive conclusion)

(Note: * indicates an insufficient or incorrect strategy.)

3) Scoring the Performances

The results of scoring the students' responses are shown in the Table 1.

Insert Table 1 here

In Question 1, most students managed to complete the task. However, they tended to use few written words to present their solution or thinking process. In Question 2, more than 70% of the students scored 3 or 4, which means the majority students could figure out the meaning of the task. Some students answered the number of odd numbers among the numbers found in Question 1. These responses were scored as "no count," as it is a problem with the directions in the stem. The stem needs to be rewritten to have a clearer direction. Since the students' responses varied more in Question 2 than Question 1, the set of scoring reasons by the raters' review of the students' responses also varied. The lack of communication skill, insufficient information to draw the conclusion, and inability to identify important elements were found in common features for the low achievers. In Question 3, many students failed to complete the tasks, although a variety of strategies were observed. The set of scoring reasons by the raters' review of the students' responses were also diverse.

4) Classifications of the Examples with Strategies by Scoring

The students' responses were classified into strategies by score. Examples of these classifications help not only in analyzing students' understandings, achievement levels, and misunderstandings qualitatively and quantitatively, but also in developing instructional strategies and materials for classroom teachers.

Question 1

The classification of the responses with strategies by scoring is shown in the Table 2. All 51 responses were classified. The students' examples are shown in Appendix B.

Insert Table 2 here

The responses which scored 1 (or 0) did not show any strategy to solve the problem. This is an important point for teachers because the low achievers need help constructing a strategy to solve it. The students' work shown in Appendix A is also very helpful for this purpose in classroom use because the work of classmates gives examples for constructing a strategy.

As for ordering achievement categories, Strategy 1 (listing all possibilities) is the most basic way. Using combinations to find all possibilities is a more abstract concept to solve the problem; hence, the strategy of combination is supposed to be a higher level of achievement than merely listing all area codes. Using a pictorial chart may or may not be crucial for this case, as we do not know whether or not the students used any charts to solve the problem. A student may not have reported using a chart in the answer, even though a chart may have been used to find the answer. Therefore, Strategy 2 and Strategy 3 can be combined together in this case. Table 3 shows the results for the various strategies used.

Insert Table 3 here

The distribution of the students are almost evenly classified into either Category 1 or Category 2.

Question 2

Forty-seven responses were examined (4 responses of the original 51 were classified as "no count" because of finding the answer among the 600's only). The results are shown in Table 4, and the students' examples are shown in Appendix C.

Insert Table 4 here

Strategy 1 is a listing of all possibilities, which is the most basic level of achievement. Strategy 2 is a listing of all odd numbers for the 600's and then finding the answer by using 10 x 8 = 80, which is a more abstract way than Strategy 1. Strategy 3 is a listing of all odd numbers for the 60 case and finding the answer by using 5 x 2 x 8 = 80. This strategy is almost the same as Strategy 2, but is a bit more abstract than that. Therefore, Strategy 2 and Strategy 3 can be combined as a middle level of transition from listing to combinations. Strategy 4 is using combinations with a pictorial chart, and Strategy 5 is using combinations without a chart; hence, these two can also be combined into one for the same reason as Question 1. This category, using the concept of combination, appears to be the most abstract way among the three categories. The results are shown in Table 5.

Insert Table 5 here

Comparing the results from Table 5 with Table 3, we can see that a cognitive shift occurred in students from Question 1 to Question 2. Only 6% of the students listed all 80 possibilities, while 41% of the students moved up to the middle level of strategy which is more abstract and involved the use of combinations.

Taking a closer look at a student response (see Example 1 in Appendix C), the response using a higher level strategy was scored a "2" because of a misconception. The numbers of the second and third digit were found correctly, but the number of the first digit was found incorrectly because the student calculated the difference between 9 and 2 instead of determining the actual numbers possible that begin with 2 through 9. This misconception was observed quite often among the students. This common error is also detectable in the performance-based task, and it is useful to correct the students' misunderstanding if the example of a student's response is used in classroom instruction.

This error also showed there is some confusion when to shift up conceptually. The learners tried to use a new method, but they tended to apply it incorrectly. This often happens between stages of learning. However, this is a good chance to recognize common mistakes if the case is used in classroom teaching.

Question 3

This problem contains three stages to solve it. All students, however, do not proceed through all three stages. Some performed all stages, but some did parts of them. Therefore, the categorical classification of the strategies does not include all responses at each stage.

1) Stage 1: Finding All Possibilities

Eighteen out of 51 responses functioned at Stage 1. The distribution of the responses with strategy by score is shown in Table 6 and the examples of responses are in Appendix D-1.

Insert Table 6 here

Both Strategy 2 and Strategy 3 used the formula 18 x 8 = 144 for finding the number of all possibilities, but Strategy 2 used the result of Question 1, while Strategy 3 did not. Therefore, these two can be combined into one category, which is a transition level from listing to combinations. Strategy 4 and Strategy 5, in which the use of combinations observed with or without a pictorial chart, respectively, can be classified into a category. The result of this categorical classification is shown in Table 7.

Insert Table 7 here

More than one-half of the students were located in a transition level. Many of the students using combinations failed to solve the problem.

2) Stage 2: Finding the Total Number of a Multiple of 3

Thirty-three out of 51 responses functioned at Stage 2. The distribution of the responses for the strategy is shown in Table 8 and the examples are in Appendix D-2.

Insert Table 8 here

Strategy 3, strategy 4, and strategy 5 contain insufficient and/or incorrect logic to find the answer; however, the accuracy of the usage of mathematical reasoning was not stressed here for the scoring. If a student showed some knowledge about a concept for solving the problem, the points were given for the work. This scoring criterion meets the "inference" policy of performance-based measurement in which an achievement test should be an inference of a learner's understanding.

Strategy 2, in which a property of a multiple of 3 was mentioned to find the total number of multiples of 3, contains higher insight toward the problem. Example D-2 in Appendix D-2 displays a further investigation to the problem. The awareness of the property is an important aspect in this problem. Although the occurrence of this strategy was very small, it was classified as a category.

The strategies taken at this stage were classified into three: listing, checking with a property of a multiple of 3, and an insufficient way. The results are shown in Table 9.

Insert Table 9 here

3) Stage 3: Finding the Probability

Twenty-five responses out of 51 appeared at this stage, but 36% of the responses just wrote "$\frac{1}{3}$" for their answer without any explanation. These responses were scored as a level "1." These responses can been considered as an intuitive solution, since the students answered $\frac{1}{3}$ because of the probability of a multiple of 3. In this scoring criterion, the reasoning or explanation is stressed; therefore, they were scored a "1" because of no explanation. The results are shown in Table 10, and some examples of their responses are in Appendix D-3.

Insert Table 10 here

The responses classified in Strategy 1 using a formula $48 \div 144 = \frac{1}{3}$ showed most of the understanding needed to solve this problem. The responses having insufficient or intuitive strategy contained a lack of understanding this problem. Therefore, Strategy 1 formed a category (mathematical solution), and Strategy 2 and Strategy 3 combined to form a category (intuitive/insufficient solution). The results are shown in Table 11. Those who were classified into the category of mathematical solution scored a "3" or "4", while those who were in the category of intuitive/insufficient one scored low.

Insert Table 11 here

4) Overall Performance

In summary, the level of students' performance on this task was at a low level. The primary reason for this result was that there were many students solving this question intuitively without any explanations, which were scored as "1". This is a serious deficiency in a performance-based task. An assessment task should be able to detect the discrepancy between an intuitive solution and a mathematical solution. Having a conflict between these two, students can realize the usefulness of a mathematical solution as well as mathematics learning. Thus, this task doesn't meet one of the criteria for "good" assessment tasks.

Moreover, each table showed too much variability, although the categorical classification showed some trends. For our purpose to develop an ordered categorical classification for measuring cognitive change and shift, this task appeared too complex to identify the cognitive levels.

Conclusions

The first conclusion from this investigation is about criteria of "good" performance-based tasks. The tasks were developed intending four criteria:

1. having a real-world context,
2. having multiple strategies and representations,
3. connecting multiple concepts to solve a problem, and
4. depicting achievement levels by explaining solving strategies verbally.

The first criterion is reflecting an authentic situation for solving a problems in daily life. The second criterion is intending open-ended situation, more than one correct answer or solution for the problem. The third criterion is intending to measure an ability to construct relations among different concepts taught in different units in the school curriculum. The last criterion is also important for performance-based tasks intending to measure communication skills. An assessment task of achievement testing must be distinguished from a puzzle or a game. For example, a magic square problem contains many strategies to solve it. It is good for classroom use to play with numbers, but it is not suitable for an achievement problem because the verbal explanation of the strategies for the solving problem does not reflect achievement levels or mastery levels of a content area. Therefore, the task is not appropriate for assessing achievement levels (Suzuki, 1993).

This study also demonstrated other criteria for the performance-based items. First, a task does not need to have a real-life context exactly, but it should model real-world phenomena. Phenomena in real-life often contain some exceptions or too complicated situations, so they are sometimes inappropriate for learning materials or assessment tasks. In education, some simplification is needed for understanding a basic concept. This is a crucial distinction between learning materials and real-world situations. Therefore, we need to change the first criterion for "good" assessment tasks; a task should model real-world phenomena.

For the second criteria, having multiple strategies and representations, is not sufficient to measure levels of conceptual maturity. For example, in Question 2, there were mainly two strategies: listing all 80 possibilities and using the concept of combinations. Although there is nothing wrong mathematically for listing all possibilities, the strategy of using combinations is more abstract and shows higher level of maturity in mathematical understanding than the other strategies. In order to assess levels of maturity, classification of ordered categories may be useful. Ordered categories can measure maturity levels as well as cognitive changes or shifts of strategies.

Question 3 suggested other criteria for assessment tasks. An assessment task should be able to detect the discrepancy between an intuitive solution and a mathematical solution. Moreover, the task complexity should match the scaling system used. If a task is too complicated to score with 0-4 scaling, the score does not have good information for representing achievement levels for educational use. However, this conclusion raised other problems. This issue is addressed later in a consideration of scaling system.

Based on these considerations, the following seven criteria are drawn for "good" performance-based items.

1. modeling real-world phenomena
2. having multiple strategies & representations
3. having ordered categories for measuring maturity levels and cognitive shifts
4. connecting several concepts to solve
5. depicting achievement levels by explaining solving strategies in words
6. detecting the discrepancy between an intuitive solution and a mathematical solution
7. matching complexity of task with a scaling system

The second consideration is about measuring maturity levels of mathematical understanding. The current scaling system does not distinguish the differences of the strategies. This study demonstrated the use of developing ordered categories for this purpose. Ordered categories can be used not only for inferring the maturity levels of understanding but also for finding cognitive changes or shifts.

The last consideration is about scaling. A scaling system should be easy to use and sufficient to measure the multidimensionality of an ability. The current scoring system does not measure the maturity levels of strategies. Also, the scaling system should provide information for improving communication levels of thinking. Finally, the scaling should match the task complexity. For the current scaling system, the complexity of Question 1 and Question 2 appeared appropriate, but Question 3 seemed too complex to measure the levels of achievement. Should we then exclude tasks from achievement testing which require the higher level of thinking to organize some mathematical concepts? Should we exclude from mathematics education complex tasks which require the organization of different concepts? The problem of task complexity raised other problems for developing achievement testing.

Implications and Further Considerations

This study revealed that students tended to lack communication skills as measured in their written responses. Since the objectives and criteria of scoring in performance-based assessment should be clearly understood by examinees, communication skills should also be taught and stressed in the classroom. Meanwhile, for active learning in a classroom, communication should be stressed to construct concepts. Since the scoring criteria should tie up with instructional goals, perfoımance-based assessment can provide "good" information for teachers to conduct active learning in the classroom.

The other implication to instruction is about the wav learning occurred. The analyses of students' responses demonstrated cognitive shifts for s. ving strategies between each question. However, the shift is not stable for learners; it is going back and forth between categories. This result showed the stages of maturity levels of students' understanding are not stable, which suggests that teachers use flexible instructional strategies to improve students' achievement levels in class.

The consideration of task complexity raised some questions for developing achievement testing. Since the task complexity is restricted by a scaling system, we do need to develop both assessment tasks and a scaling system simultaneously. However, how can we determine the appropriateness of task complexity? As for an educational goal, we expect students to have the ability to solve a complex problem in a real-world context which requires them to analyze and organize complex situations in order to find a solution. Then, what ability should we assess in a performance-based assessment task? How can we assess the ability which organizes different concepts in a complex situation? What scaling is appropriate for measuring complex ability? These are further questions for developing both scalings and tasks.

Finally, the consideration for measuring levels of mathematical maturity also raised questions for developing scaling systems. What are the expected skills for solving problems? How can we order mathematical understanding levels for various strategies? How can we measure the expected maturity in mathematical strategies of students? These questions also need to be investigated in further research.

In summary, this study is a step for improving our understanding of the following important questions;

- what are the criteria for a "good" item on a performance-based assessment?
- how can we measure cognitive complexity?
- what are the meaningful scales for measuring multidimensionality of an ability?

Acknowledgments

The authors wish to thank Dr. Viji Sundar (California State University at Stanislaus), Dr. Kenneth Travers (University of Illinois at Urbana-Champaign), and the reviewers of the proposal for their comments and assistance in this study.

References

Bloom, B. S., Hastings, J. T., & Madaus, G. F. (1971). Handbook on formative and summative evaluation of student learning. New York: McGraw-Hill.

Fitzpatrick, R., & Morrison, E. J. (1971). Performance and product evaluation. In R. L. Thorndike (Ed.), Educational Measurement (2nd ed., pp. 237-270). Washington, DC: American Council on Education.

Harnisch, D. L. (1994a). Performance assessment in review: New directions for assessing student understanding. International Journal of Educational Research, 21, 341-350.

Harnisch, D. L. (1994b). Performance assessment in mathematics. Final report to the Illinois Goal Assessment Program, University of Illinois at Urbana-Champaign.

Harnisch, D. L., & Hanson, A. (1994). An examination of multiple representations from a problem solving task: Implication for testing and instruction. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Harnisch, D. L., & Mabry, L. (1993). Issues in the development and evaluation of alternative assessment. Journal of Curriculum Studies, 25(2), 179-187.

Johansson, B., Marton, F., & Svensson, L. (1985). An approach to describing learning as change between qualitatively different conceptions. In L. H. West & L. A. Pines (Eds.), Cognitive structure and conceptual change (pp. 233-257). Orlando, FL: Academic Press.

Lane, S. (1993). The conceptual framework for the development of a mathematics performance assessment, Educational Measurement: Issues and Practice, 12(2). 16-23.

Lane, S., Liu, M., Stone, C. A., & Ankenmann, R. D. (1993, April). Validity evidence for QUASAR's mathematics performance assessment. Paper presented at the annual meeting of the American Educational Research Association; Atlanta, GA.

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. Educational Researcher, 20(8), 15-21.

Magone, M. E., Cai, J., Silver, E. A., & Wang, N. (1994). Validating the cognitive complexity and content quality of a mathematics performance assessment. International Journal of Educational Research, 21, 317-340.

Magone, M. E., Wang, N., Cai, J., & Lane, S. (1993, April). An analysis of the cognitive complexity of QUASAR's performance assessment: Tasks and their sensitivity to measuring changes in students thinking. Paper presented at the annual meeting of the American Educational Research Association, Atlanta, GA.

Masters, G. N., & Mislevy, R. J. (1993). New views of student learning: Implications for educational measurement. In N. Frederiksen, R. J. Mislevy, & I. I. Bejar (Eds.), Test theory for a new generation of tests. Hillsdale, NJ: Lawrence Erlbaum Associates.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. Educational Researcher, 23(2), 13-23.

Mislevy, R. J. (1995). Evidence and inference in educational assessment (Research Rep. No. RR-95-8-ONR). Princeton, NJ: Educational Testing Service.

Parke, C., & Lane, S. (1993, April). Designing performance assessments: An examination of changes in task structure on student performance. Paper presented at the meeting of the National Council on Measurement in Education, Atlanta, GA.

Silver, E. A. (1991). Quantitative understanding: Amplifying student achievement and reasoning. Pittsburgh, PA: Learning Research and Development Center.

Silver, E. A., & Cai, J. (1993, April). Schemes for analyzing student responses to QUASAR's performance assessments: Blending cognitive and psychometric considerations. Paper presented at the meeting of the National Council on Measurement in Education, Atlanta, GA.

Suzuki, K. (1993, December). An investigation of performance-based tasks. Unpublished manuscript, University of Illinois at Urbana-Champaign, College of Education, Champaign, IL.

TABLE 1
Distribution of the Score of Student Responses
(n=51)

	Score				
Problem	4	3	2	1	0
(1)	51 %	37 %	8 %	4 %	0 %
(2)#	33 %	41 %	12 %	4 %	2 %
(3)	1 %	20 %	14 %	35 %	19 %

TABLE 2
Distribution of the Students' Responses for Question (1):
Strategy × Score
(n=51)

	Strategy		
Score	1 (listing)	2 (combination with a chart)	3 (combination without a chart)
4	24 %	9 %	18 %
3	22 %	8 %	7 %
2	8 %	0 %	0 %
1	4 % (no categorical classification observed)		
0	0 %		

TABLE 3
Distribution of the Stunts' Responses for Question (1):
Categorical Classification
(n=51)

	Category	
Score	A (Listing)	B (Combination)
4	24 %	27 %
3	22 %	17 %
2	8 %	0 %
1	4 % (no categorical classification observed)	
0	0 %	

TABLE 4
Distribution of the Students' Responses for Question (2):
Strategy × Score
(n=47)

	Strategy				
Score	1 (listing)	2 (10 x 8 =80)	3 (5 x 16 = 80)	4 (combination with a chart)	5 (combination without a chart)
4	2 %	14 %	0 %	9 %	14 %
3	4 %	14 %	2 %	11 %	11 %
2	0 %	9 %	2 %	0 %	2 %
1	6 % (no categorical classification observed)				
0	0 %				

TABLE 5
Distribution of the students' responses for Question (2):
Categorical Classification
(n=47)

	Category		
Score	A (listing)	B (middle)	C (combination)
4	2 %	14 %	23 %
3	4 %	16 %	22 %
2	0 %	11 %	2 %
1	6 % (no categorical classification observed)		
0	0 %		

TABLE 6
Distribution of the Students' Responses for Question (3):
Strategy × Score
Stage 1: All Possibilities
(n=18)

	Strategy				
Score	1 (listing)	2 (18 x 8 = 144 using the result of (1))	3 (listing the cases of 2 □□, 18 x 8 =144)	4 (combination with a chart)	5 (combination without a chart)
4	4 %	17 %	4 %	0 %	4 %
3	0 %	23 %	0 %	0 %	0 %
2	12 %	0 %	0 %	12 %	12 %
1	0 %	12 %	0 %	0 %	0 %
0	0 %	0 %	0 %	0 %	0 %

TABLE 7
Distribution of the Students' Responses for Question (3):
Categorical Classification
Stage 1: All Possibilities
(n=18)

	Category		
Score	**A** (listing)	**B** (middle)	**C** (combination)
4	4 %	21 %	4 %
3	0 %	23 %	0 %
2	0 %	0 %	24 %
1	12 %	12 %	0 %
0	0 %	0 %	0 %

TABLE 8
Distribution of the Students' Responses for Question (3):
Strategy × Score
Stage 2: The Total Number of a Multiple of 3
(n=33)

	Strategy				
Score [illegible]	**1** (listing)	**2** (checking with a property of a multiple of 3)	**3*** (6 x 8 = 48)	**4*** (listing the possibilities in 200's and 300's, then 6 x 8 = 48)	**5*** (144 +3 = 48)
4	15 %	3 %	6 %	0 %	0 %
3	15 %	3 %	3 %	10 %	3 %
2	12 %	3 %	6 %	0 %	3 %
1	0 %	0 %	3 %	12 %	3 %
0	0 %	0 %	0 %	0 %	0 %

(* indicates an insufficient /incorrect strategy)

TABLE 9
Distribution of the Students' Responses for Question (3):
Categorical Classification
Stage 2: The Total Number of a Multiple of 3
(n=33)

	Category		
Score [illegible]	**A** (listing)	**B** (a property of a multiple of 3)	**C*** (insufficient or incorrect)
4	15 %	3 %	6 %
3	15 %	3 %	16 %
2	12 %	3 %	9 %
1	0 %	0 %	18 %
0	0 %	0 %	0 %

(* indicates an insufficient /incorrect strategy)

TABLE 10
Distribution of the Students' Responses for Question (3):
Strategy × Score
Stage 3: Probability
(n=25)

	Strategy		
Score	1 $\left(\frac{48}{144}=\frac{1}{3}\right)$.	2* $\left(6+18=\frac{1}{3}\right)$	3* (intuitively)
4	28 %	0 %	0 %
3	24 %	4 %	0 %
2	0 %	4 %	4 %
1	0 %	0 %	36 %
0	0 %	0 %	0 %

(* indicates an insufficient /incorrect strategy)

TABLE 11
Distribution of the Students' Responses for Question (3):
Categorical Classification
Stage 3: Probability
(n=25)

	Category	
Score	A $\left(\frac{48}{144}=\frac{1}{3}\right)$	B* (intuitive /insufficient)
4	28 %	0 %
3	24 %	4 %
2	0 %	8 %
1	0 %	36 %
0	0 %	0 %

(* indicates an insufficient /incorrect strategy)

Appendix A : Scoring Rubrics

SCORE LEVEL	MATHEMATICAL KNOWLEDGE Knowledge of mathematical principles and concepts which result in a correct solution to a problem	STRATEGIC KNOWLEDGE Identification of important elements of the problem and the use of models, diagrams and symbols to systematically represent and integrate concepts	COMMUNICATION Written explanation and rationale for the solution process
4	* shows complete understanding of the problem's mathematical concepts & principles * uses appropriate mathematical terminology & notation (e.g. labels answer as appropriate)[1] * executes algorithms completely and correctly	* identifies all the important elements of the problem and shows complete understanding of the relationships among elements * reflects an appropriate and systematic strategy for solving the problem * gives clear evidence of a complete and systematic solution process	* gives a complete written explanation of the solution process employed; explanation addresses what was done, and <u>why</u> it was done * if a diagram is appropriate, there is a complete explanation of all the elements in the diagram
3	* shows nearly complete understanding of the problem's mathematical concepts and principles * uses nearly correct mathematical terminology and notations * executes algorithms completely; computations are generally correct but may contain minor errors	* identifies most of the important elements of the problem and shows general understanding of the relationships among them * reflects an appropriate strategy for solving the problem * solution process is nearly complete	* gives a nearly complete written explanation of the solution process employed; may contain some minor gaps * may include a diagram with most of the elements explained
2	* shows some understanding of the problem's mathematical concepts and principles * may contain major computational errors	* identifies some important elements of the problem but shows only limited understanding of the relationships among them * appears to reflect an appropriate strategy but application of strategy is unclear * gives some evidence of a solution process	* gives some explanation of the solution process employed, but communication is vague or difficult to interpret * may include a diagram with some of the elements explained
1	* shows limited to no understanding of the problem's mathematical concepts and principles * may misuse or fail to use mathematical terms * may contain major computational errors	* fails to identify important elements or places too much emphasis on unimportant elements * may reflect an inappropriate strategy for solving the problem * gives minimal evidence of a solution process; process may be difficult to identify * may attempt to use irrelevant outside information	* provides minimal explanation of solution process; may fail to explain or may omit significant parts of the problem * explanation does not match presented solution process * may include minimal discussion of elements in diagram; explanation of significant elements is unclear
0	* no answer attempted	* no apparent strategy	* no written explanation of the solution process is provided

[1] "As appropriate" or "if appropriate" relate to whether or not the specific element is called for in the stem of the item.

Adapted from Lane (1993)

Appendix B : Examples of Reponses for Question (1)

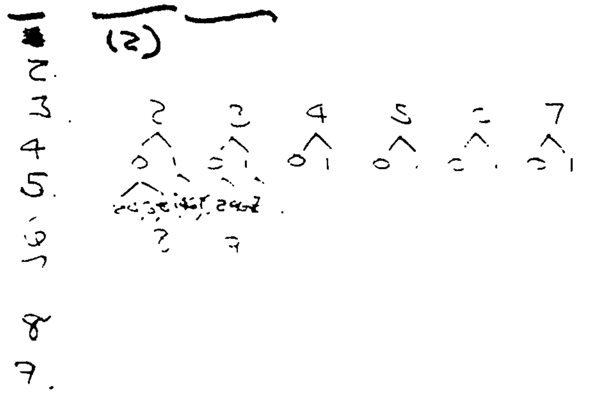
- **Strategy 1 (listing), Level 2**

601, 602, 603, 604, 605, 606, 607, 608, 609, 612, 611, 613
614, 615, 616, 617, 618, 619

Level 2	This student showed understandings of the problem's mathematical concepts. However, she simply listed all cases, and didn't integrate them to answer the question of how many area codes are possible.

- **Strategy 2 (combination with a chart), Level 4**

1) We drew a chart
6 ___ ___
we knew there was 2
possibilities in the first blank
+ 9 possibilities in the Second
So we multiplied these two



- **Strategy 3 (combination without a chart), Level 4**

18

1x2x9 There is 1 possibility for the first number because it has to be 6, 2 for the second (0, 1) and 9 for the 3rd 1-9)
Then multiply 1x2x9 = 18

Appendix C : Examples of Reponses for Question (2)

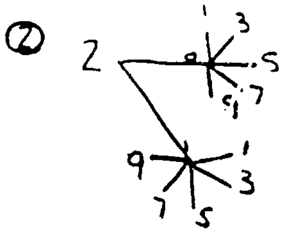
- **Strategy 1 (listing), Level 4**

80

2) How I did this was I wrote out all the possible answers then I went through them and circle all the odd ones.

- **Strategy 4 (combination with a chart), Level 3**

80



10 • 8 = 80

- **Example 1 :**

Strategy 5 (combination without a chart), Level 2

70 I multiplied seven by 2 [illegible] 5 for the this digits. 5

7
2
14
5
70

first I just took the last digit but then I remembered that it was combinations.

Appendix D - 1 : Examples of Reponses for Question (3)

Stage 1: All Possibilities

- **Strategy 5 (combination without a chart), Level 3**

Stage 2 : Strategy 4 (listing the possibilities in 200's and 300's, then $6 \times 8 = 48$)

Stage 3 : Strategy 1 $\left(\frac{48}{144} = \frac{1}{3}\right)$

① 201
② 204
③ 207
④ 213
⑤ 216
6. 219
⑦ 303
⑧ 306
9 309
⑩ 312
⑪ 315
⑫ 318

48 out of 144 → 4 out 12 → 1 to 3

$$\begin{array}{r} 9 \times 8 = 72 \\ \times\ 2 \\ \hline 144 \end{array}$$

I noticed a pattern in that there were 6 #s for every first number. I figured with 8 #s it would be (3)(8) = 48 out of a total of 144, which is 1 : 3.

Appendix D - 2 : Examples of Reponses for Question (3)

Stage 2: The Total Number of a Multiple of a Multiple of 3

- **Strategy 1 (listing), Level 3**

$$\left\{\begin{array}{l}\text{Stage 1 : not performed}\\ \text{Stage 3 : not performed}\end{array}\right\}$$

201
204
207
213
216
219

303
306
309
312
315
318

402
405
408
411
414
417

501
504
507
513
516
519

603
606
609
612
615
618

702
705
708
711
714
717

801
804
807
813
816
819

903
906
909
912
915
918

All multiples of 3 that fit the requirements for the area code form

These are 48 possibilities

- **Strategy 3 (6 x 8 = 48: failure), Level 2**

$$\left\{\begin{array}{l}\text{Stage 1 : Strategy 5 (combination without a chart)}\\ \text{Stage 3 : Strategy 1}\left(\frac{48}{144}=\frac{1}{3}\right)\end{array}\right\}$$

8
12
16
9
144

201
204
207
213
216
219
6

P. 6:144
3:72
1:24

I took the number of possibilities to the number of the ones that is a multiple of 3

BEST COPY AVAILABLE

- **Exapmle 2 :**

Strategy 2 (checking with a property of a multiple of 3), Level 4

$$\left\{\begin{array}{l}\text{Stage 1 : no explanation}\\ \text{Stage 3 : Strategy 1}\left(\dfrac{48}{144}=\dfrac{1}{3}\right)\end{array}\right\}$$

(B) the probability that an area code is 1 out of 3.

There are 48 numbers that are a multiple of 3.

There are 144 possible area codes all together 48/144 = $\frac{1}{3}$

You know that is the sum of the digits of a number equal to a multiple of 3 them that number is a multiple of 3.

603 6+0+3=9÷3=3

2 = 201, 204, 207, 213, 216, 219

3 = 303, 306, 309, 312, 315, 318

4 = 402, 405, 408, 411, 414, 417

5 = 501, 504, 507, 513, 516, 519

6 = 603, 606, 609, 612, 618, 615

7 = 702, 705, 708, 711, 714, 717

8 = 801, 804, 807, 813, 816, 819.

9 = 903, 906, 909, 912, 915, 918.

Notice how 3, 6, and 9 have the #'s of 3, 6, 9, 12, 15, and 18 in their area codes that are multiples of 3.

2, 5, and 8 have the numbers 1, 4, 7, 13, 16, and 19 in common

4 and 7 have the numbers 2, 5, 8, 11, 14, and 17

All the numbers from 2-9 have 6 possible multiples of 3 out of all their possible area codes.

Appendix D - 3 : Examples of Reponses for Question (3)

Stage 3: Probability

- **Strategy 1** $\left(\frac{48}{144}=\frac{1}{3}\right)$**, Level 4**

$$\begin{Bmatrix}\text{Stage 1: Strategy 2 } (18\times 8=144 \text{ using the result of (1))}\\ \text{Stage 2: Strategy 1 (listing)}\end{Bmatrix}$$

201
204
207
303
306
309
402
405
408
501
504
507

603
606
609
702
705
708
801
804
807
903
906
909

213
216
219
312
315
318
411
414
417
513
516
519

612
615
618
711
714
717
813
816
819
912
915
918

$\frac{48 \text{ multiples of } 3}{144 \text{ numbers}}$

Probability of an area code that is a multiple of 3 = $\frac{48}{144}=\frac{4}{12}=\frac{1}{3}$

$\boxed{\text{1 out of 3}}$

To find the answer to this problem, I started out by writing out all the area codes that were divisible by three and I got 48. I then used the number I got from the first problem (18) and multiplied it by 8 to get all the number of all possible area codes. I then put 48 over 144 because probibility is the # favorable ove the number total, and reduced it and I got $\boxed{\text{1 out of 3}}$

- **Strategy 2** $\left(\frac{6}{18}=\frac{1}{3}\right)$**, Level 2**

$$\begin{Bmatrix}\text{Stage 1: not performed}\\ \text{Stage 2: Strategy 4 (listing the possibilities in 600s: partially performed)}\end{Bmatrix}$$

603, 606, 609, 612, 615, 618,

$\frac{6}{18}$ $\boxed{\frac{1}{3}}$ 1;3