

National Assessment of Educational Progress (NAEP): Eighth-Grade Mathematics

PURPOSE OF THE ASSESSMENT

The NAEP mathematics assessments are nationally administered tests designed to assess what students should know and should be able to do in mathematics in 8th grade. Governed by the National Center for Education Statistics (NCES), the resulting reports focus on the strengths and weaknesses in students' understanding on 8th grade-appropriate mathematics and the ability of students to apply their understanding of math in problem-solving situations. Group data is reported by geographic region of the country (Northeast, Southeast, Central, and West), race/ethnicity, type of community, and examines trends over time to evaluate students' progress and report relationships between student proficiency and various background variables. Reports disclose state, regional, and national results. This information is made available (in the report, *The Nation's Report Card*) to citizens, school administrators, curriculum specialists, teachers, and policymakers (at the national, state, and local levels) and is used to ascertain whether progress is being made toward national educational goals.

Beginning in 1973, this test has been nationally administered every four years and a long-term trend assessment is also administered nationally to examine long-term trends. The long-term trend assessment uses instruments that were developed in the mid-1980s and is administered in a form identical to the original one, allowing NAEP to measure trends from 1973 to the present. Beginning in 1999, the long-term trend assessment will be administered on a four-year schedule and in different years from the main assessments.

HISTORY AND INTENTION OF THE FRAMEWORK OF THE ASSESSMENT

In 1988, Congress created the National Assessment Governing Board (NAGB) and charged it with developing the objectives and test specifications for the assessment. It was mandated that this be achieved through a national consensus approach to identify the appropriate achievement goals for each age and grade. NAGB awarded a contract to the Council of Chief State School Officers (CCSSO) to design a framework for the assessment. To achieve this task, the CCSSO focused their attention on the objectives and frameworks as were specified at the state level; examined curricular frameworks at the state-, district-, and school-levels; and consulted with leaders in the mathematics educational field.

In 1991, The College Board was awarded a contract to develop the assessment and item specifications. They were charged with creating a structure for describing what students should know and be able to do in mathematics and develop specifications of the items, paying close attention to the mix of item types, the distribution of the content areas, and the testing conditions under which the test was administered. In order to attain consensus on the test specifications, a national mail review was conducted and The College Board convened focus groups in several states to gather input on their committees' recommendations.

Once the framework was agreed upon, the NAEP planning committee chose a model that was to use the following five major content strands:

1. Number Sense, Properties, and Operations
2. Measurement
3. Geometry and Spatial Sense
4. Data Analysis, Statistics, and Probability
5. Algebra and Functions

These five strands were reflective of the recommendations made by the National Council of Teachers of Mathematics. The framework specified the percentage of items in each of the content strands. This model is reflective of traditional skills but also of broad algebra- and geometry-oriented skills at the 8th grade level.

Families of tasks and items were created to probe nature of the students' understanding and their depth of understanding across content strands. These families can also cross content areas as a means to assess students' strengths and/or weaknesses. Constructed-response items are included in this model as a means to see students' abilities to reason, connect, and communicate their knowledge of mathematics. Manipulative materials such as rulers or protractors are provided to measure students' ability to represent their understanding in geometry-based items. Figures 1a and 1b collectively represent an example of a family of items. Table 1 summarizes the Content Strands, the aspect of the Content Strand being measured, and the intended Mathematical Abilities expected to be utilized by students solving the items in this family.

Table 1

Item No.	Content Strand	Aspect of Content Strand	Mathematical Ability
1	Geometry and Spatial Sense	Subtopic of describing, visualizing, drawing, and constructing geometric figures	Conceptual Understanding
2	Geometry and Spatial Sense	Subtopic of investigating and predicting results by combining, subdividing, and changing shapes (i.e., paper-folding, rearranging pieces of solids)	Problem Solving
3	Geometry and Spatial Sense	Subtopic of identifying the relationship between a figure and its image under a transformation (i.e., lines of symmetry, flips, turns, and slides)	Problem Solving
4	Geometry and Spatial Sense	Subtopic of identifying the relationship between a figure and its image under a transformation (i.e., lines of symmetry, flips, turns, and slides)	Problem Solving
5	Measurement	Subtopic of estimating, calculating, and comparing perimeter, area, volume, and surface area	Problem Solving
6	Measurement	Subtopic of estimating, calculating, and comparing perimeter, area, volume, and surface area	Problem Solving
7	Data Analysis, Statistics, and Probability	Subtopic of reading, interpreting, and making predictions using tables and graphs.	Problem Solving

Figure 1a

With this test booklet, you will receive a packet of 6 pieces: 2 each of shape N , shape P , and shape Q . You will use these pieces in answering some of the questions. You can turn the pieces in any way or flip them over. You may use drawings to help explain your answers.

1. Laura was asked to choose 1 of the 3 shapes N , P , and Q that is different from the other 2. Laura chose shape N . Explain how shape N is different from shapes P and Q .

Answer:

2. You will need the 2 pieces labeled Q . Please find those 2 pieces now.

Use the 2 pieces labeled Q to make a square. Trace the square and draw the line to show where the 2 pieces meet.

3. Use the 2 pieces labeled Q to make a 4-sided shape that is not a square. Trace the shape and draw the line to show where the 2 pieces meet.

4. For this question you will need some of the pieces labeled N , P , and Q .

Use 4 of the 6 pieces labeled N , P , and Q to make the shape shown below. Draw the lines to show where the pieces meet and label the pieces.

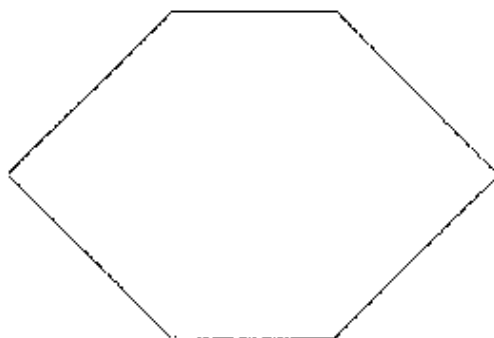


Figure 1b

5. Bob, Carmen, and Tyler were comparing the areas of N and P . Bob said that N and P have the same area. Carmen said that the area of N is larger. Tyler said that the area of P is larger.

Who was correct? _____

Use words or pictures (or both) to explain why.

6. Which of the shapes N , P , and Q has the longest perimeter (distance around)?

Shape with longest perimeter: _____

Use words or pictures (or both) to explain why.

7. This question refers to pieces N , P , and Q .

In Mr. Bell's classes, the students voted for their favorite shape for a symbol. Here are the results.

	Class 1	Class 2	Class 3
Shape N	9	14	11
Shape P	1	9	17
Shape Q	22	7	2

Using the information in the chart, Mr. Bell must select one of the shapes to be the symbol. Which one should he select and why?

The shape Mr. Bell should select: _____

Explain:

DELIVERY OF THE ASSESSMENT

NAEP assessments are currently delivered exclusively in paper-and-pencil form. However, research studies are currently being conducted (at ETS) to ascertain what math skills can and cannot be effectively assessed on a computer, how students from different population groups perform on computer versus paper-and-pencil tests, what the costs of paper versus computer delivery are, and what logistical issues are associated with computer delivery. The platform being used will be delivery via the Internet.

SCORING OF THE ASSESSMENT

Evidence of students' performance is separated into three levels of achievement: basic, proficient, and advanced. These ratings are used throughout all NAEP assessments as a means of maintaining continuity throughout all subjects and grade-levels. Students with an achievement level of basic have a partial mastery of the prerequisite knowledge and skills that are fundamental for proficient work in 8th grade. Students with an achievement level of proficient demonstrate a solid mastery of more challenging subject matter, can apply that knowledge to real-world situations, and can demonstrate the analytical skills necessary to solve the mathematical situations presented. Students whose performance is rated as advanced demonstrate superior performance in all grade-appropriate material presented. This reflects that the students' proficiency is assessed over a broad range of mathematical ideas and skills and how students make the connections using those skills.

ADMINISTRATION OF THE ASSESSMENT

Because NAEP is a large group assessment, each student takes only a small part of the overall assessment. In most schools, only a small number of the total grade enrollment is selected to take the assessment and these students may not reliably or validly represent the total school population. Only when the student scores are aggregated are the data considered to be reliable and valid estimates of what students know and can do in the content area.

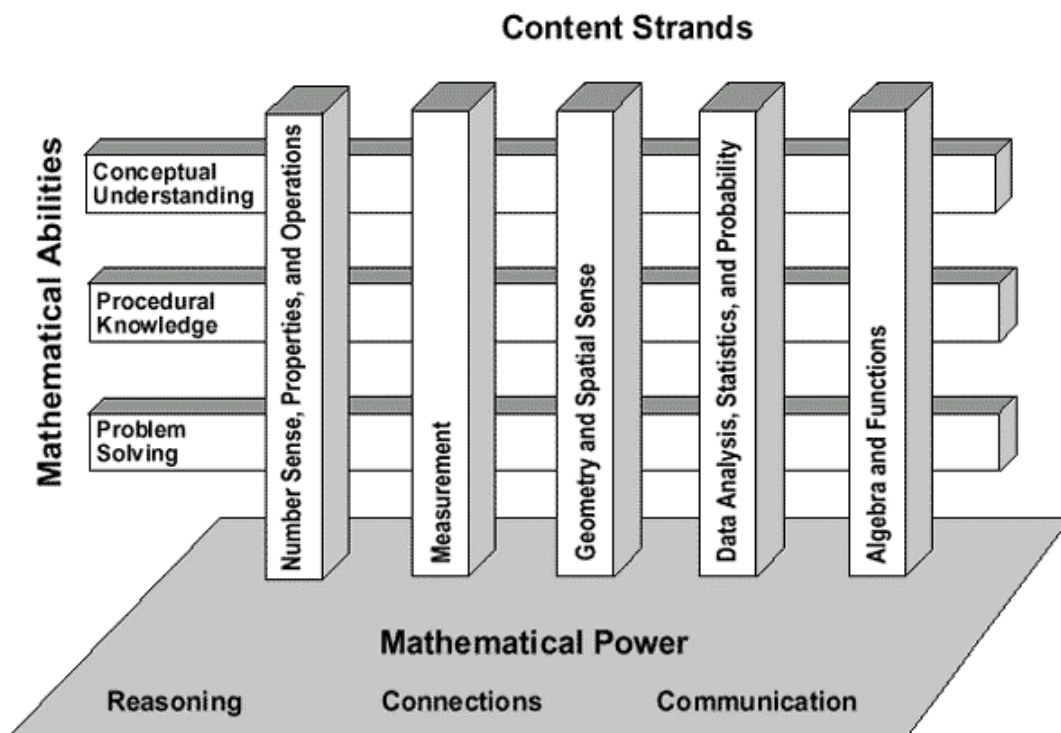
Westat is the current data collection contractor and is comprised of a national network of educators trained to collect assessment data in a uniform manner. This staff is responsible for all assessment activities including scheduling, drawing the random sample of students to be assessed, providing materials for the assessment, administering the assessment, and shipping test materials to the scoring facility.

THE FRAMEWORK

The following diagram, taken from the publication, *Mathematics Framework for the 1996 and 2000 National Assessment of Educational Progress*, illustrates the skills and behaviors that the developers of the framework sought to measure as well as how each of these components are interrelated.

Figure 2

Mathematical Framework for the 1996 and 2000 Assessments



All behaviors to be measured by the assessment are grouped into one of three areas: Content Strands, Mathematical Abilities, and Mathematical Power. Every item in the assessment is intended to not only measure the students' abilities in one of the content areas but is also intended to measure attributes in found in either (or both) of the other two groupings. Further, it was the opinion of the frameworks committee that student proficiency in mathematics was the result of a broad experience in forming a network of connections (schemas and mental models) among the various mathematical ideas and skills. Thus, as an example, they felt that it could be implied that an item that measured a student's ability in algebra could at the same time measure his/her skill in problem solving.

THE DIMENSIONS OF THE FRAMEWORK AND THE PSYCHOLOGICAL PERSPECTIVES

As mentioned previously, the first grouping, **Content Strands**, was designed to measure students' proficiencies in five mathematical content areas:

- Number Sense, Properties, and Operations
- Measurement
- Geometry and Spatial Sense
- Data Analysis, Statistics, and Probability
- Algebra and Functions

Clearly, this grouping was designed to measure specific mathematical constructs displayed by the student. This grouping is concerned with the student outcomes (i.e., number of test items answered correctly), and fits into the trait psychology perspective.

The second grouping, **Mathematical Abilities**, is concerned with the general mental abilities associated with mathematics, which are separated into three categories:

- Conceptual Understanding
- Procedural Knowledge
- Problem Solving

This grouping utilizes the cognitive psychological perspective as it is concerned with the cognitive abilities of students, i.e., the student's ability to recognize and understand what an item is asking, knowing how to approach the situation put forth before him/her (by using the appropriate mental models and schemas that they possess), solving the problem put forth in the item, and reflecting upon the solution that he/she derived. It was intended that students' demonstration of these types of thinking could be evoked by their demonstration of whether or not they answered multiple-choice items correctly and through their responses to constructed-response items. However, in reality, these types of behaviors cannot be measured directly by the work products produced by students in this assessment. This will be further discussed later in this paper.

The third grouping, **Mathematical Power**, also utilizes both cognitive and situative psychological perspectives, examining the students' abilities to reason in mathematical situations; connect that information with related mathematical knowledge and information learned from other disciplines; and communicate the perceptions and conclusions drawn from that mathematical situation. Students' ability to attempt to solve a problem by way of connecting it to a similar context or re-evaluating the way to solve a problem (with a known solution) a new way is examined as well as the process a student utilizes when he/she is unsuccessful at solving the problem (i.e., arrives at the wrong solution) and the way he/she attempts to rework the problem in a more productive fashion. The constructed-response items act as a vehicle for students to communicate their ideas and perceptions to others as they relate to the situations presented in the items and as they are related to mathematics.

Finally, Mathematical Power is intended to be measured through the use of multiple-choice items and through the analysis of the ways in which students develop their responses on the open-ended and extended open-ended items. In sum, Mathematical Power is concerned with the process the student uses to revising his/her approach using reasoning skills, gathering new information, making connections with other types of mathematical ideas, and communicating those ideas.

THE STUDENT MODEL AND STUDENT MODEL VARIABLES

This particular assessment has many student model variables. However, evidence of many of these variables is only implied and cannot be directly measured. The only student model variables measured in the statistical model are those of a very coarse grain-size, i.e., each of the five categories classified as dimensions in the Content Strands. Thus, only overall mathematical classifications are measured and reported. The student model variables found in the Mathematical Abilities are so

highly correlated that they cannot be isolated and measured individually. Further discussion of Mathematical Abilities and Mathematical Power will be discussed in the section about the Statistical Model of the assessment.

Keeping in mind that this NAEP assessment only measures and reports overall mathematic categories, the student model variables found in the following descriptions include those variables that the frameworks committee felt were important and were intended to be measured by the assessment. This is to say most of these variables are not directly evidenced, only inferred.

Content Strand: Number Sense, Properties, and Operations

This strand intends to focus on students' understanding of positive and negative numbers; properties and operations involving whole numbers, fractions, decimals, integers, and rational numbers; estimation, the use of ratios and proportional thinking to represent quantities and their application to real-world situations; and their knowledge of scientific notation to represent small and large numbers. Further, this strand was intended to ascertain a students' understanding of relative size, equivalent forms of number, and their ability to use numbers to represent the attributes of real-world quantities and objects.

Content Strand: Measurement

This strand intends to focus on students' understanding of the process of measurement and the use of numbers and measures to compare mathematical and real-world objects. It implies that students are also expected to understand the concepts of length, mass/weight, volume, surface area, time, money, and temperature.

Content Strand: Geometry and Spatial Sense

This strand intends to focus on students' understanding of spatial relationships and geometry, including an understanding of the properties of angles and polygons. Students are expected to be able to apply their reasoning skills to make and validate conjectures about transformations and combinations of shapes. It is intended that the students must also understand how to extend proportional thinking to similar figures and indirect measurement.

Content Strand: Data Analysis, Statistics, and Probability

This strand intends to focus on students' skills of collecting, organizing, reading, representing, and interpreting data in a variety of contexts to reflect the pervasive use of these skills in dealing with information. It implies that students must understand basic probability concepts and the application of these concepts in problem-solving and decision-making situations as well as have some understanding of sampling and prediction.

Content Strand: Algebra and Functions

This strand intends to focus on students' basic algebra skills including their knowledge of variables and algebraic notation, as well as their understanding of functions as a representational tool in algebra and geometry. It implies that they must also possess an understanding of equations as a modeling tool.

Mathematical Abilities: Conceptual Understanding

Conceptual Understanding intends to reflect students' abilities to reason in settings involving careful application of concept definitions, relations, or representations of either.

Mathematical Abilities: Procedural Knowledge

Procedural Knowledge intends to be reflected by students' use of algorithms to solve specific problems. It was intended to also reflect students' ability to read and produce graphs and tables, execute geometric constructions, and perform non-computational mathematical skills.

Mathematical Power

Mathematical Power is identified as a function of students' prior knowledge and experience and the ability to connect that knowledge in productive ways to new contexts. It tries to ascertain a student's overall ability to gather and use mathematical knowledge through solving non-routine problems; through exploring, conjecturing, and reasoning logically; through communicating about and through mathematics; and through connecting mathematical ideas in one context with mathematical ideas in another context or with ideas from another discipline in the same or related contexts.

THE EVIDENCE MODEL

As previously mentioned, there are many student model variables built into the framework. So although the types of evidence sought have been identified, the assessment in its current state provides little ground to clearly accumulate evidence for many of those variables. This is a result of the implied nature of these variables and their high inter-correlation. Thus, they cannot be isolated and measured individually.

However, the evaluation rules (evidence rules) used to score this assessment are consistent. The observable variables are comprised of students' answers to multiple-choice and constructed-response items.

Responses to all item types are recorded in scannable assessment books that are scanned by an image system at NCS. The data values for the multiple-choice item responses are returned as a numeric code so they can be scored as right or wrong.

Images of the constructed-response items are scanned and saved as digitized computer files that are then distributed among the teams (of human raters) that score them, in an effort to maximize consistency and reliability. Each constructed-response item has a unique scoring guide that identifies the range of possible scores and the criteria to be used to evaluate student responses. Within the scoring guides for the more complex items, are models for awarding partial credit. All raters are trained using a benchmark method of training that includes practice scoring on sample responses. Once training has been concluded, two raters score each student's response to an item, being backread by the table leader to ensure inter-rater reliability. Figure 3 shows an example of a digitized computer file containing a student's response to Item 7, in our family of items (Figure 1b). Figure 4 shows the scoring guide for the same item.

Figure 3

7. This question refers to pieces *N*, *P*, and *Q*.

In Mr. Bell's classes, the students voted for their favorite shape for a symbol. Here are the results.

Using the information in the chart, Mr. Bell must select one of the shapes to be the symbol. Which one should he select and why?

The shape Mr. Bell should select: *N*

Explain:

<i>N</i>	<i>P</i>	<i>Q</i>
$\frac{14}{34}$	$\frac{9}{27}$	$\frac{11}{33}$

more votes

Figure 4

Scoring Guide

Conclusion:

because more students chose it.

because it was first choice in one class and second choice in the other classes.

majority" is acceptable (taken to mean most.) If student says the most classes, do not accept.

Score & Description
Correct Correct responses
Incorrect #3 Piece Q chosen, with an explanation that refers to a number of votes.
Incorrect #2 Piece N chosen, but explanation not given or is inadequate with incorrect computation.
Incorrect #1 Any incorrect response other than those described above.

THE EVIDENCE MODEL AS INTENDED BY THE CURRENT FRAMEWORK

What follows is the identification of evidence that is expected to be

accumulated about the students, as specified by the designers of the framework. It is clear that

many of these constructs reflect the attributes that are expected to be observed. The designers of the framework often use terminology that would imply that students' item responses are a direct reflection of what processes are being conducted in their minds as they provide their solutions. At best, inferences can be made about these processes based on the choices the students make in multiple-choice items, through the documentation of their work process in a constructed-response item, and through the drawings/representations that they use when asked to do so. As we will discuss, these observables don't always provide direct connections to the students' knowledge and abilities, and they often cannot be measured statistically. In addition, the designers of the frameworks and the item writers intend to measure multiple constructs in each item but it is at a very coarse level that these inferences (about Mathematical Abilities and Mathematical Power) can be made. A closer look at this problem will be discussed later in this paper.

Content Strand: Number Sense, Properties, and Operations

Students are expected to demonstrate their understanding of the constructs in the student model, demonstrate they know how to perform basic algorithms, and demonstrate that they can use calculators in appropriate ways.

Content Strand: Measurement

Students are expected to demonstrate their understanding of the constructs involving applications of measurement. They are required to demonstrate their understanding of attributes of various objects, apply measurement concepts, and communicate measurement-related ideas. They must also solve problems involving proportional thinking and perform applications that involve the use of complex measurement formulas. Through items in the Measurement Strand, students are asked to demonstrate the connections of measurement with other strands (e.g., number sense and operations, algebra, and geometry).

Content Strand: Geometry and Spatial Sense

Students are expected to demonstrate their ability to draw informal constructions and justify why they chose to represent their thought in the fashion that was employed. They must demonstrate their reasoning skills within both formal and informal situations, as presented in the items. The demonstration of extending proportional thinking to similar figures and indirect measure evidences that students can make connections with other content strands of mathematics.

Content Strand: Data Analysis, Statistics, and Probability

Students are expected to demonstrate their ability to apply the knowledge and ideas required in this strand. They will be expected to demonstrate that they can analyze statistical claims, design experiments, and use simulations to model real-world situations.

Content Strand: Algebra and Functions

Students are expected to demonstrate their ability to transform and solve number sentences and equations of increasing levels of complexity. They are expected to solve simple equations and

inequalities through a variety of methods, including both graphical and basic algebraic methods. Students are also required to demonstrate their use of open sentences and equations as representational tools.

Mathematical Abilities: Conceptual Understanding

Students' conceptual understanding in mathematics is evidenced when they know and apply facts and definitions; compare, contrast, and integrate related concepts and principles to extend the nature of concepts and principles; recognize, interpret, and apply the signs, symbols, and terms used to represent concepts; interpret assumptions and relations involving concepts in mathematical settings; demonstrate their use of and interrelation of models, diagrams, manipulatives, and varied representations of concepts; and generate examples as well as non-examples of concepts. Conceptual Understanding is expected to be further evidenced in their performance in the production of examples, common or unique representations, or in the way they communicate and manipulate central ideas about the understanding of a concept in a variety of ways.

Mathematical Abilities: Procedural Knowledge

Students' conceptual understanding in mathematics is evidenced when they select and apply appropriate procedures/algorithms for a given situation correctly; verify or justify the correctness of a procedure using concrete models or symbolic methods; extend or modify procedures to deal with factors inherent in problem settings, and describe the results of the algorithm and why that particular procedure was used to render the correct answer to the problem assigned. It is assumed to be further evidenced by whether they select the appropriate procedure for a given task and, if so, how well they execute that procedure.

Mathematical Power

Mathematical power is intended to be evidenced through student performance within a particular content strand at the conceptual, procedural, and problem-solving levels of ability. Students must display their mathematical power through the strategies they employ to solve problems and through the way they reason that that particular path was the correct path to take. This should be evidenced through their own reporting of their thinking, in how they explain their reasoning, and their ability to solve problems (in extended open-ended items). It is important to note that partial credit is awarded to the student if his/her process is correct, even if the final solution that they reach is incorrect. This is clearly demonstrative of a cognitive psychological perspective.

THE STATISTICAL MODEL WITHIN THE CURRENT EVIDENCE MODEL

Nationally representative probability samples of eighth-grade students are selected using a complex multi-stage sampling design that involves sampling schools within selected geographic areas across the country and adding sampling weights to ensure that valid inferences can be drawn between the student samples and the respective populations from which they are drawn. These weights are added to account for the disproportionate concentration of Black and/or Hispanic students in some schools; for students who attend non-public schools; and to account for the lower sampling rates for small schools. Minority students are also over-sampled, to help meet this end. To help ensure adequate sample representations for each state, NAEP provides substitutions for non-participating

public and non-public schools, so that the representation of the student population corresponds to figures from the US Census.

Analyses are conducted to determine the percentages of students who give various responses to each of the cognitive and background questions administered in the test. When computing these percentages, missing responses prior to the last intentional response (on a student's block) are treated as intentional omissions whereas missing responses at the end of the blocks administered are considered "not reached." Therefore, the percentages of responses calculated for each question are based on the number of completed questions, using this rule. For multiple-choice and short constructed-response questions, a lack of response is treated as if the student had not been administered the final question. However, if the last item of the test was an extended constructed-response question, it is treated as if the student intentionally did not answer the question.

Item response theory is used to estimate the average mathematical scale scores. These scores are computed for the nation, for various subgroups of interest within the nation, and for states and territories. The main purpose for this analysis is to provide a common scale on which performance characteristics can be compared for the various subgroups. Because a BIB-spiraling design is utilized, these composite scores are designed to represent the distribution of the performances of the population. (*Note: students are administered blocks of items from any given test and are not administered an adequate number of items to make reliable inferences of their individual performances*). A scale ranging from 0 to 500 was created to report population performances for each content strand, summarizing student performance across all three distinct assessment types (multiple-choice, short constructed-response, and extended constructed response). The scales have a mean of 250 and a standard deviation of 50. A composite scale was also created to render a weighted average across the content strands; where the middle of the scale was targeted for the eighth grade performances (whose scores in 1996 ranged between 231 and 375). It should be noted that the lower range of the overall scale pertains to the fourth-grade assessment and the upper range of the scale targets the twelfth-grade assessment.

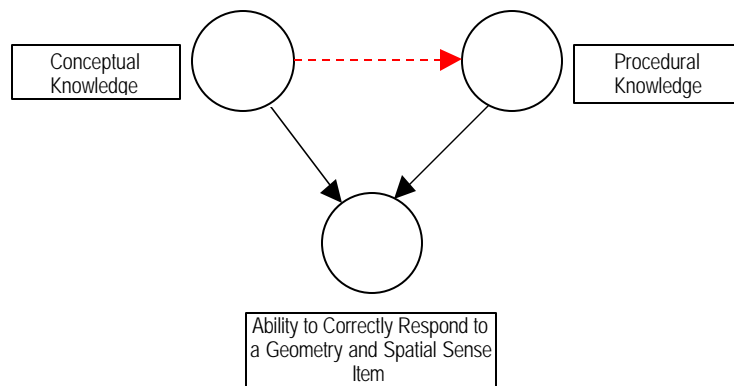
Three different IRT models are used, based on the different item types in the assessment. Multiple-choice items are scaled using the three-parameter logistics model (3PL); short constructed-response items are scaled using the two-parameter logistics model (2PL) for right/wrong responses; and the extended constructed-response questions are rated using the Muraki partial credit model (GPL), thus allowing raters to award partial credit for student answers that may be incorrect with regard to their final solution, but are correct in the cognitive processes employed to answer the question. This presumes that the work products produced by the students in the constructed-response items are rich enough to allow raters to make judgments about the correctness of the solution as well as the quality of the process used by the students to arrive at their solutions. Inherent in this type of scoring, however, is a problem when analyzing the overall scores—there is no means of qualitatively differentiating between students who receive partial credit on an item because they arrived at the right solution without necessarily employing the correct reasoning or procedures versus students who may arrive at an incorrect solution but utilized the correct procedures and reasoning. In terms of Evidence-Centered Design, a single observable score is passed along for analysis (in terms of Evidence Identification) that prevents making qualitative or deeper distinctions among students during Evidence Accumulation.

What is striking about this statistical model is that it demonstrates a departure from the student and evidence models found in the psychological model as set forth by the frameworks committee. The

IRT models employed in the statistical model are, by design, only able to capture students' knowledge within the content strands. The framework committee's intention was to objectively measure students "Mathematical Abilities" (a.k.a. cognitive abilities) as well as their "Mathematical Power" (a.k.a. reasoning and communicative skills). The problem presented here is that the NAEP items may be clear in their assessment of content domain being assessed, yet at the same time, although a student may need to have the conceptual or procedural knowledge to complete an item, there is no independent way of isolating those attributes as a separate score. These scores simply cannot make distinctions among people with different abilities. At the end of this section is a discussion regarding the changes being made to the framework.

In the current model, it seems that even in the presence of the partial-credit model (for extended constructed-response questions), the cognitive processes cannot be independent of each other, e.g., a student must have the conceptual knowledge as well as the procedural knowledge to understand how to solve a complex item. In the case of this item type, partial credit can be given, but only to the extent that the raters have identified that the correct overall cognitive processes have been utilized. In addition, the test developers have taken great effort to not only identify which content strand each item intends to measure, but have also identified the cognitive and communicative attributes the item is intended to measure as well. To this end, it appears that in the statistical model, the best probabilities that one could hope for would be conditional probability distributions where the cognitive factors being considered would be dependent upon each other, thus confounding any clear analysis of these factors. Figure 5 contains a diagram illustrating an example of this thought.

Figure 5



At best, any conditional probabilities that could be calculated, would be calculated to the extent that the overall cognitive abilities could only be measured and, as such, a classification can be made about students in rating their performance as either adequate or inadequate.

In thinking in terms of a Bayes nets, it seems that these cognitive variables may be considered conditionally independent. However, this may not hold true in all situations. If the item type were multiple-choice, a student might be able to provide the correct response because he/she understands it conceptually; but the procedural knowledge variable may not provide additional information about students in this item type. Finally, it appears that it would be difficult to elicit a

performance indicative of these behaviors alone. The scoring model for this assessment allows only for correct/incorrect ratings, and the only allowances made to document these constructs would be in the partial-credit scoring model of extended constructed-response items. If all of the “Mathematical Abilities” and “Mathematical Power” can, at best, be identified as overall concepts, their individual component behaviors cannot be isolated and rated independently by making those inferences inductively. Another problem with trying to isolate these behaviors is that it seems possible that each of these performances (children) may have multiple parents. Throughout all descriptions of the different disclosed items in the different content strands, the test developers have indicated that the items require these behaviors to produce correct responses. Again, the results are confounded.

Frank Jenkins has indicated that the only solution might be to document the evidence through the use of a join tree, by clustering the variables into overlapping cliques. However, I’m not certain that this methodology would be easily applied to the current paper-and-pencil assessment. It seems that it may work better in a computer-delivered environment where information about every keyboard or mouse action can be captured and analyzed to evidence the different cognitive constructs (e.g., procedural knowledge) in an unconfounded manner. However, I wonder shifting to the new “Technically Rich Environments” (i.e., computer test administrations for NAEP) would produce other confounding variables such as expertise of students’ computer skills. In this vein of thought, it seems to me that the proficiency that students possess would vary nationally as well as within each state, based on such factors as the economics of the school district or the financial situation of individual families. Would this added student model variable now confound the content strand variables? Would this new environment move the assessment away from the purpose of the assessment and more toward the computer abilities of the students? Does the question now become, “Are we measuring a students’ content knowledge or their ability to convey their content knowledge through computer-administered tests?”

In this vein of thinking, it seems that what is an already complex assessment design would need to become more complex in terms of what needed to be measured. Would another component need to be added to measure students’ computer abilities? If so, how should poor student outcomes in blocks measuring computer abilities be analyzed? Would poor scores in this area indicate similar performance in the content areas actually being measured? Would delivering an assessment with this added component be done best through a combination of a computer-delivered/paper-and-pencil assessment? And finally, if this is the case, is anything gained or would the costs, logistics, and confounded evidence outweigh the benefits?

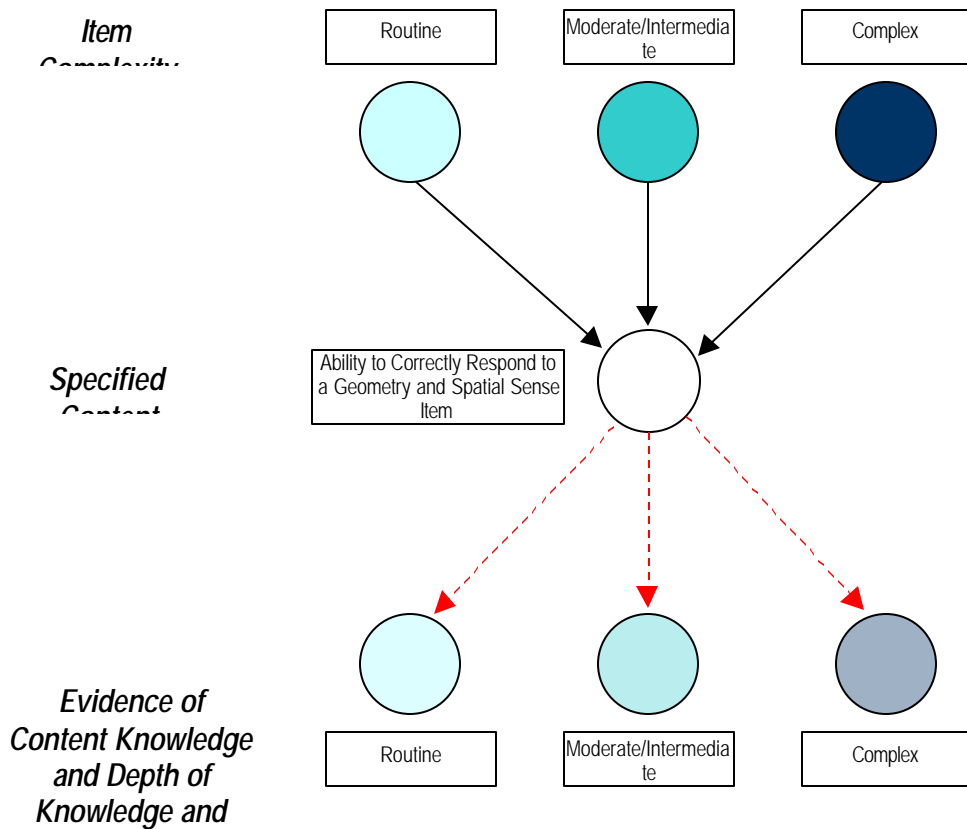
In another study, Brenda Tay-Lim ran an investigation of NAEP variables using factor analysis as well as cluster analysis. She ran her analyses on the NAEP fourth-grade mathematics assessment data. She was trying to ascertain if these types of analyses would reveal the cognitive dimensions underlying the framework of the test. The results of her cluster analysis led her to the conclusion that, “...the cluster method did not identify the underlying content or cognitive processes as specified by the test developers.” The factor analysis led her to the same conclusion. She indicated that this was because the items were not grouped according to content or cognitive processes, but were instead grouped by the difficulty of the mathematical concept and the complexity of the procedure needed to solve the item. Since the difference between the fourth-grade assessment differs from the eighth-grade assessment only in terms of task complexity and deeper content knowledge, her analysis may also apply to the eighth-grade assessment.

The Evidence Model and Refinement of the Framework

To the credit of NAGB, a new frameworks committee has been convened, re-examining the current framework and making recommendations to change the “Mathematical Abilities” to “Levels of Complexity.” This change will maintain the multi-dimensionality of the items but will allow for clearer distinctions between the items in terms of content and complexity. Although the current intention is to continue to report solely on measures of content knowledge, it does set forth the ability to identify and capture new types of evidence. After the evidence accumulation of these new dimensions, they can choose to report (statistically) their findings as these new categories can be defined with clearer boundaries. Further, it is less likely that these new categories would be correlated with each other but can, in fact, lead to stronger evidence of students’ knowledge because as these levels are paired with the content strands, inferences about the depth of students’ knowledge and implied abilities may be able to be made with more confidence, item by item. Given these new pairings, items can be specified in advance (in terms of their location on the scoring continuum) and item writers would have a better ability to write items to those specifications. In turn, evidence of students’ knowledge, abilities, and understanding would be better evidenced in their work products.

Figure 6 takes the previous diagram and illustrates this concept using the new framework dimensions.

Figure 6



There is one drawback that remains with this new framework, however. Coarse grain-size student models would still be evidenced in this model. Specifically, the continued use of the partial-credit model would still render overall scores that could not make distinctions between students of different profiles of skill and abilities. However, it seems that those overall scores would represent a more meaningful representation of students’ knowledge, skills, and understanding than those currently reported.

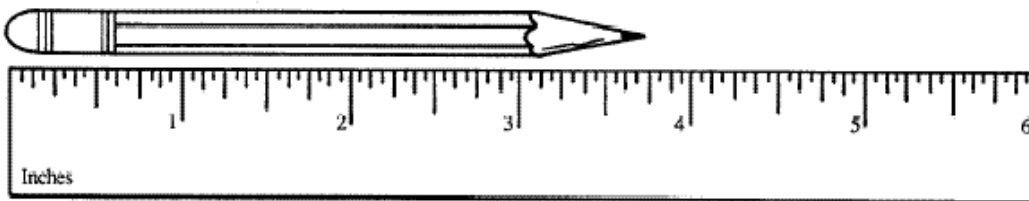
STIMULUS MATERIAL SPECIFICATIONS OF THE TASK MODEL

The NAEP assessment does not have an explicit task model. The stimuli for the task model variables are presented in the form of three different item types: multiple-choice, open-ended, and extended open-ended questions. The test specifications can be classified as the stimulus specifications, indicating the content strand to which of items that should be written, the types of content knowledge that can be solicited in an item, and the types of items that can be used to elicit the desired behaviors.

Examples of Item Types

Figure 7 is a sample multiple-choice item from the Measurement content strand. Not only is it used to evidence a student's understanding of measurement but also at the same time it is intended to provide evidence of a student's procedural knowledge, i.e., whether the student can employ the correct measurement procedures to determine the correct length of the pencil. This is a good example of the confounding of Mathematical Abilities.

Figure 7



3. What is the length of this pencil to the nearest quarter inch?
- A) $3\frac{1}{4}$ inches
 - B) $3\frac{3}{4}$ inches
 - C) $4\frac{1}{4}$ inches
 - D) 4 inches
 - E) I don't know.

Figure 8 is a sample multiple-choice item from the Data Analysis, Statistics, and Probability content strand, testing for the student's understanding of organizing and interpreting the data that is provided to them and making a decision as to which options are feasible versus those that are not. The different options presented in the item are conditionally dependent on the student's understanding of the stimulus and it is through the student's reasoning and problem-solving skills, that he/she must choose not only those options that are possible, but must also decide which seem reasonable and correct in light of their mathematical knowledge and communication skills.

Figure 8

4. A bag contains two red candies and one yellow candy. Kim takes out one candy and eats it, and then Jeff takes out one candy. For each sentence below, fill in the oval to indicate whether it is possible or not possible.

Possible Not Possible

- | | | |
|----------------------------------|-----------------------|---|
| <input checked="" type="radio"/> | <input type="radio"/> | Kim's candy is red and Jeff's candy is red. |
| <input type="radio"/> | <input type="radio"/> | Kim's candy is red and Jeff's candy is yellow. |
| <input type="radio"/> | <input type="radio"/> | Kim's candy is yellow and Jeff's candy is red. |
| <input type="radio"/> | <input type="radio"/> | Kim's candy is yellow and Jeff's candy is yellow. |

Figure 9 is a sample open-ended item from the Geometry and Spatial Sense content strand. This item is intended to also test the student for their problem-solving skills by seeking to evidence whether they can demonstrate the connection of proportional thinking to similar figures and indirect measurement, by placing each of the additional objects in the areas where they belong. Through reasoning, the student must demonstrate that they understand the nature of the geometric shapes that each of additional objects share with the provided objects (sets) as well as using the reasoning skills necessary to judge their solution to be the correct solution. However, this particular item does not seem to provide any more information about the student than a multiple-choice item would as there is no illustration or explication of how they arrived to their final solution.

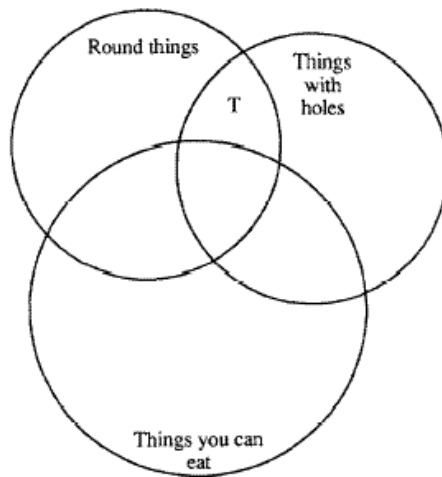
Figure 9

Each circle in the figure below represents the set of objects that have a certain property, as described in the circle. Where an object is placed in the figure depends on which of the properties it has.

Example: Note that the T (for tire) is placed in two of the circles at the same time, but not in the third. That is, it is in the circle that represents round things and the circle that represents things with holes, but it is not in the circle that represents things you can eat.



T - Tire



6. Indicate where each of the following objects should be placed in the diagram by writing its letter in the correct place. Use each letter only once.

		
M - Marble	D - Doughnut	S - Slice of Swiss Cheese

Work Product Specifications

In terms of work product specifications, the three item types require specific types of student responses. Multiple-choice items require students to read the item, reflect upon what it is asking, compute the solution, and choose the option that they believe to be the correct answer. Open-ended items require that students provide a short answer to the question asked. These answers can be a numerical result, a description of a group of mathematical objects, a drawing to illustrate an elicited concept, or a short explanation. Extended open-ended items are those items that may provide a stem containing stimulus information about one or more of the content strands requiring that the student reflect about the stimulus; understand how to solve it; choose and employ a game plan to solve the situation; and interpret the solution in terms of the situation that was provided in the stimulus. This type of problem requires that the student's response provides evidence of their work on the item and communicate their decision making process in the context of the problem.

Further, the work product rendered by the student is his/her answers to the items. This means that students are expected to "bubble-in" their answers in their test booklets for multiple-choice items and either write, draw, or complete an illustration for constructed-response items in their test booklets in the spaces provided following the stimulus.

In addition, students are required to use manipulatives (such as rulers and scientific calculators) to solve items in the Measurement content strand. The actual manipulation of these tools are not used as part of the Evidence Accumulation but instead used to produce the student's solutions to the items (i.e., their work products) in the task model and, in and of themselves, are not work products.

Some items use visual stimuli including different representations of data (such as tables and graphs) are used in the test to evoke students' use of appropriate methods of gathering data (in the Data Analysis, Statistics, and Probability content strand). Items are included that solicit arguments and evaluation of those arguments, based on data analysis. See Item 7 in Figure 1.

More About the Stimulus Specifications

The framework committee recommended that the grade 8 mathematics assessment includes 13 different 15-blocks of multiple-choice and constructed-response (both open-ended and extended open-ended) items. Some of these blocks were unique to grade 8; others overlapped either 4th or 12th grade; and a small percentage of blocks overlapped all three grades. The use of these blocks is for trend analysis.¹

Table 2 indicates the percentage of time that is to be devoted to each content strand by the students as well as the number and percentage of items that fall into those categories.

¹ 1994 National Assessment of Educational Progress Mathematics Assessment Framework and Specifications, The College Board, Contract Number RN 91084001, March 31, 1992, p. 90.
The NAEP 1996 Technical Report, National Center for Education Statistics, US Department of Education, NCES 1999-452, pp. 37-39

Table 2

Content Strand	Percentage of Time	Number of Items (In 1996 Assessment)	Percentage of Items (In 1996 Assessment)
Number Sense, Properties, and Operations	25% -60%	48	29%
Measurement	15%	27	16%
Geometry and Spatial Sense	20%	31	19%
Data Analysis, Statistics, and Probability	15%	25	15%
Algebra and Functions	25%	34	21%
	Total:	165	100%

Table 3 reflects the number of items and percentage of items authored by test developers to measure Mathematical Ability. Remember, these particular distinctions cannot be statistically measured and these are not written as separate items, however, the experienced item writers create items integrate these ideas as part of their authoring process and they deliberately think about what processes, procedural knowledge, and conceptual knowledge the students will need to solve the item correctly.

Table 3

Mathematical Ability	Number of Items (In 1996 Assessment)	Percentage of Items (In 1996 Assessment)
Conceptual Understanding	57	35%
Procedural Knowledge	46	28%
Problem Solving	62	38%
Total:	165	101%

The test specifications have meaning to experienced test developers. They are tied to the National Council of Teachers of Mathematics (NCTM) standards. Following (in Figure 10) is a sample of test specifications for the “Geometry and Spatial Sense” content strand.²

² 1994 National Assessment of Educational Progress Mathematics Assessment Framework and Specifications, The College Board, Contract Number RN 91084001, March 31, 1992, p. 90.

Figure 10

Grade 8, Strand C: Geometry and Spatial Sense

As described in the NCTM standards, spatial sense must be an integral component of the study and assessment of geometry. Understanding spatial relationships allows students to use the dynamic nature of geometry to connect mathematics to their world.

This content area is designed to extend well beyond low-level identification of geometric shapes into transformations and combinations of those shapes. Informal constructions and demonstrations, including *drawing* representations, along with their justifications, take precedence over more traditional types of compass-and-straightedge constructions and proofs. While reasoning is addressed throughout all of the content areas, this content area continues to lend itself to the demonstration of reasoning within both formal and informal settings. The extension of proportional thinking to similar figures and indirect measurement is an important connection here.

In grade 8, students are expected to have extended their understanding to include properties of angles and polygons, and they apply reasoning skills to make and validate conjectures about transformations and combinations of shapes.

C1. Topic: Describe, visualize, draw, and construct geometric figures

a. Subtopic: Draw or sketch a figure given a verbal description [open-ended items]

b. Subtopic: Given a figure, write a verbal description of its geometric qualities

Specific: Given a verbal description of a figure or a design composed of figures, draw or sketch the figure(s). (**O**, **C**onceptual, **C**ommunication)

Specific: Using appropriate geometric language, write a description of a figure or a picture composed of geometric figures. (**E**, **C**onceptual, **C**ommunication)

Specific: Identify a property that is not shared by every rectangle (**O**, **C**onceptual, **R**easoning)

(Note: *O* – Open-Ended Item; *E* – Extended Open-Ended Item)

It is clear that these specifications may have meaning to veteran NAEP mathematics test developers, but due to the short development time required to create items for each test administration, there is no time to create documentation (or a primer) for stimulus material specifications. Instead, panels of 15 teachers and teacher-educators are “taught” how to author items in workshops. Prior to the workshops, the panelists are asked to write some sample mathematics items. In the workshop, they are then taught the mechanics of item writing. As a group, they critique the items that they brought with them. Next, they pair off and try to write new items, incorporating the concepts taught in the workshop. At the end of the workshop, the items are given to ETS and the teachers are instructed to write more items at home. The items produced by these “outside” item writers then go through the internal review process at ETS to ensure that they are representative of the different content

strands, are fair and unbiased, and are grade-appropriate. As such, these principles are not embedded in schemas and the training of new item writers is “learned by example.”

Conditions for Examinee Interaction with the Assessment

Currently, students are administered the assessment in the classroom. In order to reliably test the items specified by the framework, over 100 items are developed for the assessment (160 items were developed for the 2000 assessment). In order to keep student motivation optimal, matrix sampling (specifically, BIB spiraling) is used to ensure that the testing time be kept to a reasonable amount of time for the student; that all content areas of the assessment are being administered; and that the effects of the location of the block within each booklet is minimized. Here, the items are divided into different booklets and administered to different, but equivalent, samples of students. Within each booklet, the different cognitive questions are grouped into collections of separately timed blocks. These blocks and their sequences can vary between each version of the test booklet. This type of spiraling requires that many different booklets must be printed and reduces the likelihood that students will be seated within viewing distance of someone with an identical booklet.

Students are also administered background questions, providing NAEP with the demographic information that it needs for analysis.

DISCUSSION OF THE CURRENT AND FUTURE FEATURES OF THE ASSESSMENT AND CHANGES THAT ARE PROPOSED FOR THE FRAMEWORK FOR THE 2002 ASSESSMENT.

When the frameworks committee developed following framework, it was with the intention that all of the constructs to be measured by the assessment would be grouped into one of three areas: Content Strands, Mathematical Abilities, and Mathematical Power. The frameworks committee decided that every item in the assessment should not only measure the students’ abilities in one of the content areas but also intended that it measure attributes found in either (or both) of the other two groupings. Further, it was the opinion of the frameworks committee that student proficiency in mathematics was the result of a broad experience in forming a network of connections (schemas and mental models) among the various mathematical ideas and skills. However, as the statistical model of this assessment has demonstrated, only the constructs identified in the first grouping (Content Strands) can directly be observed and measured. Direct observations cannot be made about the other two groupings (Mathematical Abilities and Mathematical Power); they can only be inferred or implied and we can only assume that students do well in these areas. Statistically, they cannot be measured.

What follows is a description of the three groupings, as designed by the Frameworks Committee (NAGB). I will also discuss any shortcomings of the framework as it relates to each group.

The first grouping, Content Strands, measures students’ proficiencies in five mathematical content areas:

- Number Sense, Properties, and Operations
- Measurement
- Geometry and Spatial Sense
- Data Analysis, Statistics, and Probability

- Algebra and Functions

Clearly, this grouping is designed to measure specific mathematical constructs displayed by the student. In this grouping, the evidence rules are concerned with whether the student answers the multiple-choice questions right or wrong, or whether the student can either partially or fully answer the open-ended questions correctly (to demonstrate partial knowledge or procedural knowledge). However, the distinctions between individual student-performances cannot be made based on the partial credit models, since the score produced is in terms of a final score – there are no diagnostics that reflect the type of knowledge that individuals may display. Again, as stressed in the statistical model, the assessment is interested in group-results, not in individual performances.

The combination of the automated scoring of the multiple-choice items and (human) rater judgments of both types of constructed-response items connects the features (i.e., the items) of the assessment with the statistical model (found within the evidence model). In terms of Evidence-Centered Design, the student responses serve as the work products; the scoring results serve as Evidence Identification. The multiple-choice items provide evidence of students' factual knowledge. Indirect evidence of problem-solving and procedural skills is more easily identified in the constructed-response items.

It is important to remember that this assessment is concerned with population performances. It uses a student-level IRT model although student-level scores are not computed. Conditional probabilities are rendered, given the values of the thetas. The group distributions and grouped patterns of responses include a link back to the student-level parameters but the conditional probabilities are not used. In this case, individuals' proficiencies and performances as they relate to the student model variables are not computed. However, this assessment can relate the group and sub-group proficiencies as they relate to the student model variables found in the Content Strands.

The second grouping, Mathematical Abilities, is concerned with the general mental abilities associated with mathematics, which are separated into three categories:

- Conceptual Understanding
- Procedural Knowledge
- Problem Solving

The frameworks committee was concerned with the cognitive abilities of students, i.e., the student's ability to recognize and understand what an item is asking, knowing how to approach the situation put forth before him/her (by using the appropriate mental models and schemas that they possess), solving the problem put forth in the item, and reflecting upon the solution that he/she derived. However, inherent in these constructs is the impossibility of isolating them as a separate measurement in a mathematical item, so in order to get an item correct, it is implied that the Mathematical Abilities required to solve the item correctly must also be good.

A few years ago, NAEP held a study comprised of teachers and teacher-educators to discover whether the Mathematical Abilities could, in fact, be directly isolated and observed from a NAEP assessment. What they discovered was that these experts highly agreed in their Content Strand classifications. In contrast, they could not reach agreement when trying to classify the items in the individual Mathematical Abilities categories. They concluded that the Mathematical Abilities

constructs were too closely correlated with each other to be treated as separate behaviors (i.e., student model variables). Further, this assessment is measuring a student's knowledge and abilities at a given point in time – it cannot make distinctions between the types of abilities the student is utilizing, be it recall, recognition of a certain type of problem, or purely creative thinking.

As a result of these findings, a new framework is being constructed replacing the current Mathematical Abilities with “Levels of Complexity.” The committee charged with developing these strands is comprised by a more complete panel of experts from Chief Council of State School Officers (CCSSO) and the Council for Basic Education (CBE). In addition, test development experts from Educational Testing Service (ETS) act as observers to these meetings, to ensure that better connections are made between the proposed framework and the specifications of the items to be authored. In their draft classifications, the mathematics items will be viewed in terms of their rigor of mathematics: “Routine,” “Moderate/Intermediate,” or “Complex.” It was the opinion of the committee that these dimensions could be more easily defined and agreed upon.

The current thinking is that statistics of these constructs will probably not be reported until they are proven psychometrically reliable. However, perhaps through an analysis of the Content Strand IRT parameters where items are lined up on a continuum in terms of difficulty, the complexities of the items will also line up in a similar fashion. If this happens, inferences may be made to better understand what a student of that grade level would likely be able to solve.

The third grouping, Mathematical Power is intended to examine the students' abilities to reason in mathematical situations; connect that information with related mathematical knowledge and information learned from other disciplines; and communicate the perceptions and conclusions drawn from that mathematical situation. Students' ability to attempt to solve a problem by way of connecting it to a similar real-world contexts or re-evaluating the way to solve a problem a new way (using a known solution) is examined as well as the process a student utilizes when he/she is unsuccessful at solving the problem (i.e., when he/she arrives at the wrong solution) and the way he/she attempts to rework the problem in a more productive fashion.

Mathematical Power cannot be directly observed but evidence of strength of it can be inferred through the use of multiple-choice items and through the analysis of the ways in which students develop their responses on the open-ended and extended open-ended items. It is concerned with the process the student uses to revising his/her approach using reasoning skills, gathering new information, and making connections with other types of mathematical ideas.

In 1996, the frameworks committee (as an afterthought) added the Mathematical Power grouping. The committee felt that the identification of these conceptual student model variables would add richness to the assessment that the Mathematical Abilities could not. Conceptually, Mathematical Power has helped ensure that when the items are developed, there are a certain number of connections within each discipline of mathematics connecting the mathematics to real-world situations and problems. It is best communicated through the use of constructed-response items. When the items are developed, the Mathematical Power ensures that test developers pay attention to the goal of the items. However, like the Mathematical Abilities, these constructs cannot be statistically measured.

Concluding Remarks

It seems that the frameworks committee is moving in the direction of identifying more salient task model features by proposing the changes in groupings from “Mathematical Abilities” to “Levels of Complexity.” Whether these new categorizations will prove to be reliable measures and seemingly not highly correlated should be demonstrated on subsequent administrations.

Clearly, a suggested improvement might be in the area of documenting the work specifications for future item writers. However, it is also clear that this assessment will continue to examine group performances, so the grain-size of what we can infer is of a courser nature than those assessments and tutoring systems used to diagnose individuals. This is clearly consistent with the purpose of the assessment – to examine the performances of populations, not individuals, and make inferences about those groups.