DOCUMENT RESUME

ED 464 917 TM 033 847

AUTHOR Schulz, E. Matthew; Lee, Won-Chan

TITLE Describing NAEP Achievement Levels with Multiple Domain Scores.

PUB DATE 2002-04-00

NOTE 81p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (New Orleans, LA, April 2-4, 2002).

PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)

EDRS PRICE MF01/PC04 Plus Postage.

DESCRIPTORS *Academic Achievement; Achievement Tests; Elementary Education; *Mathematics Tests; *Scores

IDENTIFIERS Domain Analysis; *National Assessment of Educational Progress

ABSTRACT

This study was conducted to demonstrate the potential for using multiple domains to describe achievement levels in the National Assessment of Educational Progress (NAEP) mathematics test. Mathematics items from the NAEP grade 8 assessment for the year 2000 were used. Curriculum experts provided ratings of when the skills required to answer the items correctly were introduced and when they were mastered in a standard mathematics curriculum. The ratings were used to define two or three domains within each content strand in the test framework (number sense, measurement, geometry, data analysis, and algebra). Characteristic curves based on these domains tended to be well spaced and noncrossing as a function of the NAEP mathematics scale score. Results support Guttman-style interpretations of what students at a given level of achievement can or cannot be expected to do in terms of scoring higher or lower than a percentage correct criterion score that may be chosen to define mastery of a domain. On the basis of these results, it is concluded that achievement-level descriptions based on multiple domain scores would be useful in NAEP and similar broad-based assessments. Four appendixes contain a discussion of the computation of domain characteristic curves, instructions and forms for rating instructional timing, descriptions of the domains, and number sense test item examples. (Contains 11 tables, 14 figures, and 32 references.) (SLD)

Describing NAEP Achievement Levels with Multiple Domain Scores

E. Matthew Schulz and Won-Chan Lee

ACT Inc.

2255 North Dubuque Road
P. O. Box 168
Iowa City, IA 52243-0168
(319) 337-1468
schulz@act.org

Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, April 2-4, 2002.

Abstract

This study was conducted to demonstrate the potential of using multiple domains to describe achievement levels in the National Assessment of Educational Progress (NAEP) Mathematics test. Mathematics items from the NAEP Grade 8 assessment for the year 2000 were used. Curriculum experts provided ratings of when the skills required to answer the items correctly were introduced, and when they were mastered in a standard mathematics curriculum. The ratings were used to define two or three domains within each content strand in the test framework (Number Sense, Measurement, Geometry, Data Analysis, and Algebra). Characteristic curves based on these domains tended to be well spaced and noncrossing as a function of the NAEP Mathematics scale score. Our results support Gutttman-style interpretations of what students at a given level of achievement can or cannot be expected to do—score higher or lower than a percentage correct criterion score that one may choose to define mastery of a domain. On the basis of these results, we conclude that achievement-level descriptions based on multiple domain scores would be useful in NAEP and similar broad-based assessments.

Describing NAEP Achievement Levels with Multiple Domain Scores

A very popular idea associated with the definition of achievement levels in educational assessments is the Guttman-notion that higher-level students can do whatever lower-level students can do, plus at least one more thing. Unfortunately, it is extremely challenging to define clearly just what the "things" are that students at different levels of achievement can or cannot do (Forsyth, 1991). The purpose of the present study is to suggest one way this particular objective might be achieved.

Guttman scales clearly define the meaning of score levels (Guttman, 1950). In a Guttman scale, there are as many levels as there are binary items or rating scale categories in the assessment. The marginal (overall) difficulty of the items represents the difficulty order for every examinee. The content of the items and their difficulty convey a clear understanding of what level scores mean. Applied to educational achievement, examinees are understood to have mastery of levels up to and including their reported level of achievement, and nonmastery of higher levels. Mastery of a level implies a particular score on the item and on the content domain represented by the item. Kolstad (1996) referred to a Guttman scale in describing the kinds of interpretations that tasks in the National Adult Literacy Survey were designed to support.

Traditional Guttman scaling cannot be used in educational assessments because the items in these assessments have relatively large amounts of measurement error. Traditional Guttman scaling is limited to variables like physical functioning or disability, in which mastery of a level can be reliably assessed with only one item or rating. [For examples of Guttman scales see Boulton-Lewis (1987), Fox and Tipps (1995), and Katz and Akpom (1976).]

In NAEP, there are hundreds of items, but only four levels of achievement—Below Basic, Basic, Proficient, and Advanced. This presents a challenge for making general inferences about achievement levels. Although items have not been assigned to achievement levels in NAEP, there have been Guttman-like attempts to describe levels through a post-hoc examination of the relative difficulty of individual items. Content experts have attempted to form generalizations about the things students at different levels of achievement can or cannot do by examining the items that discriminate between achievement levels (National Center for Educational Statistics, 1992). Because there is no strong relationship between item content and discrimination or difficulty, post-hoc descriptions based on item statistics tend to produce a loosely defined mixture of processes, content, principles, etc. Also, statements about what examinees can or cannot do based on specific "exemplar" items can be contradicted with reference to other, similar items (Forsyth, 1991).

Other methods have attempted to incorporate features of Guttman scaling more strictly. In Wilson (1989), and Masters, Adams, and Lokan (1994), each item in a test was associated with a level and the item difficulty ranges of levels did not overlap. IRT models that specified noncrossing IRFs (Rasch models) were used. In Wilson (1989) items were assigned to levels first based on their content. Lower and upper boundaries of levels on an IRT scale were defined with respect to a 0.8 probability of mastering, respectively, the easiest and hardest items within each level. In Masters, et al. (1994), boundaries for levels on an IRT scale were established by studying the spatial arrangement of items calibrated to the scale (an item map). Boundaries were drawn so that items with similar content were in the same level.

The approach in this study is to incorporate Guttman scaling notions into NAEP achievement level descriptions in a way that relies less heavily on individual item statistics. The relationship between item content and any statistic, such as difficulty or discrimination is, like the item score itself, unreliable and difficult to predict. Teachers, for example, cannot predict item p-values very accurately (Impara & Plake, 1998). The notion that an item's difficulty should determine what achievement level it belongs to, as in the Masters et al. (1994) and Wilson (1989) methods, may be too restrictive for many, potentially meaningful definitions of achievement levels. The domains of addition, subtraction, multiplication, and division items, for example, might represent meaningful achievement levels, but the difficulty ranges of their representative items will surely overlap.

Domain Scores

Hambleton, Swaminathan, Algina, and Coulson (1978) referred to domain scores in the context of criterion-referenced testing and reviewed then current methods of estimating domain scores. The domain score was the examinee's "true proportion correct score" or expected proportion correct score if all of the items in the domain of items measuring an objective had been administered. These authors further discussed the concept of allocating examinees to mastery/nonmastery status based on domain score estimates and a given "cut-off score".

More recently, IRT methods have been brought to bear on estimating domain scores (Bock, 1997; Bock, Thissen, & Zimowski, 1997). In order to estimate a domain score with IRT it is necessary that a) there exists a probability sample of the domain in the form of items and item weights, b) the item data fits an IRT model, and c) one has reasonably good estimates of the parameters of the model. IRT-estimated domain scores have been shown to be more accurate predictors of performance on a finite domain than were percentage correct scores on a test (Bock, et al., 1997).

Domain scores have been proposed as a means of describing, and possibly defining, levels of achievement in NAEP (Mislevy, 1998; Pommerich, Nicewander, and Hanson, 1999). Referring to representative samples of NAEP items as market-baskets, Mislevy (1998) notes that any point on the scale can be described as an estimated score on a domain, and that achievement levels can be described by domain scores. Pommerich, Nicewander, and Hanson (1999) evaluated various methods of computing average domain scores for groups of examinees, and commented on potential applications to NAEP.

Multiple domains within a test

There are precedents in the literature where test items have been classified into multiple domains, with each domain being more content-specific than the test as a whole (Schulz, Perlman, Rice, & Wright, 1992; Schulz, Kolen, & Nicewander, 1999; Janssen, Tuerlinckx, Meulders, & De Boeck, 2000). Key features of these studies are:

1) a domain is defined as a subset of items representing more specific content within the broader item pool of a test;
2) items are assigned to domains by content, not statistical, criteria; and
3) expected domain scores, not item true scores, define what examinees at different levels of achievement can or cannot be expected to do.

With these features, domains do not necessarily represent distinct statistical factors in a test. In all of these studies, the test items were calibrated, and expected domain scores estimated, with a unidimensional IRT model. In two (Schulz, et al., 1999; Janssen, et al., 2000), expected domain scores are used, or proposed, for assigning examinees to achievement levels or for making mastery/nonmastery decisions about the examinees with regard to attainment targets represented by domains.

Plots of Domain Characteristic Curves

A domain characteristic curve (DCC) in this study is similar to a test characteristic curve. It is the expected proportion correct on a set of items that belong to a particular domain. As used in this study, "domain characteristic curve" refers to expected performance on a specific set of items within the test. The items may represent more specific content than the test as a whole, and may be viewed as a sample from a universe of items representing that content.

As a method of illustrating how increases in achievement are related to performance on various constructs or areas of skill within a test, a domain characteristic curve (DCC) is more stable than item-based methods. For example, one can easily imagine a particular addition item being harder than a particular multiplication item for a given examinee, but a representative domain of addition items should be easier than a representative domain of multiplication items for any examinee. This difficulty-order relationship can be illustrated with a plot showing the characteristic curves of both domains, one for addition items and one for multiplication items, as a function of achievement. One would expect this order relationship to be the same across all levels of achievement and to be independent of irrelevant details such as the type of item format one uses.

In line with the above example, a key objective of this study is that the characteristic curves of multiple domains within a given instructional or content strand will be noncrossing. The advantages of noncrossing domain characteristic curves are similar to those of noncrossing item characteristic curves (Andrich, 1985; Masters, et al., 1994;

Wilson, 1989; Wright and Stone, 1979). A set of noncrossing characteristic curves supports inferences of Guttman patterns of mastery. The relative difficulty of the skills, corresponding to the curves, is presumably the same for every examinee, regardless of their scale score. Also, a uniform developmental path to mastery of the domains is implied—the domains are presumably mastered in the same order by all examinees. This uniformity stimulates insights about the meaning of the achievement variable and facilitates descriptions of different points and levels of achievement.

Instructional Timing

Classification systems based on a combination of instructional sequence and content are good candidates for yielding domain scores that support decisions about curriculum and instruction. In Schulz, et al. (1992), the reading curriculum for Grades 3 to 8 was defined by a total of ninety-six objectives, sixteen objectives per grade. A domain of ten items represented each objective within grade. Objectives were organized into strands, such as "uses a dictionary", that spanned grade levels. Domains within the same instructional strand increased in difficulty across grade levels. This regularity provided a useful perspective on whether the learning objectives were appropriately matched to grade levels and whether a 70% correct standard for 'mastery' was appropriate for each objective.

Similarly, the domains in Janssen, et al. (2000), represented reading attainment targets in primary education. These investigators ventured that "the ordering of attainment targets on the basis of their difficulty may provide relevant information regarding the order in which these attainment targets should be taught in the curriculum."

The NAEP content strands, topics, and subtopics (Chapter 3, NAGB, 2000) appear to have an instructional sequence. For example, 20 of the 28 Number Sense subtopics

(71%) covered at Grade 8 can also be represented at Grade 4, while only 13 of the 25 Algebra subtopics (52%) covered at Grade 8 can be represented at Grade 4. This suggests that Number Sense subtopics are introduced earlier in an instructional sequence than are Algebra subtopics. Topics that are listed first within a strand are generally represented for the first time at a lower grade level (Grade 4 or 8) than are topics listed last (Grade 8 or 12). For example, Topic 1 in the Algebra strand includes six subtopics represented at Grade 8, five of which (83%) are also represented at Grade 4. Of topics 8 through 14 of Algebra (the last seven topics listed in the framework), only four can be represented at Grade 8; none of these four are represented at Grade 4. This pattern suggests that instructionally-ordered subdomains can be created within each content strand.

Mere indications of instructional order among strands, topics and subtopics in the NAEP math assessment are not sufficient for defining instructionally ordered domains, however. The instructional timing of individual items must be considered. Some topics and subtopics may be assessed at all three grades of the NAEP math assessment (Grades 4, 8, and 11). This means items within at least some subtopics and topics could be dispersed across the grade level continuum in terms of when the skills they measure are introduced or mastered. A significant facet of the present study, therefore, is to estimate the instructional timing of each test item and to incorporate this information, along with other item-level information such as difficulty, into the definition and description of multiple domains within a test.

Methods

Assessment Items

The Grade 8 NAEP Mathematics assessment for 2000 involved a total of 159 items whose IRT parameters were estimated. These were organized into thirteen test administration blocks numbered 3 through 15. (The first two blocks in each test booklet are used for collecting administrative and demographic information.)

Table 1 shows the existing NAEP item classifications used in this study. Content strand classifications were based on a spreadsheet supplied by the NAEP test development contractor. The spreadsheet also contained topic and subtopic classifications (within strand). These were used informally when domains were defined in this study. Item type classifications were based on the scoring rubrics and the item statistic file obtained from the contractor.

Table 1. NAEP Item Classifications

Classification Type and Category	Number of Items
Content Strand	
Number Sense, Properties, and Operations	44
Measurement	23
Geometry and Spatial Sense	29
Data Analysis, Statistics, and Probability	23
Algebra and Functions	40
	159
Item Type	
Multiple Choice	98
Short Answer	34
Extended Response	27
	159

Domain Characteristic Curves

Appendix A contains a detailed, technical presentation of how domain characteristic curves were constructed. In brief, an expected score was computed for each item as a function of the NAEP score scale. The domain score is the sum of expected item scores over the items in the domain, divided by the total possible score on those items. For comparability purposes, domain scores in this study are converted to expected proportion correct (EPC) scores. The domain characteristic curve shows the EPC, which is also called the domain score, as a function of the NAEP composite scale score.

Curriculum Panel

The curriculum panel consisted of three consultants drawn from a network of consultants involved with the NAEP Mathematics assessment. The size and composition of the panel reflected the exploratory nature of this study. By using panelists who had experience with NAEP, we minimized the amount of training and time it would take to ground panelists to the items and test framework, procedures in handling secure test items, etc.

Ratings of Instructional Timing

Round 1: For each item, the panelists were asked to identify the most difficult or complex, curriculum-based skill needed to answer the item correctly and to rate the instructional timing of the skill in terms of Introduced (I), Reinforced (R), and Mastered (M). Instructional timing was to be based on a traditional mathematics curriculum, as experienced by an average student. All ratings were made on a seven-category rating scale ranging from "below grade 5" to "above grade 9". An initial set of directions and definitions of these concepts was provided to the panelists (see Appendix B).

After the panelists rated two blocks of items (Blocks 13 and 14, a total of 20 items), a conference call was held to answer questions and develop a consensus about key facets of the rating procedures. It was agreed that Introduced ratings would be based on the introduction of the skill at a conceptual level, not necessarily coverage of the specific computation called for by the item; Mastery ratings were to reflect the grade at which eighty percent of students would have mastery of the skill. Mastery was not specified in terms of an expected score or probability of correct response, but it was assumed that the skill would no longer be specifically taught or reinforced in the curriculum.

Round 2 Round 1 ratings were analyzed for consistency among raters and other characteristics, such as floor and ceiling effects thought to be important to the study. Floor effects were noted for ratings of Introduced—a relatively large proportion of items were rated as introduced "below Grade 5." Because all raters indicated that reinforcement of a skill occurred in every grade between Introduced and Mastery, ratings of "Reinforced" were redundant, and were dropped from further consideration in the study. Due to limitations of time and resources, we decided to live with the floor effects of the rating scale rather than ask the curriculum panelists to revisit their ratings of "Grade 5 or lower."

To prepare for a second round of ratings, Introduced and Mastery ratings were flagged if they were two or more categories (grades) away from their respective median. [With three raters, the median rating was the middle rating (if all raters differed), or the mode (if at least two raters agreed.)] In addition, the difference between each rater's Mastery and Introduced ratings was computed. This difference was flagged (for a given rater) if it differed from the median difference by two or more grades or rating scale categories. Across raters, a total of 21 Introduced ratings, 30 Mastery ratings, and 36 differences (Mastery minus Introduced) were flagged. These numbers were 6.5% to 9% of the ratings.

A spreadsheet containing all ratings was returned to the panelists for review. Flagged ratings and differences were highlighted. Each panelist reviewed ratings that were flagged. Raters were not required to revise their ratings, but were free to do so. Raters indicated their changes directly on the spreadsheet and returned the spreadsheet via email. A majority of the ratings were changed. After incorporating raters' changes, a total of 11 Introduced ratings, 2 Mastery ratings, and 2 differences (Mastery minus Introduced ratings) were flagged by the same criteria listed above. No further changes were made to the ratings.

Classification of Items into Domains

To provide the curriculum panel with a starting point for this task, a mathematics teacher created an initial set of domains. The teacher had experience teaching pre-algebra and algebra courses to grade school and high school students. The teacher was introduced to the methods and goals of the study. She was given a list of the items ordered by mean Introduced, then mean Mastery rating, within each content strand (Number Sense, Measurement, etc., (see Table 1)). She was also given a 3-ring binder with the text of the items in the same order as they were listed in the spreadsheet. The NAEP topic and subtopic of the item were also shown. It took the teacher approximately 15 hours to classify the items into three or four domains within each content strand. The teacher provided brief, written, general descriptions of the skills each domain represented.

The Teacher domains were then reviewed, and revised in some cases by a "Reviewer." The Reviewer was a mathematics test development specialist at ACT. The Reviewer provided a different set of Algebra domains, including brief domain descriptions and indicated some general dissatisfaction with the Teacher's Number Sense domains.

A set of "Baseline" domains was then developed or chosen by the principal investigator from the Teacher and Reviewer input. For Algebra and Number Sense the baseline domains were more numerous than those developed by the teacher or reviewer in order to take account of both sources of input. For Geometry, Measurement, and Data Analysis, the baseline domains were the Teacher's or the Reviewer's unaltered in any way.

The baseline domains were given to the panelists as a starting point from which to construct their own domains. Materials included a spreadsheet showing the item classifications into baseline domains. The item block, sequence within block, NAEP content topic and subtopic (if applicable), mean Introduced and mean Mastered ratings were shown for each item. Items were sorted by mean Introduced rating, then by mean Mastered rating within baseline domain within strand. Baseline domains were ordered by relative difficulty. The panelists also received the text of the items in the order items were listed in the spreadsheet.

Identical sets of materials, including spreadsheets and item text in corresponding order, were provided for each set (strand) of Teacher and Reviewer domains that were not adopted as baseline domains. The panelists could thus see all ways that had been conceived for defining domains within content strands.

The spreadsheet given to the panelists showing the baseline domains also contained columns for them to indicate their agreement with the baseline domains or, alternatively, to indicate a different set of within-strand domains. They were encouraged to develop their own within-strand domains. The panelists were also asked to provide domain descriptions for any new domains they created, and to review the descriptions provided for any domains that they chose to adopt from the baseline domains.

All experts (Teacher, Reviewer, and panelists) were encouraged to leave any item unclassified if they felt it did not fit into any of the three or four domains they defined within a content strand. [We do not feel that a multiple domain method of describing achievement requires one to classify every item in the assessment into a domain.] Fewer than three items within any strand were left unclassified by any one expert.

After receiving the domain definitions of the panelists, two levels of 'final' domain classifications, F1 and F2, were developed by the principal investigator. F1 domains tended to be smaller and more detailed. They represented as much consensus as possible among the five sources of expert input (Teacher, Reviewer, and three panelists). At least two, and often all five sources would agree with an item's F1 domain classification, based on their input.

F2 domains consisted of the larger F1 domains, plus domains that were created by merging two or more smaller F1 domains within the same strand together. Merging was necessary to achieve the sample size and/or stability of DCCs across item types (see Figures 8 through 12). Descriptions of the F1 and F2 domains were developed by the principal investigator using relevant portions of the domain descriptions provided by the experts.

This study is currently incomplete with regard to defining and describing the F1 and F2 domains and documenting consensus among the experts. The teacher is scheduled to review and revise if necessary the current F1 item classifications and domain descriptions. (This will take place the week of March 18 to 22nd.) The Reviewer will check the teacher's work. F1 and F2 domains will then be sent to the curriculum panelists for final review and input.

We do not anticipate the teacher or panelists will make any substantial changes to the current final (F2) domains. It is possible that some additional items will be left unclassified in order to accommodate panelists' input. We anticipate that the domain descriptions will be improved and will become slightly longer.

Results

Ratings of Instructional Timing

Inter-rater Agreement. One index of agreement among raters is the range, or the difference between the maximum and minimum rating of a given item. If all raters give the item the same rating, this difference is zero. Table 2 shows the cumulative relative frequency distribution of this difference, based on the final ratings. For "Introduced" ratings, the difference was one or less for 84% of the items. For "Mastery" ratings, the difference was one or less for 78% of the items. There were no differences of four and only one difference of three for both "Introduced" and "Mastery" ratings.

Table 2. Rater Agreement

Range (Max minus Min)	Cumulative Relative Frequency by Type of Rating	
	Introduced	Mastery
0	.35	.29
1	.84	.78
2	.99	.99
3	1.00	1.00

Another index of agreement is the correlation between raters. For Introduced ratings, pairwise correlations (Pearson) among the three raters were .76, .80, and .82. For Mastery ratings the pairwise correlations were .84, .85, and .89.

Table 3. Distribution of Item Mean Rating

	Frequency by Type of Rating (%)	
Mean Rating	Introduced	Mastery
Exactly 4.0	38 (24%)	0
“> 4.0” to “< 5”	33 (21%)	14 (9%)
“≥5” to “<6”	35 (22%)	22 (14%)
“≥6” to “<7”	36 (23%)	29 (18%)
“≥7” to “<8”	13 (8%)	26 (16%)
“≥8” to “<9”	4 (3%)	41 (26%)
“≥9”	0	27 (17%)
Total:	159 (100%)	159 (100%)
Avg. Mean Rating:	5.3	7.2
Std. Dev.	1.1	1.6

Mean Ratings. To compute the mean rating for an item, ratings of “below 5” were coded ‘4’, and ratings of “above 9” were coded ‘10’. Table 3 summarizes the distributions of the mean rating across items. For example, the row labeled (“>4” to “<5) would include an item whose three Introduced ratings from the three panelists were 4, 4, and 5 (mean = 4.33) and an item whose Mastery ratings from the panelists were 4, 5, and 5 (mean = 4.67). Frequency counts are shown separately for Introduced and Mastery ratings. Mean Introduced ratings were distributed across items with mean 5.3 and standard deviation 1.1. Mean Mastery ratings were distributed across items with mean 7.2 and standard deviation 1.6.

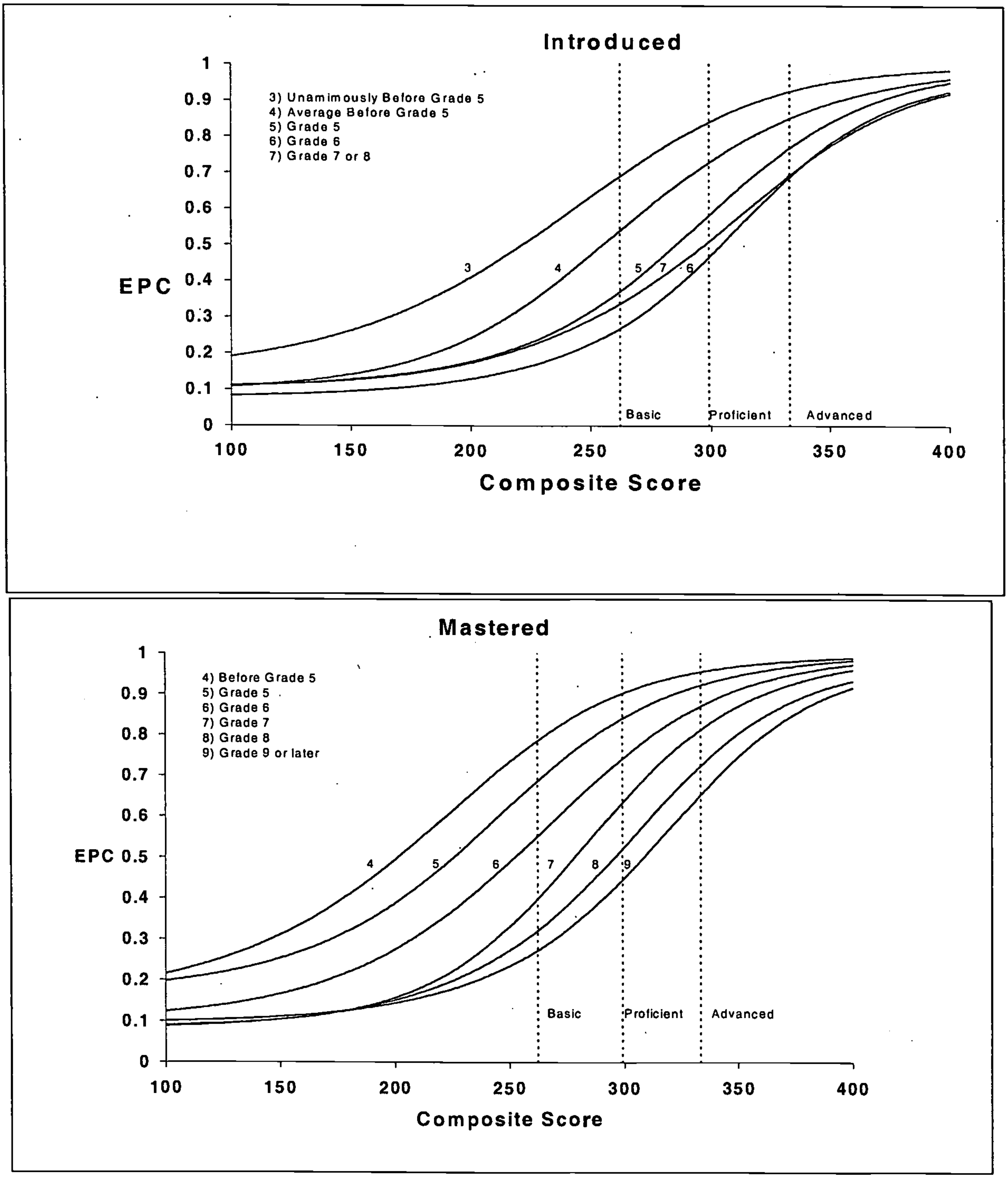
Figure 1 shows that if domains were based solely on ratings of instructional timing, domain characteristic curves would, with one exception, have the properties sought in this study. The curves in Figure 1 tend to be well-separated and non-crossing. The grade level at which items are expected to be mastered completely accounts for the difficulty order of corresponding characteristic curves in the lower panel of Figure 1. These curves do not overlap in any region of the scale. In the upper panel, the grade level at which items are introduced accounts for the difficulty order of all but two curves. Below the Advanced level of proficiency, items introduced at Grade 7 or 8 (curve "7") tend to be slightly easier than items introduced at Grade 6 (curve "6"). At the Advanced level, the two categories of items are equally difficult.

The difficulty reversal of the Grade 6 versus Grades 7&8 curves for "Introduced" ratings is partly explained by the effects of item type. Thirty-one percent of the items introduced in grade 6 were extended response items. Only eighteen percent of the items introduced at grades 7 or 8 were extended response. When extended response items were removed from both of these item-groups, the corresponding curves were practically indistinguishable.

In Figure 1 and other figures, NAEP achievement level boundaries for Grade 8 are indicated by dashed vertical lines. Lower boundaries for Basic, Proficient, and Advanced levels of achievement were established in a separate standard setting process (American College Testing, 1993). The lower boundaries on the NAEP composite scale are 262 for Basic, 299 for Proficient, and 333 for advanced. The estimated national mean and standard deviation for Grade 8 in the Year 2000 Assessment were, respectively, 275 and 37. The estimated percentage of Grade 8 students at or above each achievement level was

66 for Basic, 27 for Proficient and 5 for Advanced (National Center for Education Statistics, 2001). The figures in this paper show a 100 to 400 score range, but the full range of the NAEP composite scale for Grade 8 is 0 to 500.

Figure 1. Instructional Timing Characteristic Curves



The effect of item type, illustrated in Figure 2, indicates that one must be careful in attributing differences between domain characteristic curves to content or instructional timing. Below 225 on the NAEP scale, the characteristic curve of multiple choice items is clearly distinguished from those of short answer and extended response items. Only the multiple-choice items have a non-zero c-parameter (a lower asymptote suggestive of guessing). Extended response items are distinguished from multiple choice and short answer items by being more difficult everywhere on the NAEP scale.

Figure 2. Item Type Characteristic Curves

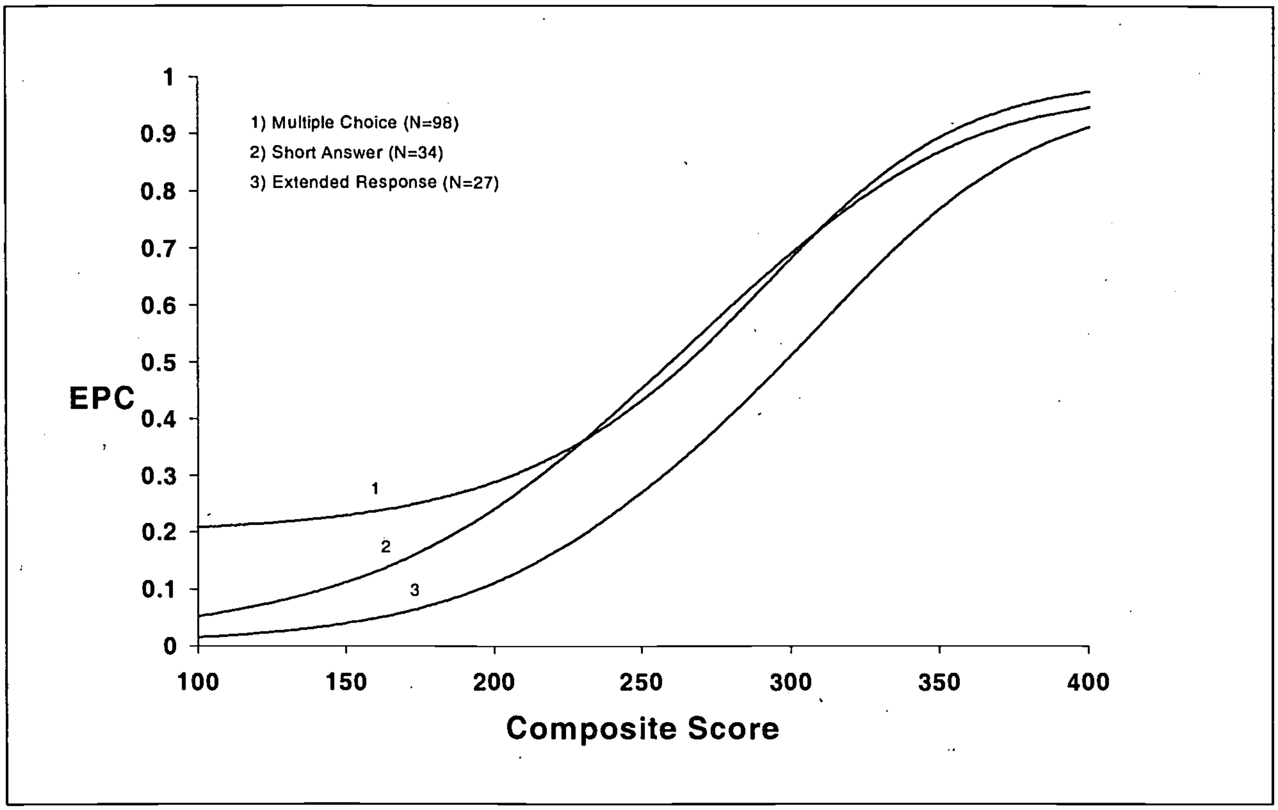


Table 4 shows the instructional timing of the item types. Skills measured by short answer items tend to be introduced earliest (4.8) followed by multiple choice items (5.4) and then extended response items (5.7). Item types are even more differentiated by Mastered ratings. Mean Mastered ratings by item type are 6.3, 7.3, and 8.0 for skills represented by, respectively, short answer, multiple choice, and extended response items. [The floor effects associated with "Introduced"ratings might have limited the extent to which these ratings are associated with item type.] Differences among item types in the mean item b-parameter are predictable from the instructional timing ratings, with short answer items being easiest (-.30) followed by multiple choice items (.09) and extended response items (.50). Mean b-values may not be directly comparable across item types due to differences in other parameters of the items.

Table 4. Instructional Timing by Item Type

		Means		
Type of Item	N	Introduced	Mastered	b-value
Multiple Choice	98	5.4	7.3	.09
Short Answer	34	4.8	6.3	-.30
Extended Response	27	5.7	8.0	.50
	159			

Within-Strand Domains

A total of fourteen domains were defined, three within each content strand except for the Measurement strand, which contained two domains. The domains will be referred to by both number and content strand. For example, Domain 1 in the Number Sense strand is not the same domain as Domain 1 in the Algebra strand. Domain descriptions are contained in Tables 5 through 9. These are preliminary descriptions, based upon a synthesis of expert domain descriptions.

Table 5. Number Sense Domain Descriptions

Domain	Description
NS-1	Whole number computation, place-value, rounding, and 2-dimensional representation of fractions. In some cases, the whole number computation occurs in a problem situation, but the numbers needed for the computation are directly given in the text. In others, the problem is completely set up for the student. Computation includes addition, subtraction, multiplication and simple division. Rounding problems do not require computation (i.e., Block 7, Item 3). Students must be able to equate fractions with the amount of shaded area in a two-dimensional figure. Place value problems involve knowledge of position for representation of ones, tens and hundreds.
NS-2	Fractions, ratios and decimals. These items involve computation using fractions and decimals and understanding of ratios. The computation occurs in a problem situation or may be set up for the student.
NS-3	Rates and percents, elementary number theory, exponents, and scientific notation. These items involve using percents or proportions to solve problems. Scientific notation items are included here, too. The numbers are usually not just given for the student to perform. They usually occur within the context of a story problem. Elementary notions of number theory such as even and odd numbers or prime numbers are included in this domain.

Table 6. Measurement Domain Descriptions

Domain	Description
M-1	One dimension, angles and simple area. Items in this domain involve linear, one-dimensional measures, including length, weight, reading scales, and also the concept of perimeter. Also included are items that require knowledge of basic terms and properties of angles, including concepts of a right angle and knowing what a protractor is for. Simple area problems are restricted to rectangles, with the numbers given in a diagram, or problems that require spatial visualization, but not computation, of area.
M-2	Problems involving complex area, surface area, and volume. The area problems in this domain involve more complex geometric figures or arrangements than simple rectangles and more computation. Area may have to be found by computing several areas and adding them together, or by applying spatial-visualization skills and computation. Computation of surface area and volume may be required

Table 7. Geometry Domain Descriptions

Domain	Description
G-1	Basic properties of figures. Items in this domain involve simple properties, including angles of less than 90 degrees, classification of triangles and quadrilaterals, identification of figures. An item involving the measure of a missing angle in a triangle when the other two are known is included here.
G-2	Transformations and Spatial-visual items. This domain consists of items that involve single transformations, including reflections, translations and rotations. Items ask for completion of figures to have a certain symmetry, finding reflections, or drawing lines of symmetry. Spatial-visual items involve abstract visualization, including the rearranging of or mentally transforming of shapes into different shapes, subdividing figures, and visualizing three-dimensional shapes -- how they look flattened out, and how flattened out shapes fold into three-dimensional shapes.
G-3	Properties of figures, high level. These items involve the use of properties of figures to do more complicated problems. Included are problems using the Pythagorean theorem, finding missing information in more complex settings, and using similarity of figures. These items may involve creating a figure that satisfies a set of properties

Table 8. Data Analysis Domain Descriptions

Domain	Description
DA-1	Representing data. Items in this domain test a student's ability to represent information in a simple histogram or frequency chart. This domain may also contain items requiring very basic computational skill and/or understanding of two-dimensional representations of fractions.
DA-2	Using graphs and charts. Items in this domain require the student to use information in graphs and charts (e.g. pie charts, frequency histograms, bivariate line charts). The student may have to decide which information in the chart is relevant to the question and interpret the chart to obtain the numbers needed to answer the question. The student may have to perform arithmetic operations on the numbers to get the right answer. Students must understand how to apply or obtain percentages, proportions, and rates of change.
DA-3	Probability, statistical computation, and meaning of statistical terms. Items in this domain test a student's ability to represent information in a simple histogram or frequency chart. require the student to use information in graphs and charts (e.g. pie charts, frequency histograms, bivariate line charts). The student may have to decide which information in the chart is relevant to the question and interpret the chart to obtain the numbers needed to answer the question. The student may have to perform arithmetic operations on the numbers to get the right answer. Students must understand how to apply or obtain percentages, proportions, and rates of change. Items testing meaning of statistical terms do not require computation, but require understanding of mean, median, mode, and range—including whether one of these statistics is more appropriate than another for a given purpose.

Table 9. Algebra Domain Descriptions

Domain	Description
A-1	Basic operations, logic, simple patterns, and grids. These items require the student to apply basic computational skills (addition, subtraction, multiplication and division) to find solutions to simple equations. Students must recognize basic symbols such as "<" and know the correct order in which to perform multiple operations. Solutions and problems involve only whole numbers. Simple pattern items require the student to recognize numeric or logical patterns. The problems involve patterns of frequency, shading, or simple relationships between numbers. Logic items require the student to establish the order of three or more objects when given two or more pieces of partial information. Grid items require the student to understand how information such as rate of change or direction from a starting point is represented on a grid. The solutions involve counting grid units.
A-2	Coordinate systems. These items use graphical displays but are more abstract than the grid problems in Domain 1, and involve computation. Solutions require understanding of coordinate systems, using coordinates of points to find coordinates of other points or solutions, computing rates from a graph, and extending one or more patterns of change represented graphically.
A-3	Understanding variables in expressions, solving equations, and complex problem solving. These items require the student to understand how quantity is represented by a variable, the use of variables to form expressions, and the evaluation of expressions to check on a quantity. Students may be required to solve equations with one or more variables and to use symbols to represent solutions. Students must be proficient in computation to recognize and correctly extend a pattern in some of these problems.

Table 10 shows summary statistics for the domains. Only four domains contained fewer than ten items. The largest domain—Domain 1 in Number Sense—contained eighteen items. In all but one case, the order of domains by their mean item b-value matched their order by mean Introduced and mean Mastery ratings. For example, Domains 1 to 3 in the Number Sense strand had mean item b-values of, respectively, -1.23, 0.0, and 0.84; mean Introduced ratings of, respectively, 4.1, 4.9, and 6.1; and mean Mastery ratings of, respectively, 4.9, 6.6, and 8.5. One exception to this result is that Domain 3 in Data Analysis has a lower mean Introduced rating (5.8) than Domain 2 in Data Analysis (5.9), but higher mean Mastery (8.2 versus 8.0) and higher mean item b-value (.66 and .43). None of these differences are significant, however.

Table 10. Domain Summary Statistics

			Means		
Content Strand	Domain	Number of Items	Introduced	Mastered	b-value
Number Sense	NS-1	18	4.1	4.9	-1.23
	NS-2	16	4.9	6.6	0.00
	NS-3	13	6.1	8.5	-.84
Measurement	M-1	11	4.1	5.6	-.72
	M-2	10	5.7	8.0	1.10
Geometry	G-1	5	4.3	5.9	-.38
	G-2	16	4.8	6.6	-.04
	G-3	8	7.0	9.4	1.02
Data Analysis	DA-1	5	4.1	5.3	-1.39
	DA-2	8	5.9	8.0	.43
	DA-3	10	5.8	8.2	.66
Algebra	A-1	15	4.6	6.6	-.83
	A-2	10	6.4	8.4	.54
	A-3	14	6.6	8.9	.67

In a few other cases, there were not large differences between the instructional timing and difficulty of domains within a content strand. Domains 2 and 3 in Algebra and Domains 1 and 2 in Geometry strand were not widely separated in instructional timing or difficulty. These domains are distinguished from each other primarily by content.

Figure 3. Instructional Timing and Difficulty of Items by Number Sense Domain

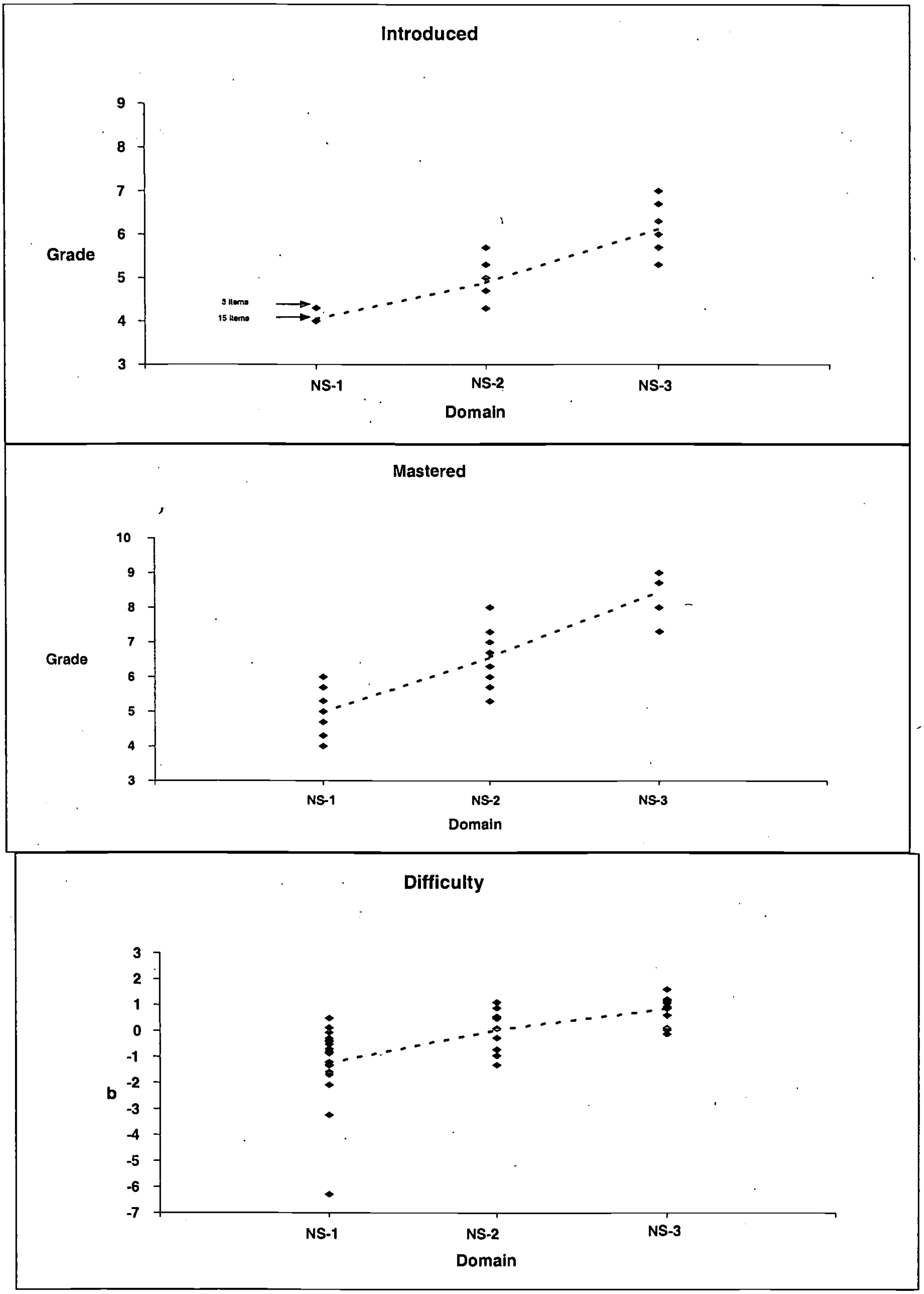


Figure 4. Instructional Timing and Difficulty of Items by Measurement Domain

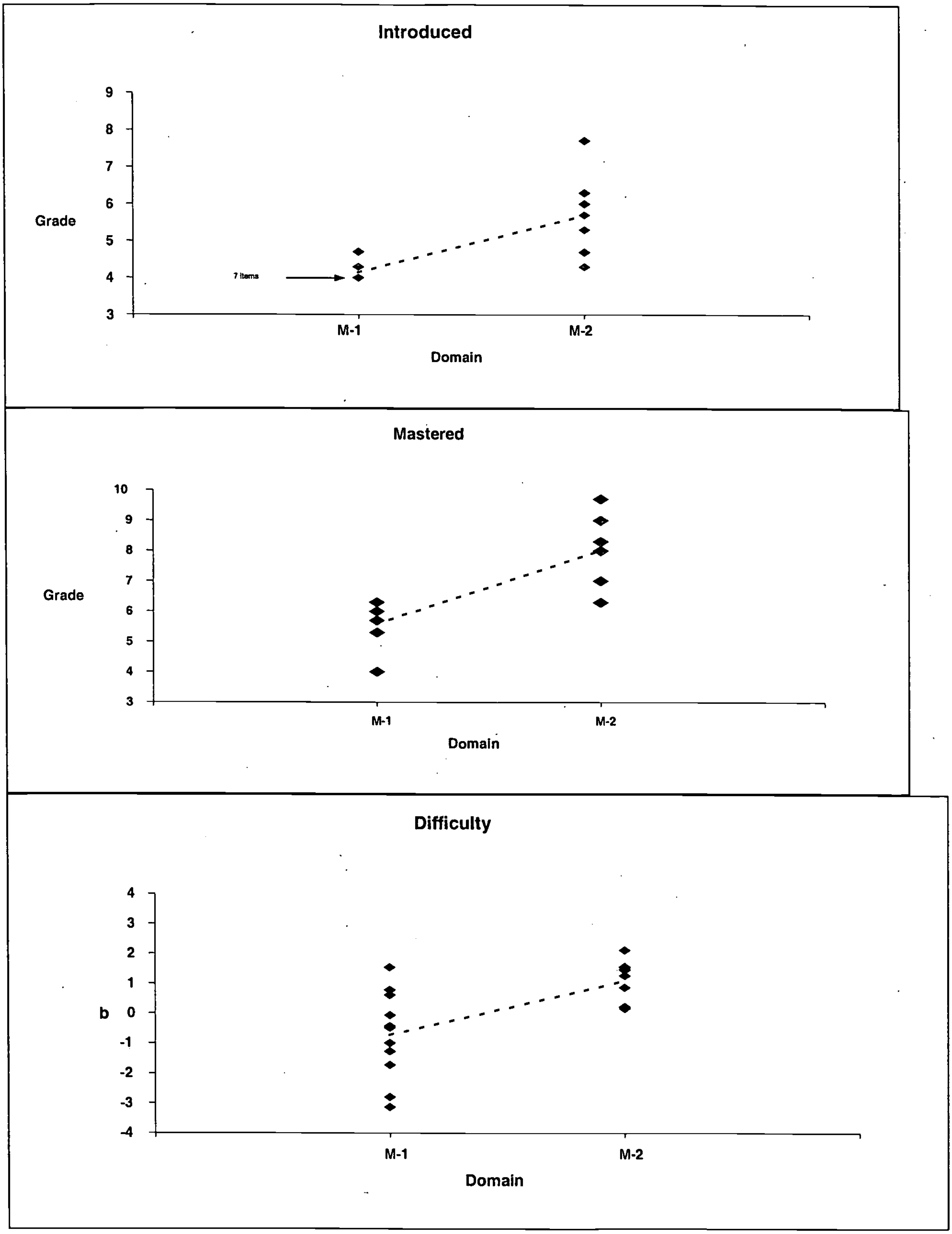


Figure 5. Instructional Timing and Difficulty of Items by Geometry Domain

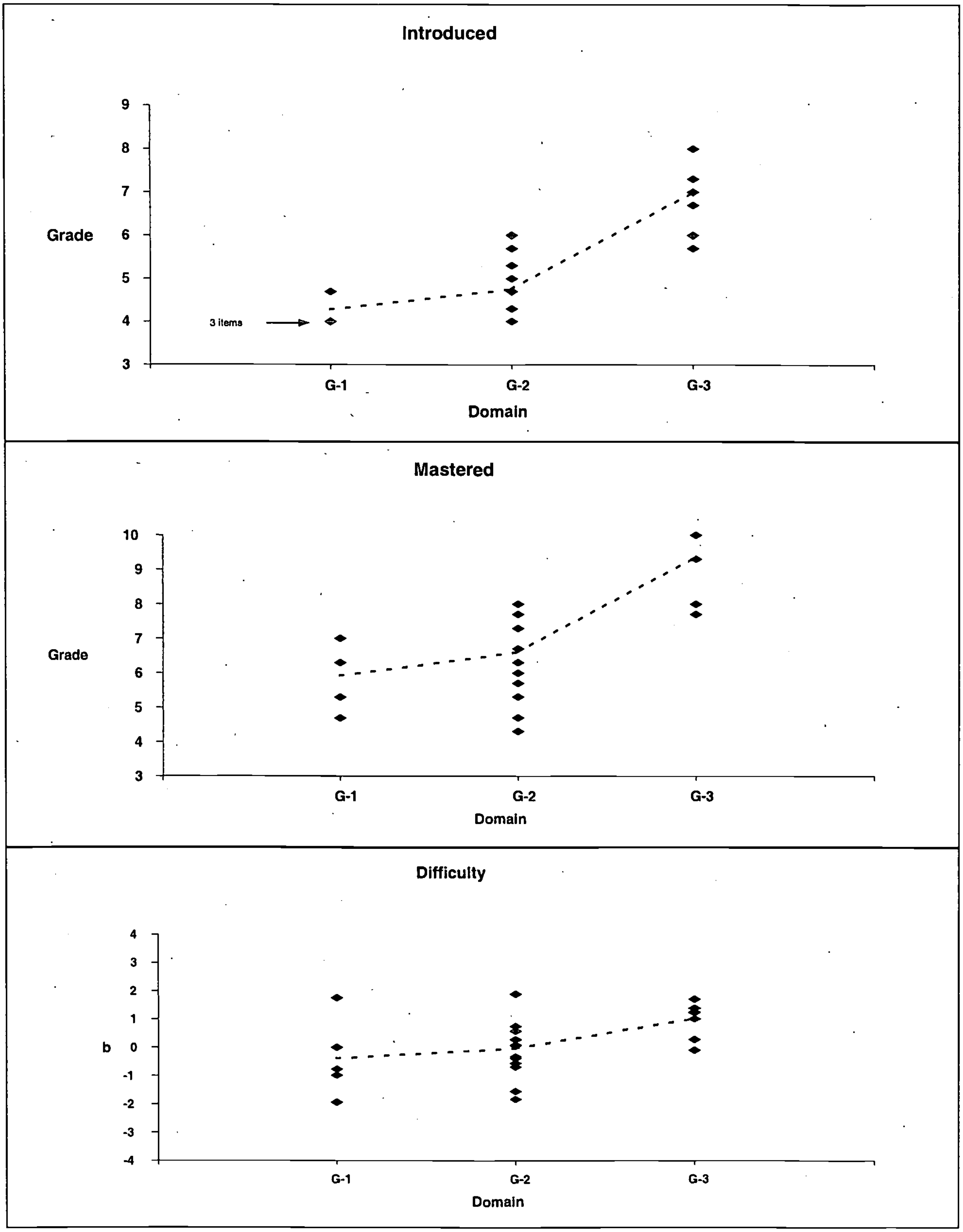


Figure 6. Instructional Timing and Difficulty of Items by Data Analysis Domain

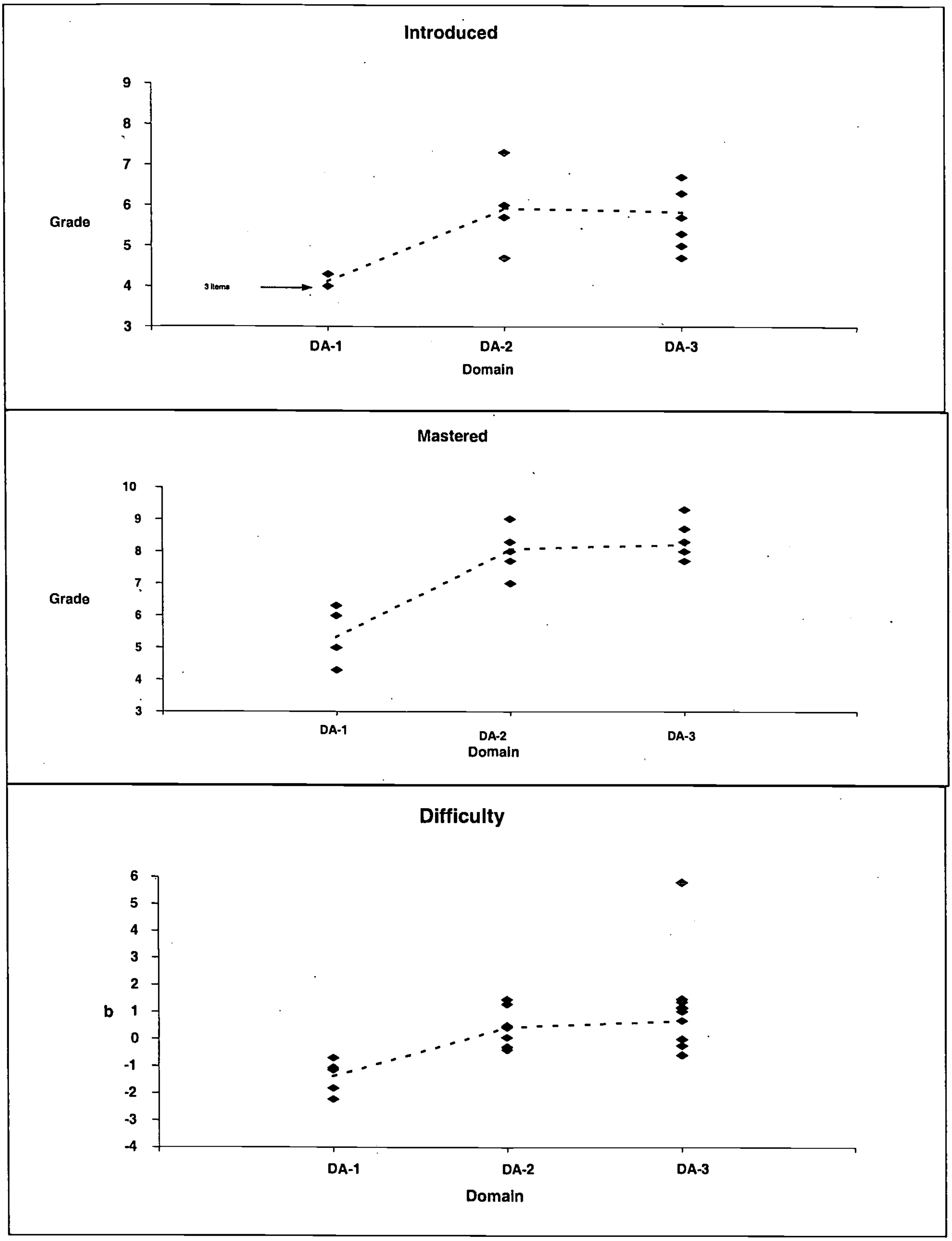
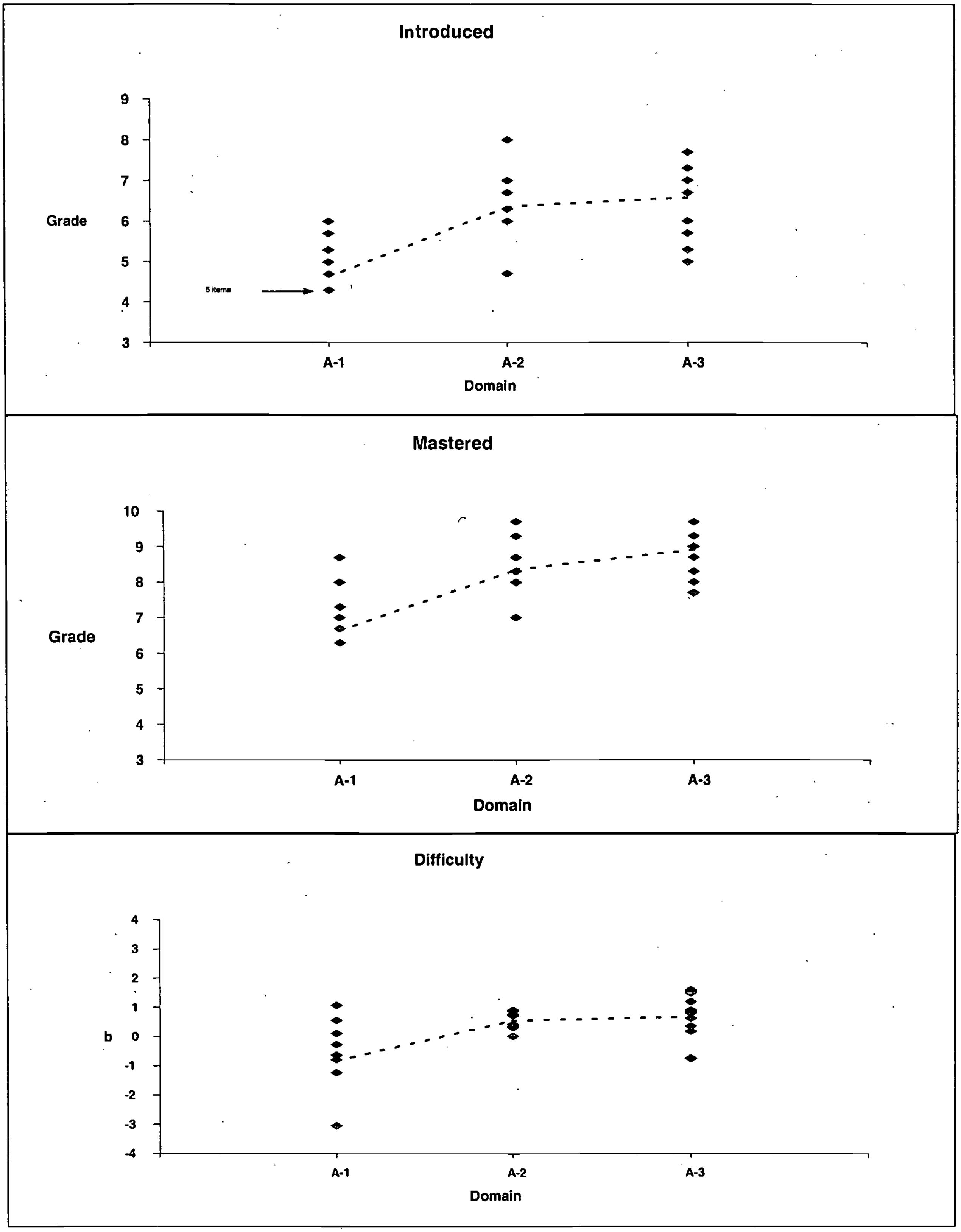


Figure 7. Instructional Timing and Difficulty of Items by Algebra Domain



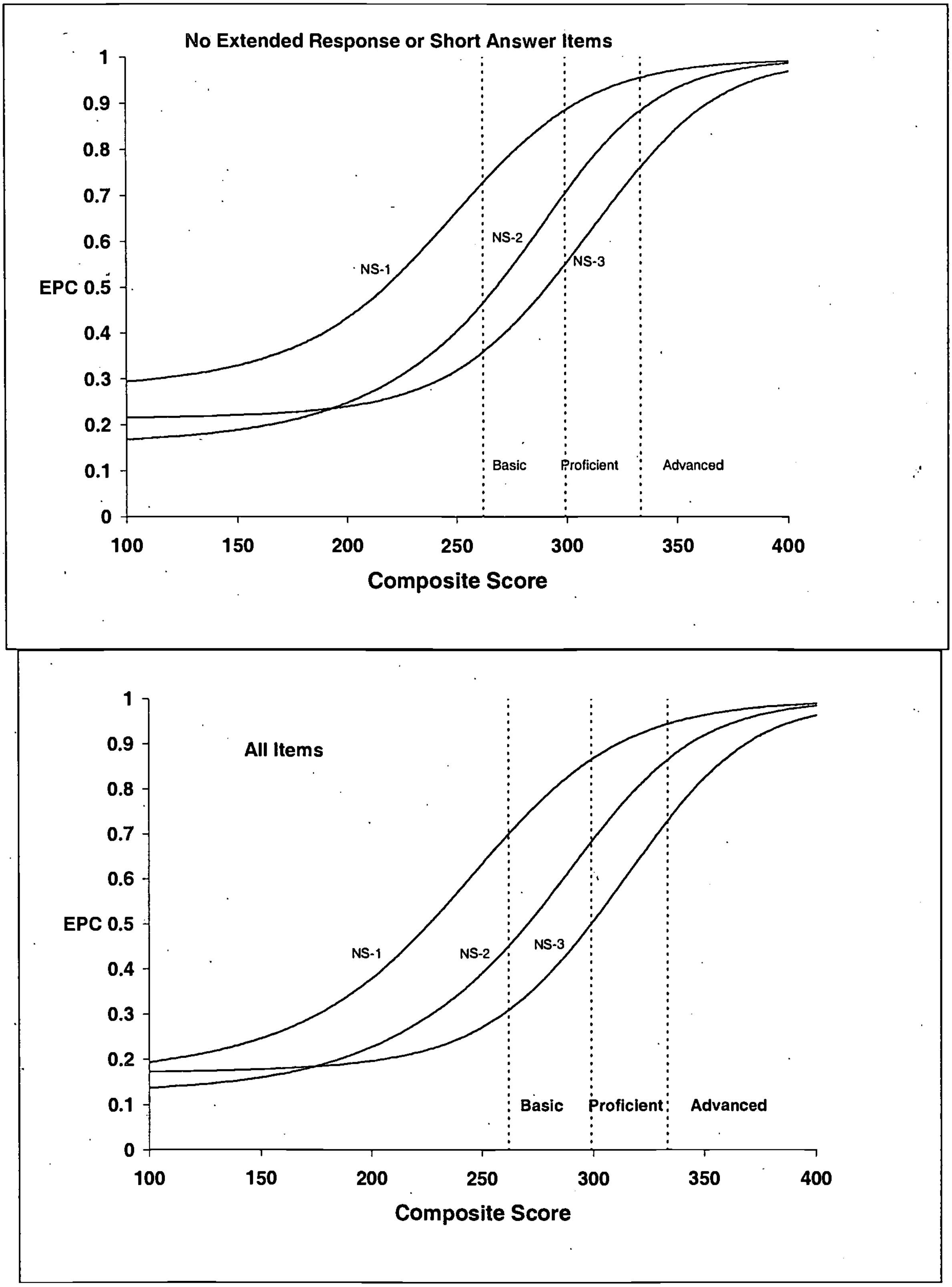
Figures 3 to 7 illustrate the dispersion and overlap in the instructional timing and difficulty of items across domains. Figure 3, for example, plots the instructional timing and difficulty of Number Sense items by their domains. The dashed lines within these figures connect the means shown in Table 10. Items are generally more tightly dispersed within than across domains. Items within the first domain of each strand tend to show little dispersion due to the floor effect of the rating scale mentioned above. A single point corresponding to "Grade 4" in these figures, typically represents more than one item.

Within-Strand Domain Characteristic Curves

Figures 8 through 12 show the characteristic curves of the within-strand domains. There is one figure per content strand. The upper panel in each figure shows the domain characteristic curves (DCCs) based on all items in the domain. The lower panel shows the DCCs with extended response items excluded. Since the Number Sense strand (Figure 8) contained only one extended response item, the lower panel for this strand shows DCCs with short-answer items, as well as the extended response item, excluded—that is, DCCs based only on multiple choice items.

Figures 8 through 12 are consistent with regard to two important results. First, none of the DCCs in these figures overlap except in the tail regions associated with guessing and differences between item types. Second, the difficulty order of the domains is the same in both the upper and lower panels. That is, exclusion of extended response items did not change the difficulty order of the domains. Thus, there may be a general basis for saying that one domain is more difficult than another and that the DCC curves show the order in which the domains are mastered in the curriculum, given any reasonable EPC criterion for mastery.

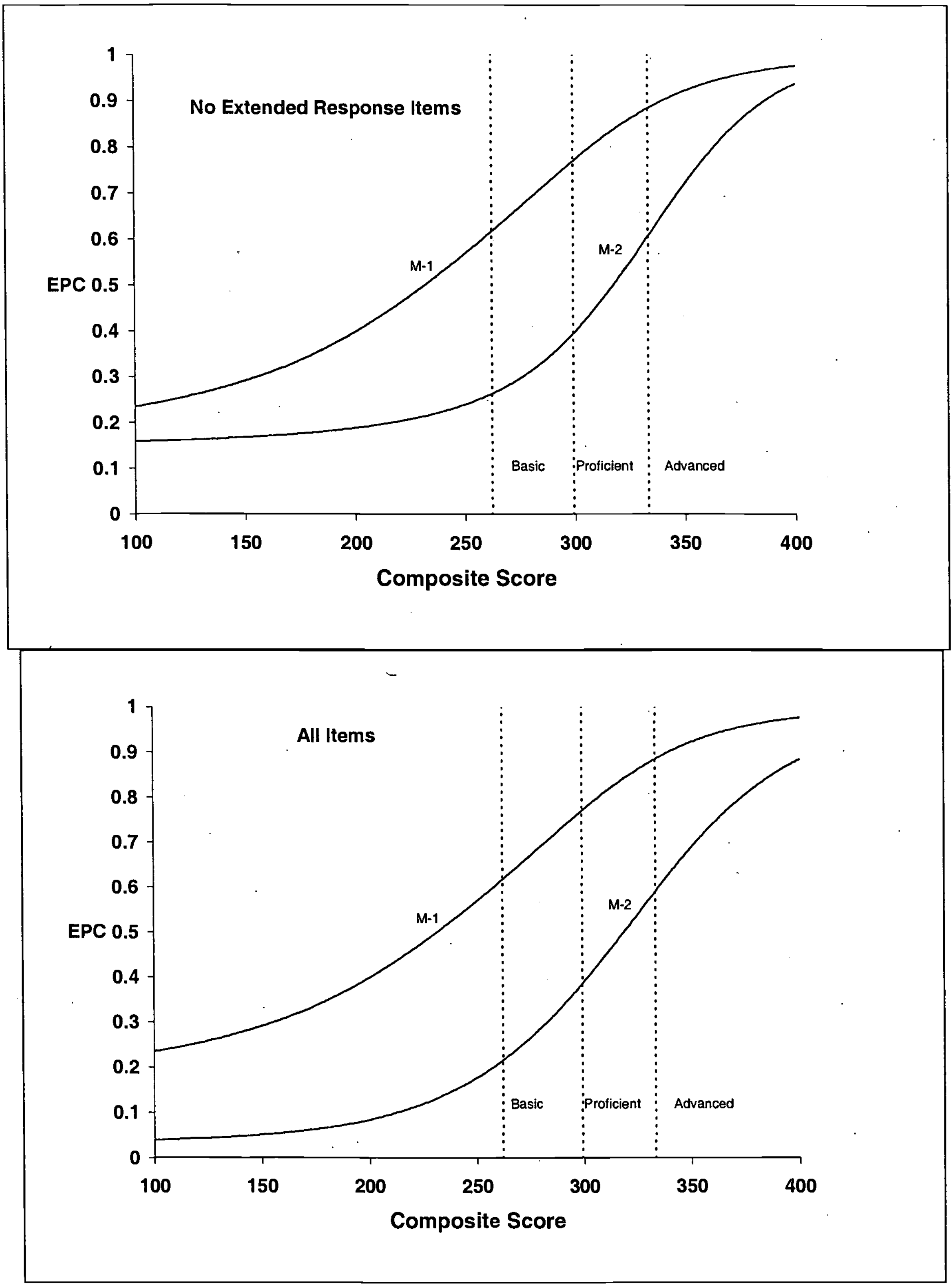
Figure 8. Number Sense Domain Characteristic Curves



No Extended Response or Short Answer Items

All Items

Figure 9. Measurement Domain Characteristic Curves



34

Figure 10. Geometry Domain Characteristic Curves

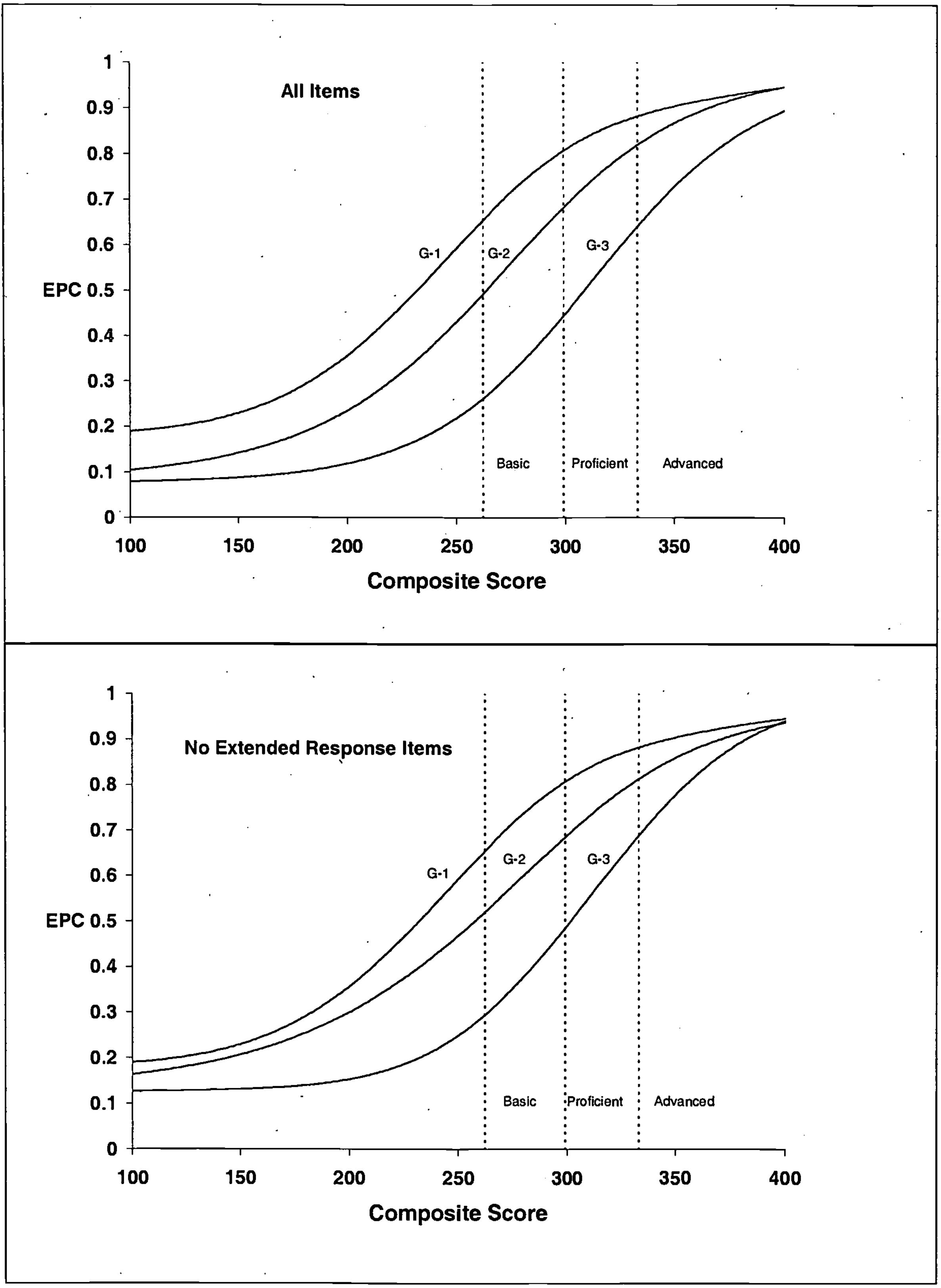


Figure 11. Data Analysis Domain Characteristic Curves

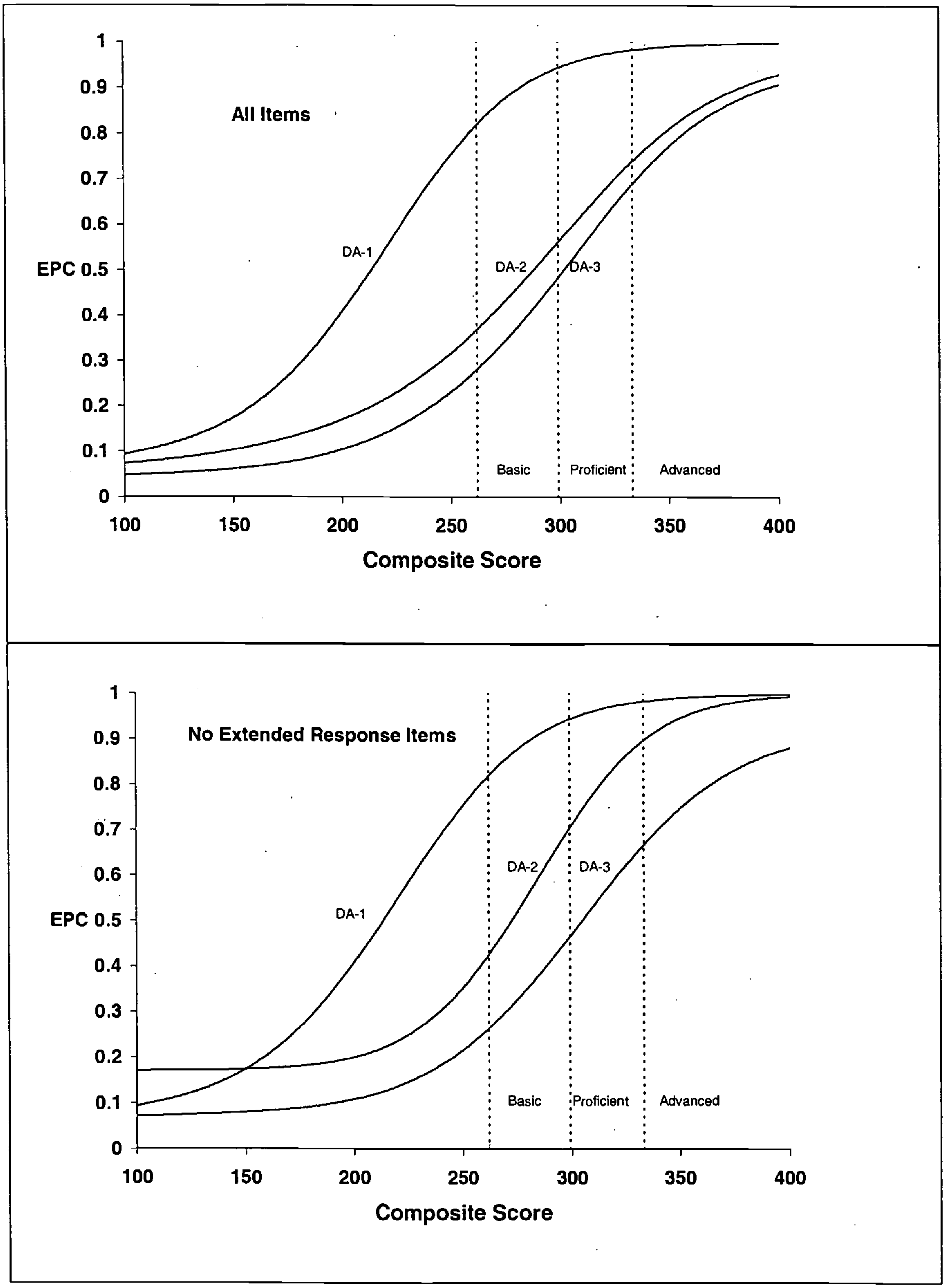


Figure 12. Algebra Domain Characteristic Curves

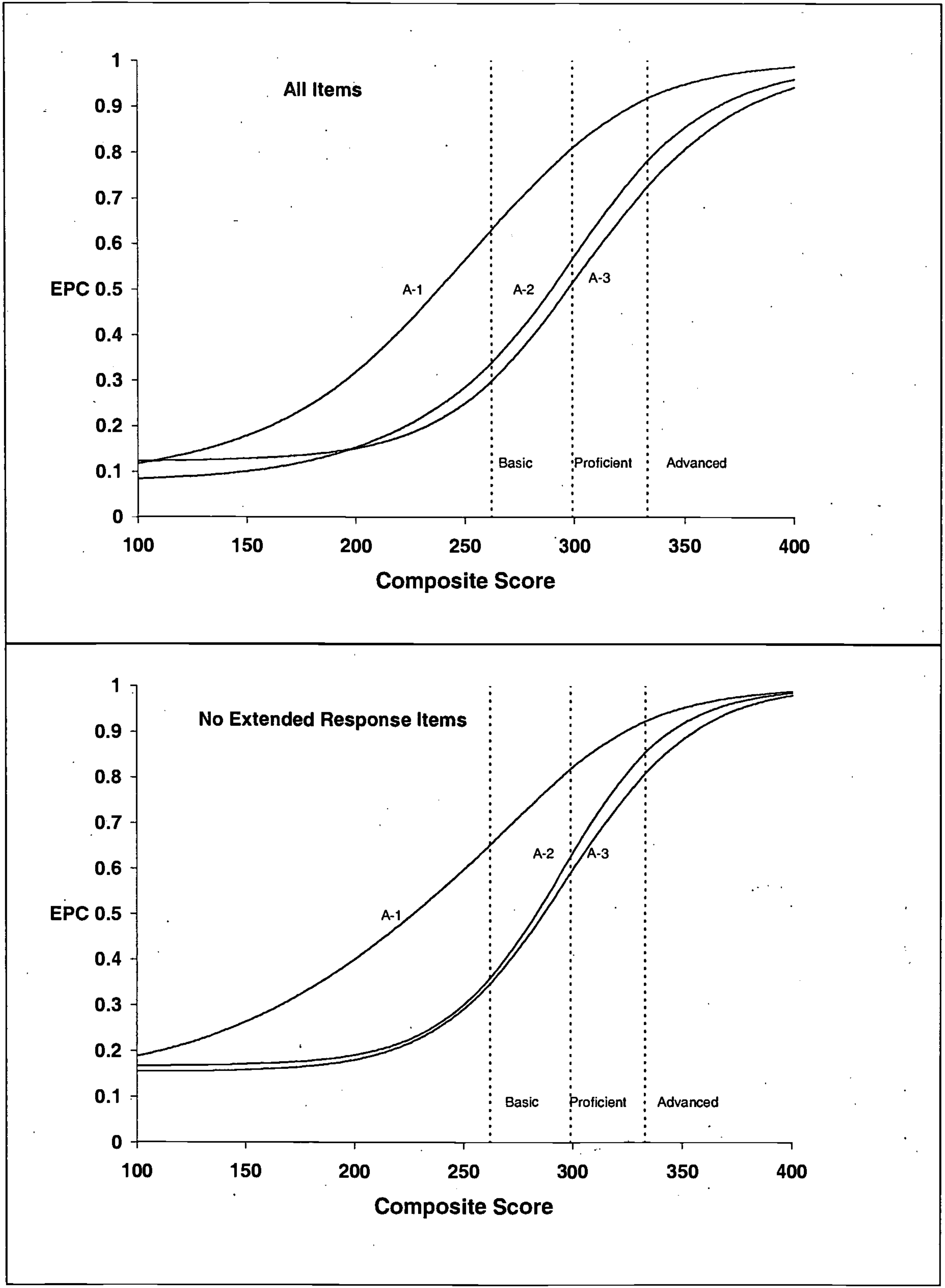


Table 11 shows the expected domain scores associated with the achievement level boundaries. These scores correspond to the top panel of Figures 8 through 12—domain scores with all item types included in the domain. The pattern of shaded versus non-shaded domain scores in Table 11 shows classifications of mastery/nonmastery pertaining to the lower boundaries of the achievement levels if the criterion for mastery were .65 or 65% on the domains. [This is a purely hypothetical criterion.] Students at the lower boundary of the Basic level would have mastery of three domains: Number Sense 1, Geometry 1, and Data Analysis 1. Students at the lower boundary of the Advanced level would have mastery of all but two domains: Measurement 2 and Geometry 3. Students at the lower boundary of the Proficient level would have mastery of half of the domains.

Strictly speaking, a domain score in this study is the mean of a hypothetical population of students defined by a given NAEP composite scale score. Half of the students in this population would have a higher domain score; half would have a lower domain score than the one associated with the given NAEP composite true scale score. This interpretation is due to the composite nature of the NAEP scale and the fact that the domain scores correspond to only one of the scales (strands) that are used to form the composite.

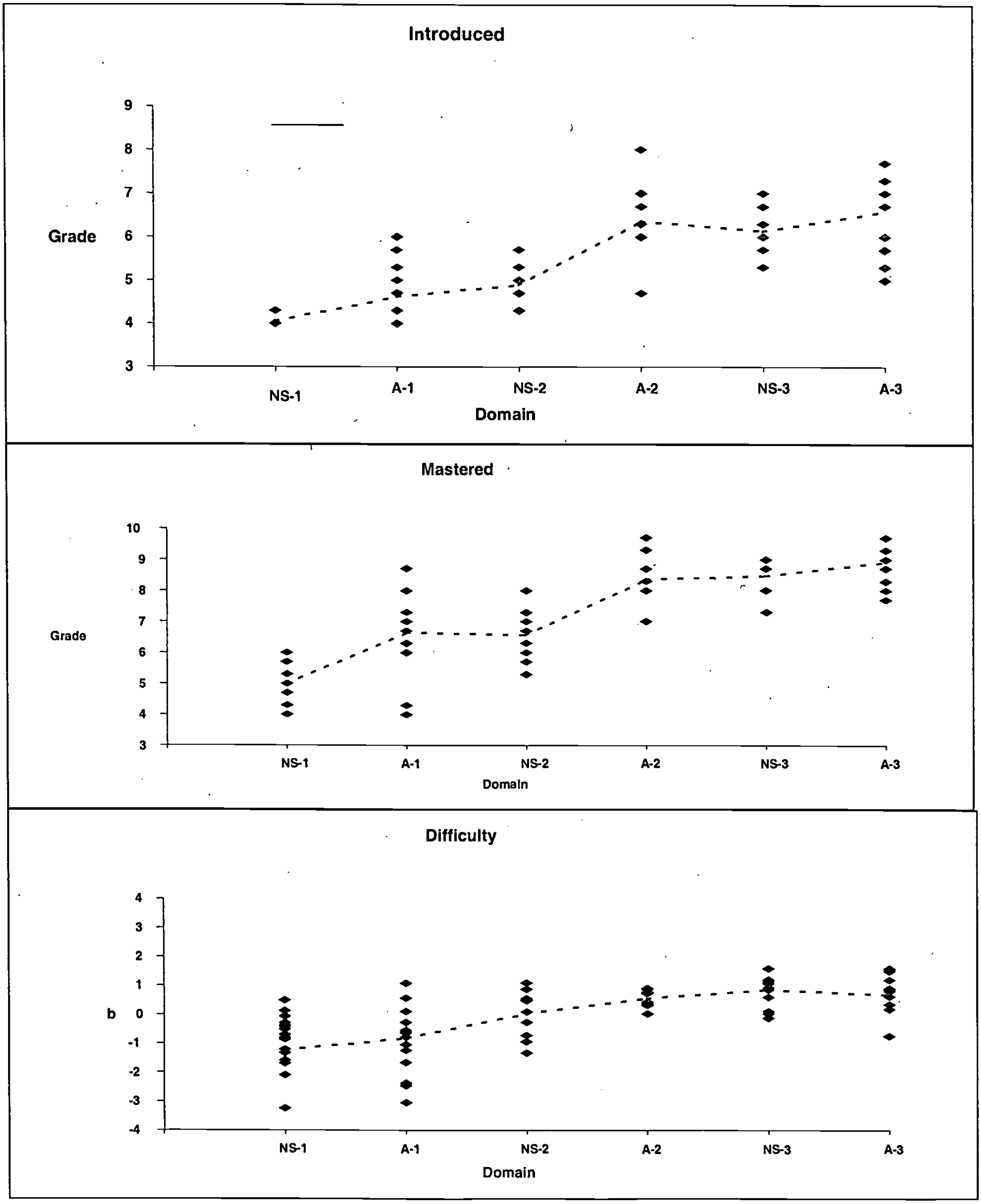
Figures 13 and 14 show that useful order relationships might exist among domains from different content strands. For example, in Figure 13 we see that skills measured by items within Number Sense Domain 1 tend to be introduced earlier, mastered earlier, and have lower difficulty (b-values) than skills measured by items within Algebra Domain 1. In Figure 14, we see that the characteristic curve for Number Sense Domain 1 is lower or 'easier' everywhere on the NAEP scale than the curve for Algebra Domain 1. It is conceivable that skills within Number Sense Domain 1 might be considered essential for mastering the skills within Algebra Domain 1. Other order relationships shown in Figures 13 and 14 invite similar speculation.

Table 11. Expected Percent-correct Domain Scores at Boundaries of Achievement Levels

		Achievement Level Boundary		
Content Strand	Domain	Basic	Proficient	Advanced
Number Sense	NS-1	70	86	94
	NS-2	45	68	86
	NS-3	31	50	73
Measurement	M-1	62	77	88
	M-2	21	38	59
Geometry	G-1	65	81	88
	G-2	49	68	82
	G-3	26	44	64
Data Analysis	DA-1	82	94	98
	DA-2	37	56	74
	DA-3	28	48	69
Algebra	A-1	63	81	92
	A-2	34	57	78
	A-3	30	30	73

Note: Shaded domain scores correspond to "mastery" of a domain based on a criterion domain score of 65% or .65.

Figure 13. Instructional Timing and Difficulty of Items by Within-Strand Number Sense and Algebra Domains

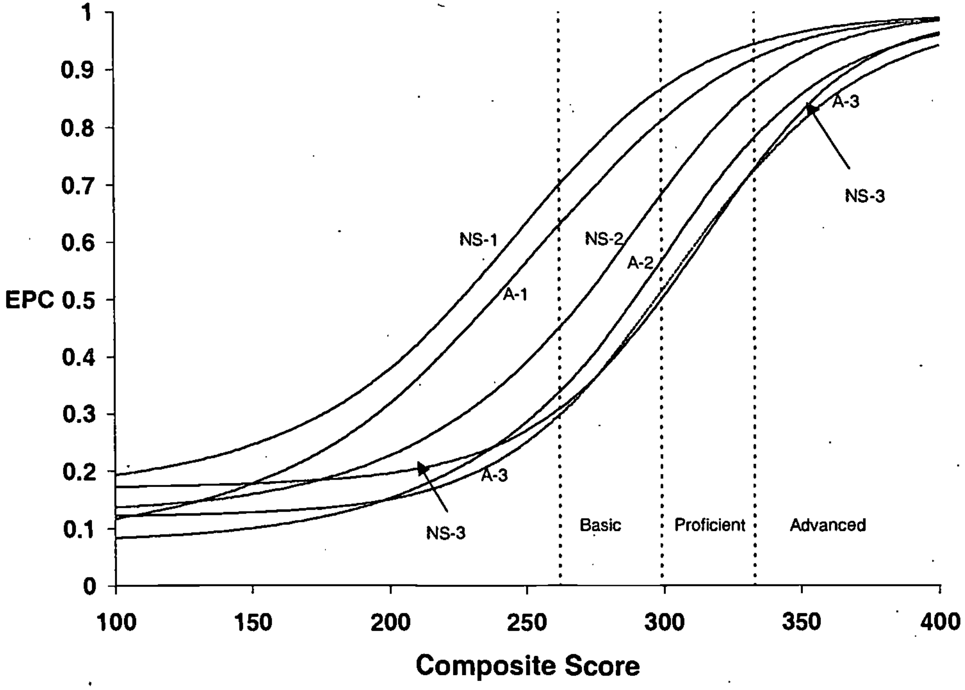


Introduced

Mastered

Difficulty

Figure 14. Within-Strand, Number Sense and Algebra Domains



Other order relationships may be less clear or essential. Figure 14 shows that Domain 3 within the Number Sense strand (NS-3) is comparable in difficulty to Domain 3 within the Algebra strand (A-3) everywhere on the scale (except in lower tails). However, such results might also be interesting because, in combination with Figure 13, they identify different domains of skills, not necessarily hierarchically related, that are taught and mastered at about the same time in a standard mathematics curriculum.

Discussion

Some of the basic elements in how multiple domains in this study were defined are consistent with what experts in the field of criterion-referenced testing have advocated as a way to improve the instructional consequences of criterion referenced tests. The post-hoc, item-by-item verification of relevance to a particular domain, based on a general description of a paragraph or two, as performed by the curriculum panelists in the present study, is in line with the "postcursive judgements" recommended by Popham (1994). Popham proposed that a new stage, which he called a "verification procedure," be added to criterion-referenced test development. He states:

> "I recommend we spend less energy on precursive delineation of what can be assessed and, instead, give more systematic attention to postcursive judgements regarding the degree to which test items yield supportive evidence concerning the score-based inference at issue." (p 30).

The postcursive judgements could be thought of as a validity check on the test development activities.

Popham recommended that the postcursive process involve a panel of 10-20 qualified judges. The present study, with only 3 curriculum panelists, one teacher, and one reviewer, can only be regarded as a step in this direction. In addition to using more experts, more definitive work should incorporate an explicit definition of the population of experts on which the domains are based and use a careful sample of the population. Assuming the work of the judges is to determine item membership in the domains, as opposed to defining the domains, clear operational definitions can be used to assign items to domains based upon the ratings. The definition of the rater population and the operational definitions for assigning items to domains would be as important a component of the domain descriptions as the summary paragraphs describing the cognitive skills represented by the domains.

Popham also recommended that the intellectual essence of what is required by the test's items be provided in a paragraph or two and that "this boiled down, general description of what's going on in the successful examinee's head be accompanied by a set of varied, but not exhaustive illustrative items." (pp. 17-18). The domain descriptions in this study are essentially the "paragraph or two" that Popham recommends. We intended to accompany our final domain descriptions with illustrative items. Unfortunately, none of the items from this particular assessment (Year 2000, Grade 8 mathematics) are approved for release into the public domain. However, from the NCES website, we have printed over 100 items from previous Grade 8 NAEP assessments. From these, we expect to be able to find items that illustrate at least some of the domains defined in this study.

The notion that items from other NAEP assessments (i.e., 1996, 1992, 2004) could illustrate the domains in this study is consistent with the concept of a domain as a construct to which a "universe" of items apply. Taking this notion a step further, it is conceivable that items from other assessments could be classified into the domains of this study and, if their item calibrations can be brought onto the same theta scale, domain scores could be re-estimated with these additional items. It might be possible, then, to define domains at a greater level of detail—perhaps to separate out the F1 domains that were judged to be too small in the present study.

We have not attempted to assess the measurement error associated with the domain scores and characteristic curves in this study. If one wishes to generalize the results of this study beyond the specific items in this assessment, then the item parameters must be viewed as random effects and one must consider sampling error. If one wishes to associate the domain scores only with the items used in this study, one must still consider the error inherent in the item statistics. This is not an exhaustive list of sources of error that could substantially affect the appearance of the domain characteristic curves reported here.

One source of error that was taken into account in this study concerns the effect of item type. Popham (1984) commented that score-based inferences must be supported by a variety of item formats, not just a specific type of item conforming to narrow item specifications and item-writing guides. We specifically took the effect of extended response items into account because we did not want to define domains whose order relationships depended on whether extended response items were included. But it is also possible that extended response items represent a particular kind of cognitive ability inherent in the meaning of some domains. Exclusion of extended response items could alter the essential character of a domain.

One of the curriculum panelists cautioned that the within-strand domains defined in this study may not be applicable to the next NAEP assessment. There is no guarantee that domains defined by a post-hoc analysis of items will be represented by items in the next assessment. In this sense, our method is different from what Popham envisioned, as he imagined that item writers would use the domain descriptions in test development. Perhaps domains like those defined in this study should ultimately be incorporated into the assessment framework. But we would also argue that domains should be defined broadly enough to accommodate minor changes in test plans and item-writing guides. This argument is consistent with Popham's notion that instructionally relevant, score-based inferences are better supported by general descriptions of domains, like those given in this study, and a postcursive process than by detailed test specifications.

Our results support the idea of using information about instructional timing to define domains. We do not suggest that ratings of instructional timing be the sole, or even primary consideration in constructing such domains. This is obvious from the extensive overlap in the instructional timing of items across our domains (Figures 3 through 7). But ratings of instructional timing were important for resolving differences in domain definitions and item classifications across experts. If the instructional timing of items within a given NAEP topic or subtopic, or within a newly-created domain, varied widely, the experts were more willing to accept that the items differed in some important respect and to assign them to different domains.

A specific result supporting the use of instructional timing is that the difficulty order of our domains generally agrees with their average instructional timing ratings. Such consistency may be useful in helping educators develop expectations about how their students would perform on NAEP, given the mathematics curriculum the students are exposed to.

Characteristic curves based solely on ratings of instructional timing (Figure 1) tended to have the desirable properties that make domain characteristic curves a useful descriptive tool. That is, with one exception (introduced at grade 6 versus grades 7 or 8), they were logically ordered and non-crossing. This result suggests that ratings of instructional timing will prove useful in defining domains that have non-overlapping characteristic curves.

The consistency of ratings of instructional timing among the curriculum panelists encourages the use of similar rating methods in future studies. The panelists were free to use any text book series (as some did) and any other materials they desired to provide their ratings. They were also given considerable freedom with regard to what their Mastered ratings meant. Given the freedom panelists had with their ratings, plus the fact that there is probably natural variation across schools, states, and text book series in the instructional timing for any skill, even in a standard curriculum, the inter-rater consistency found in this study is truly noteworthy. It is not clear to us that narrower operational definitions of what raters mean by "Introduced" or "Mastery" are of value if they do not result in better agreement among raters.

Although consensus about domain has not been completely documented at this point, it appears likely that a reasonable consensus and level of comfort will emerge among experts. This also is an important result of the study—the idea that a consensus is possible. There are many ways items within a test can be classified into domains. Items can be organized differently for different purposes. The value of a particular set of domains depends on how easily understood and widely shared their meaning is, and on statistical features of the domains as a set, such as the appearance of a plot of their characteristic curves.

The tendency of the domain characteristic curves in this study to be noncrossing shows that the descriptive power of a Guttman scale can be realized at a higher level of analysis, and perhaps more persuasively, than previously thought. Previously, Guttman-consistent patterns of mastery have been recognized in IRT only when the response functions of individual items did not cross (Andrich, 1985; Masters, et al., 1994; Wilson, 1989; Wright & Stone, 1979). These measurement theorists interpreted noncrossing item response functions as evidence for Guttman patterns of mastery among all examinees. If the IRT model for noncrossing item characteristic curves fit the data, a Guttman pattern of mastery was presumed to hold true for every student. This inference is not persuasive to those who question the fit of the model to the data or who give the item true score a population-level [conditional on ability], rather than student-level, interpretation (e.g.

Holland, 1990). A student-level interpretation of a domain score is not only more plausible, it is observable. In particular, teachers and other school-level consumers of achievement test information, are probably most familiar with percentage correct scores—to which domain scores immediately relate—as a form of reporting achievement test results. Also, as shown in this study by the use of item statistics from IRT models that allow item characteristic curves to cross, the fit of the IRT model to the data may be less of an issue in assessing the validity of noncrossing domain, as opposed to item, characteristic curves. The use of an IRT model that allows item characteristic curves to cross, allows domain characteristic curves to cross. So if noncrossing domain characteristic curves are observed, they are a function of the data, not the model. On this basis, the discovery of noncrossing domain characteristic curves may be more persuasive.

Despite the importance we attach to the fact that most of the domain characteristic curves did not cross in this study, overlapping curves would still support a pattern of mastery classifications like the one suggested in Table 11. They would support, for example, the notion that a student in the Advanced level has mastery of all domains mastered by a student at the Proficient level, plus at least one more. The only difference is that if characteristic curves crossed, the order in which domains are mastered would depend on the percentage correct criterion used for mastery. This could raise doubt about the meaning and general usefulness of the domains.

If multiple domains are adopted as a way of describing or defining achievement levels in broad based assessments, we recommend that plots of domain characteristic curves be used as a "point of entry" into the meaning of achievement test results for lay users. They show the relationship between two units of measurement having immediate meaning or consequence for lay users—scale score units and percentage (or proportion) correct score units. They communicate the conditional, if not overall, difficulty order relationship among the domains and they suggest criterion-referenced interpretations of what students at a given level of achievement can or cannot be expected to do. Vertical separation between multiple DCCs at any point on the scale corresponds to a range of percentage correct, or mastery scores. With this meaning, a plot of domain characteristic curves leads naturally to a more detailed presentation of test results, such as might be conveyed by the additional use of item maps, verbal descriptions, and various summary statistics including the instructional timing of items associated with the domains.

In sum, a multiple-domain score method of describing achievement on a broad-based assessment is recommended by this study. The most important component of the method involves using domains, rather than items, as the unit of analysis for representing the relationship between content and achievement. Domains represent a higher level of analysis for summarizing the relationship between item content and item statistics by means of formal sampling procedures and operational definitions. The second most important component of the method recommended here is to represent the relationship between content and achievement through a plot showing the characteristic curves of multiple domains. Vertical separation of the curves at any point on the scale communicates immediately what examinees can or cannot be expected to do—score higher or lower than a given percentage correct score on the domains. Percentage correct scores are readily understood by teachers and are easy to understand in terms of a particular student having partial mastery of content. Such plots will encourage users to invest time in learning more about the domains and, as a result, understanding the curricular and instructional implications of the test's results. . If domains represent a curricular sequence, even approximately, the curves are likely to be noncrossing. In that case, the domains will for all practical purposes comprise a Guttman scale.

References

American College Testing (1993, February). *Description of mathematics achievement levels-setting process and proposed achievement level descriptions (Volume 1).* Iowa City, IA: Author.

Allen, N. L., Carlson, J. E., and Zelenack, C. A., (1999). *The NAEP 1996 Technical Report.* NCES 1999-452. Washington DC: U.S. Department of Education, Office of Educational Research and Improvement.

Andrich, D. (1985). An elaboration of Guttman scaling with Rasch models for measurement. In N. Tuma (Ed.), *Sociological Methodology* (pp 33-80). Jossey-Bass.

Beaton, A. E. & Allen, N. L. (1992). Interpreting scale through scale anchoring. *Journal of Educational Statistics*, *17*, 191-204.

Bock, R. D. (1997). A brief history of item response theory. *Educational Measurement: Issues and Practice*, *16*(4), 21-33.

Bock, R. D., Thissen, D., & Zimowski, M. F. (1997). IRT estimation of domain scores. *Journal of Educational Measurement*, *34*(3), 197-211.

Boone, W. J. (April, 1990). *How psychometric analysis can help teachers realize a curriculum.* Paper presented at the Annual Meeting of the American Educational Research Association. Boston, MA.

Cliff, N. (1983). Evaluating Guttman Scales: Some old and new thoughts. In H. Wainer & S. Messick (Eds.), *Principles of modern psychological measurement: A festschrift for Frederick M. Lord.* Hillside, NJ.

Donoghue, J. R. (March, 1997). *Item mapping to a weighted composite scale.* Paper presented at the annual meeting of the American Educational Research Association; Chicago, IL.

Forsyth, R. A. (1991). Do NAEP scales yield valid criterion-referenced interpretations? *Educational Measurement: Issues and Practice*, *10*(3), 3-9.

Educational and Psychological Measurement: Issues and Practice, *10*(3), 3-9.

Fox, J. E. & Tipps, R. S. (1995). Young children's development of swinging behaviors. *Early Childhood Research Quarterly*, *10*, 491-504.

Guttman, L. (1950). The basis for scalogram analysis. In S. A. Stouffer, L. Guttman, E. A. Suchman, P. A. Lazarsfeld, S. A. Star, & J. A. Clausen (Eds.), *Measurement and Prediction* (pp 60-90). Princeton: Princeton University Press.

Hambleton, R. K., Swaminathan, H., Algina, J., & Coulson, D. B. (1978). Criterion-referenced testing and measurement: A review of technical issues and developments. *Review of Educational Research*, *48*(1), 1-47.

Holland, P. W. (1990). On the sampling theory foundations of item response theory models. *Psychometrika*, *55*, 577-601.

Impara, J. C. & Plake, B. S. (1998). Teacher's ability to estimate item difficulty: A test of the assumptions of the Angoff standard setting method. *Journal of Educational Measurement*, *35*, 69-81.

Janssen, R., Tuerlinckx, F., Meulders, M., & Boeck, P. (2000). A hierarchical model for criterion-referenced measurement. *Journal of Educational and Behavioral Statistics. 25*, 285-306.

Katz, S., & Akpom, C. A. (1976). A measure of primary sociobiological functions. *International Journal of Health Services*, *6*(3), 493-507.

Kolstad, A. (April, 1996). *The response probability convention embedded in reporting prose literacy levels from the 1992 National Adult Literacy Survey.* Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Lewis, D. M., Mitzel, H. C., & Green, D. R. (June, 1996). Standard setting: A bookmark approach. In D. R. Green (Chair), *IRT-based standard setting procedures using behavioral anchoring*. Symposium presented at the Council of Chief State School Officers National Conference on Large-Scale Assessments. Phoenix, AZ.

Masters, G. N., Adams, R. & Lokan, J. (1994). Mapping student achievement. *International Journal of Educational Research*, *21*, 595-609.

Mislevy, R.J. (1998). Implications of market-basket reporting for achievement-level setting. *Applied Measurement in Education.* *11*(1), 49-63.

National Assessment Governing Board (2000). *Mathematics Framework for the 1996 and 2000 National Assessment of Educational Progress.* Washington DC: Author.

National Assessment Governing Board (1990). *Setting appropriate achievement levels for the National Assessment of Educational Progress: policy framework and technical procedures.* Washington DC: Author.

National Center for Education Statistics (1997). *NAEP 1996 Mathematics Report Card for the Nation and the States.* Washington DC: Author.

National Center for Education Statistics (2001, August). *The Nation's Report Card: Mathematics 2000.* Washington DC: Author.

National Council on Measurement in Education) (1992). 1992 NCME Annual Meeting Highlights. *Educational Measurement: Issues and Practice*, *11*(2), 30.

Pommerich, M., Nicewander, W.A., & Hanson, B. A. (1999). Estimating average domain scores. *Journal of Educational Measurement*, *36*, 199-216.

Popham, W. J. (1994). The instructional consequences of criterion-referenced clarity. *Educational Measurement: Issues and Practice*, *13* (4).

Schulz, E.M., Kolen, M. & Nicewander, W.A. (1999). A rationale for defining achievement levels using IRT-estimated domain scores. *Applied Psychological Measurement.* *23*, 347-362.

Schulz, E. M., Perlman, C., Rice, W. K., Jr., & Wright, B. D. (1992). Vertically equating reading tests: An example from Chicago Public Schools. In M. Wilson (Ed.), *Objective Measurement: Theory into Practice* Vol 1, (pp 138-154). Norwood, NJ: Ablex Publishing Corporation.

Wilson, M. (1989). A comparison of deterministic and probabilistic approaches to measuring learning structures. *Australian Journal of Education*, *33*(2), 127-140.

Wright, B. D., and Stone, M. H. (1979). *Best Test Design.* Chicago: MESA Press.

APPENDIX A

Computation of Domain Characteristic Curves

Let η represent the NAEP mathematics true scale score, and let D represent the score on a domain represented by two or more items. The goal is to compute and plot $E(D \mid \eta)$ --the expected domain score conditional on η. The item parameters obtained from the NAEP test development contractor pertain only to the subscales, since the NAEP mathematics data were analyzed separately within each of five content areas. Ability on subscale *j*, *j*=1,...,5 is denoted θ_j. The NAEP composite math scale, η is related to the subscale $\theta = \{\theta_1, ..., \theta_5\}$ through the transformations:

$$\boldsymbol{y} = A\theta + \boldsymbol{b},$$

and

$$\eta = \boldsymbol{w}'\boldsymbol{y}, \tag{2}$$

where A is a diagonal matrix of constants, $\boldsymbol{b}$ is a column vector of constants, and $\boldsymbol{w}$ is a column vector of weights summing to 1.

To represent dichotomous and polytomous items without distinction let

$$E(U_{ij} \mid \theta_j) = \sum_{h=0}^{m_i} \left[hP(U_{ij} = h \mid \theta_j) \right]$$

be the conditional-on-θ_j expected score on any item, *i*, calibrated to subscale *j*, where m_i is the number of categories for item *i*, and *h* is the item score having values ranging from 0 to $m_i - 1$. $P(U_{ij} = h \mid \theta_j)$ was obtained from the three-parameter logistic IRT model for multiple choice items (*h*=0,1). For open-ended items, the two-parameter logistic IRT model was used for items scored with only two categories (h=0,1), or a generalized partial credit model (Muraki, 1992) was used for items scored with more than two categories (h=0,1,..., $m_i - 1$).

In mapping NAEP items to the composite scale, Donoghue (1997) describes the direct regression method, which models the regression of item scores on η. The conditional-on-η expected score on item i can be defined as:

$$E(U_{ij} \mid \eta) = \int_{-\infty}^{\infty} E(U_{ij} \mid \theta_j) f(\theta_j \mid \eta) d\theta_j .$$

Given $E(U_{ij} \mid \eta)$, for all items in the domain, the expected conditional domain score is:

$$E(D \mid \eta) = \sum E(U_{ij} \mid \eta) .$$

We assume that both the conditional and marginal distributions of the subscale θ_j are normal density functions: $\theta_j \mid \eta \sim N(\mu_{j|\eta}, \sigma^2_{j|\eta})$ and $\theta_j \sim N(\mu_j, \sigma^2_j)$. To compute $E(U_{ij} \mid \eta)$ in Equation (4), values of $E(U_{ij} \mid \theta_j)$ were averaged over $f(\theta_j \mid \eta)$ by numerical quadrature. The mean, $\mu_{j|\eta}$, and standard deviation, $\sigma^2_{j|\eta}$, of the conditional θ_j distribution were computed based on Equation (6) below. Using $\mu_{j|\eta}$ and $\sigma^2_{j|\eta}$, the minimum and maximum quadrature points were found as $\mu_{j|\eta} \pm 3\sigma^2_{j|\eta}$, and 40 equally spaced quadrature points were employed. Although η ranges from 0 to 500, we used 100 through 400 with an increment of one.

Under the normality assumption, Donoghue (1997) shows that

$$f(\theta_j \mid \eta) = N\left(\mu_j + \frac{\sigma_j \rho_{j\eta}(\eta - \mu_\eta)}{\sigma_\eta}, \sigma_j^2(1-\rho_{j\eta}^2)\right), \tag{6}$$

where μ_η and σ_η are the mean and standard deviation of ηs over the population of interest and $\rho_{j\eta}$ is the correlation between θ_j and η.

The values of $\rho_{j\eta}$ needed to compute $\mu_{j|\eta}$ and $\sigma_{j|\eta}^2$ in Equation (6) were obtained as follows. Let $\boldsymbol{c}' = \boldsymbol{w}'\boldsymbol{A} = \{c_1, ..., c_5\}$. Then,

$$\begin{aligned} Cov(\theta_j, \eta) &= \sum_{k=1}^{5} c_k Cov(\theta_j, \theta_k) \\ &= \sum_{k=1}^{5} \frac{c_k \rho_{jk}}{\sigma_j \sigma_k}, \end{aligned} \tag{7}$$

where ρ_{jk} is the correlation between θ_j and θ_k. It follows that

$$\rho_{j\eta} = \frac{Cov(\theta_j, \eta)}{\sigma_j \sigma_\eta}. \tag{8}$$

Values needed to compute the domain characteristic curves as outlined above were obtained from the NAEP test development contractor. Values used for the mean and standard deviation of η were, respectively, 275 and 37. Other values are shown in Tables A1 and A2. The numbered subscales are identified with content strands as follows: 1) Number Sense, 2) Measurement, 3) Geometry, 4) Data Analysis, and 5) Algebra. All values reflect marginal distributions of theta-metric plausible values.

Table A1. Marginal Subscale Theta Distributions and Transformation Constants

Subscale	Theta		Transformation constants and weights to form composite		
	Mean	S.D.	Slope	Intercept	Weight
1	.0261	1.0197	36.152	275.240	.25
2	.0179	1.0648	45.172	272.541	.15
3	.0337	1.0334	32.841	271.461	.20
4	.0563	1.0274	41.128	275.740	.15
5	.0464	1.0071	35.790	275.974	.25

Covariances and correlations for each pair of scales, based on theta-metric plausible values are shown in Table A2.

Table A2. Subscale Covariances and Correlations

Scales	Covariance	Correlation
1,2	.9720	.8954
1,3	.9103	.8639
1,4	.9774	.9331
1,5	.9506	.9258
2,3	.9971	.9062
2,4	.9873	.9026
2,5	.9790	.9129
3,4	.9284	.8745
3,5	.9412	.9045
4,5	.9590	.9268

APPENDIX B

Instructions and Form for Ratings of Instructional Timing

Item Rating Instructions

Please rate all of the items within a block before proceeding to another block. For each item:

1) Read the item. Review the scoring key or the scoring rubric and sample response for each item to be certain that you know the correct response and how the scoring rubrics were applied.

2) Think of the knowledge and skills that are required to answer the item correctly or to receive full credit. Review the framework to refresh your memory of the content strands and cognitive skills for the Mathematics NAEP. Make notes on the item pages to help with the next step. If you think that more than one knowledge area or skill is required to correctly answer an item, please make note of that.

3) Rate each item on the basis of your knowledge of the *traditional mathematics curriculum* for an *average student* in the United States and the <u>*most* difficult or complex skill</u> needed to answer the item correctly,as identified in Step 2.

- Place an "I" at the grade level the skill is introduced.
- Place an "R" at the grade level the skill is reinforced.
- Place an "M" at the grade level the skill is mastered.

Please record an I and an M for every item even if they occur in the same grade. An R may be used more than once, but should be indicated only in grades between an I and an M, if applicable. Mastery implies the skill will not subsequently be specifically taught or reinforced in the curriculum.

Indicate your ratings on the following pages. You may make copies of these pages, or use a pencil and eraser, to revise your ratings if desired. To the extent possible, base your ratings on *your own evaluation* and do not consider information that you may have available about the items' NAEP framework classifications for your ratings.

Block 3

Item	Ratings						
No.	below 5th	5th	6th	7th	8th	9th	above 9th
1							
2							
3							
4							
5							
6							
7							
8							
9							

[Additional rating forms, one page per block, accompanied these materials

APPENDIX C

F1 Domains

Number Sense Strand

1) Basic Computations with Positive Whole Numbers.

These items give positive whole numbers on which the students are directed to perform a computation (add, subtract, multiply, or divide). The numbers do not have to be extracted from a context or problem situation. If items require multiple operations, parentheses are used to specify the order of the operations.

2) Rounding, Place Value, and Contextual Addition and Subtraction of Integers.

The items in this domain present integers (negative or positive whole numbers) in situations requiring addition and subtraction. Also included in this domain are items requiring rounding (i.e., of decimals to whole numbers) and understanding of place value (ones, tens, hundreds).

3) Models for Numbers

These items utilize area and number line models for rational numbers.

NS-1

4) Contextual Multiplication and Division of Positive Whole Numbers; Basic Multiplication and Division with Integers; Order of Operations.

Items in this domain present positive whole numbers in problems that require multiplication or division for solving. Some items might only ask the student to specify the operation multiplication or division) needed to solve the problem. Non-contextual items (numbers are simply given) with negative whole numbers, and items requiring the performance of multiple operations in correct order are included.

---

5) Decimals

Items in this domain involve basic computation (add, subtract, multiply, or divide) with decimals. The numbers may or may not occur in a context (problem solving). These problems involve more than rounding. They require the student to know or correctly place the decimal in the solution.

6) Fractions and Ratios

NS-2

These items require understanding or skill in the use of fractions, ratios, and proportions. Items involving problem solving and comparing fractions are included. Problems involving time (hours and minutes) are included in this domain because they require equating number of minutes to a fraction of an hour.

---

7) Rates and Percents

These items involve the solution of problems using the percents and rates.

8) Number Theory

These items involve working knowledge of elementary number theory concepts, including odd/even numbers, primes, multiples, integers, and consecutive integers. It also includes involving reasoning with numbers in the context of largest/smallest.

NS-3

9) Scientific Notion

These items involve working with numbers in scientific notation, and recognizing them in equivalent forms.

Measurement Strand

1) One dimensional-Measurement

Items in this domain involve one-dimensional measures, including length (distance, perimeter), weight, and reading scales, with attention to units.

2) Angle

Items in this domain use the concept of angle, both measurement and construction. Included are any items involving comparisons of sizes of given angles.

M-1

3) Understanding Area

Items in this domain require understanding of area and equivalent areas. They may require finding the area of simple figures by adding whole and half-units together. They may require knowing how to compute the area of a square or rectangle, or estimating the area of a familiar rectangular object or place. But they differ from domain 4 in that little, if any computation is required beyond that of a rectangle.

4) Computation of Area

Items in this domain require more complex computations to solve area problems. They may require knowledge of formulas for triangles. They may require finding ways to subdivide a given area into smaller areas, or drawing figures of equivalent or proportional areas where the solution is based on more than similarity of shape.

5) Surface Area, Volume

M-2

These items contain problems which use the concepts of surface area and volume. They may involve computing several areas and adding them together, or higher level spatial-visualization skills.

Geometry Strand

1) Properties of Figures

Items in this domain involve knowledge of simple properties, including angle measure, classification of triangles and quadrilaterals, and identification of figures.

G-1

---

2) Transformation and Symmetry

This domain contains items that involve single or multiple transformations, including reflections, translations, and rotations. Items can ask for completion of figures to have certain symmetry, finding reflections, or drawing lines of symmetry.

3) Spatial-Visual

These items involve abstract visualization, including the rearranging of or mentally transforming of shapes into different shapes, subdividing two-dimensional figures, and visualizing three-dimensional shapes – how they look flattened out, and how flattened out shapes fold into three – dimensional shapes.

G-2

4) Geometric Reasoning

These items involve more complex reasoning about properties of geometric figures or geometric situations themselves, but the reasoning is still based on manipulatives or visual comparison of figures. These items typically require combining manipulatives, altering figures, or drawing new figures that have geometric properties specified in the problem. Item 12, in Block 6 could be in Domain 4 or 5, but is placed in domain 4 because the formal definition of "similar" triangles is not actually needed to solve the problem. From the figure given, it is easy to visually recognize the proportionalities needed to solve the problem.

---

5) Problem Solving Using Formal Properties of Figures

These items involve the solution of problems using formal properties and formulas for geometric figures. Computation is typically required. Item 10 in Block 7 is in this domain because it requires computation of area. Item 17, Block 4 is in this domain because it can be solved without reference to the figure by knowledge of formal properties of a parallelogram.

G-3

Data Analysis Strand

1) Representing Data

Items in this domain test a student's ability to represent information in a simple histogram or frequency chart.

DA-1

2) Using Graphs and Charts

Items in this domain require the student to use information in graphs and charts (e.g. pie charts, frequency histograms, bivariate line charts). The student may have to decide which information in the chart is relevant to the question and/or interpret the chart to obtain the numbers needed to answer the question. The student may have to perform arithmetic operations on the numbers, or apply or obtain percentages, proportions, and rates of change.

DA-2

3) Probability and Statistics

Items in this domain require the student to analyze data using probability or statistical computation. The student may also have to give a sample space for a given situation or determine appropriate sampling technique. Some items test understanding of statistical terms including mean, median, mode, and range and determining whether one of these is more appropriate than another for a given situation; computation may or may not be required for these items.

DA-3

Algebra Strand

1) Basic Operations

Items in this domain require the student to apply basic computational skills to find solutions to simple open sentences. Knowledge of basic symbols (e.g. inequality symbols) and order of operations are necessary. No particular algebra skills are necessary for these items. Problems may be modeled on a number line.

2) Logic, Simple Patterns, and Algebraic Reasoning

A-1

These items require students to recognize simple logic or numeric patterns, above the level of simple arithmetic sequences. Logic problems include synthesis of given information to make conclusions. Algebraic reasoning items include Item 10 in Block 4, in which symbols (circle, triangle, square) are associated with different amounts of weight and the solution is associated with balancing a scale.

---

3) Coordinate Systems

A-2

Items in this domain involve the use of coordinate systems, both one- and two dimensional. Locating points with numbers or ordered pairs is a necessary skill. Some items include continuing a given pattern(s) in order to solve a more complicated problem. Items may require reasoning given information about points on a coordinate system.

---

4) Variables in Expressions

This domain consists of items which present the student with expressions or situations with variable expressions, or require the student to interpret such expressions.

A-3

5) Solving Equations

These items require the solving of equations and inequalities which may involve algebraic manipulation. The student may have to set up the equation or formula based on given information.

6) Pattern Solving Requiring Complex Reasoning and Computational Skill

These items involve pattern recognition in more difficult situations, including fractional patterns that are not arithmetic, or patterns involving exponents.

APPENDIX D

NAEP Items Classified by F1 domains

(Number Sense Only)

Number Sense. F1-1: Basic Computations with Positive Whole Numbers.

These items give positive whole numbers on which students are directed to perform a computation (add, subtract, multiply, or divide). The numbers do not have to be extracted from a context or problem situation. If items require multiple operations, parentheses are used to specify the order of the operations.

Examples

4. (150 ÷ 3) + (6 × 2) =

A) 10
B) 58
C) 62
D) 112

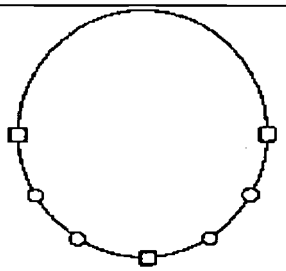
20. 503 - 207 =

A) 206
B) 296
C) 304
D) 396

Number Sense F1-2. Rounding, Place Value, and Contextual Addition and Subtraction of Integers.

The items in this domain present integers (negative or positive whole numbers) in situations requiring addition and subtraction. Also included in this domain are items requiring rounding (i.e., of decimals to whole numbers) and understanding of place value (ones, tens, hundreds).

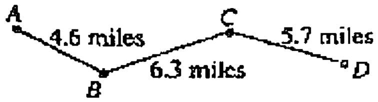
Examples



Each □ costs 6¢
Each ○ costs 4¢

13. If the string does not cost anything, how much does the necklace above cost?

A) 10¢
B) 24¢
C) 28¢
D) 34¢



14. Carol wanted to estimate the distance from A to D along the path shown on the map above. She correctly rounded each of the given distances to the nearest mile and then added them. Which of the following sums could be hers?

A) 4 + 6 + 5 = 15
B) 5 + 6 + 5 = 16
C) 5 + 6 + 6 = 17
D) 5 + 7 + 6 = 18

15. There are 50 hamburgers to serve 38 children. If each child is to have at least one hamburger, at most how many of the children can have more than one?

A) 6
B) 12
C) 26
D) 38

17. By how much would 217 be increased if the digit 1 were replaced by a digit 5?

A) 4
B) 40
C) 44
D) 400

21. What number is four hundred five and three-tenths?

A) 45.3
B) 405.3
C) 453
D) 4,005.3

31. Which of the following is closest to 15 seconds?

A) 14.1 seconds
B) 14.7 seconds
C) 14.9 seconds
D) 15.2 seconds

32. The census showed that three hundred fifty-six thousand, ninety-seven people lived in Middletown. Written as a number, that is

A) 350,697
B) 356,097
C) 356,907
D) 356,970

34. The length of a dinosaur was reported to have been 80 feet (rounded to the nearest 10 feet). What length other than 80 feet could have been the actual length of this dinosaur?

Answer:____________________ feet

36. Which of the following numbers, when rounded to the nearest thousand, becomes 27,000?

A) 26,099
B) 26,490
C) 27,381
D) 27,550
E) 27,640

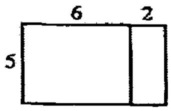
Number Sense F1-3. Models for Numbers

These items utilize area and number line models for rational numbers.

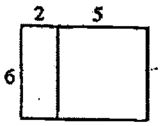
Examples

2. Which of the following figures best illustrates the statement
$5 \times (6 + 2) = (5 \times 6) + (5 \times 2)$?

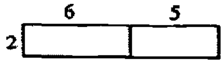
A)



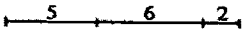
B)



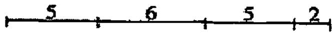
C)

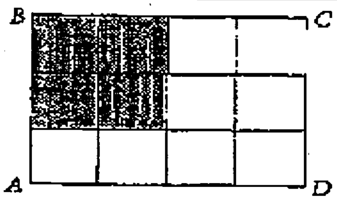


D)



E)





37. In the figure above, what fraction of rectangle ABCD is shaded?

A) 1/6

B) 1/5

C) 1/4

D) 1/3

E) 1/2

Number Sense F1-4. Contextual Multiplication and Division of Positive Whole Numbers; Basic Multiplication and Division with Integers; Order of Operations.

Items in this domain present positive whole numbers in problems that require multiplication or division for solving. Some items might only ask the student to specify the operation multiplication or division) needed to solve the problem. Non-contextual items (numbers are simply given) with negative whole numbers, and items requiring the performance of multiple operations in correct order are included.

Examples

18. Christy has 88 photographs to put in her album. If 9 photographs will fit on each page, how many pages will she need?

A) 8
B) 9
C) 10
D) 11

23. Jill needs to earn $45.00 for a class trip. She earns $2.00 each day on Mondays, Tuesdays, and Wednesdays, and $3.00 each day on Thursdays, Fridays, and Saturdays. She does not work on Sundays. How many weeks will it take her to earn $45.00?

Answer:______________________

29. Raymond must buy enough paper to print 28 copies of a that contains 64 sheets of paper. Paper is only available in packages of 500 sheets. How many whole packages of paper will he need to buy to do the printing?

Answer:______________________

33. A club held a car wash and washed 21 cars. If the club raised $84, how much did it charge per car?

A) $0.25
B) $4.00
C) $5.00
D) $1,764.00

38. (-5)(-7) =

A) -35
B) -12
C) -2
D) 12
E) 35

41. Anita is making bags of treats for her sister's birthday party. She divides 65 pieces of candy equally among 15 bags so that each bag contains as many pieces as possible. How many pieces will she have left?

A) 33
B) 5
C) 4
D) 3
E) 0.33

Number Sense F1-5. Decimals

Items in this domain involve basic computation (add, subtract, multiply, or divide) with decimals. The numbers may or may not occur in a context (problem solving). These problems involve more than rounding. They require the student to know or correctly place the decimal in the solution.

Examples

5. What is the product of 3.12 and 83?

Answer: ____________________

9. Ground beef costs $2.59 per pound. What is the cost of 0.93 pound of ground beef?

A) $3.52
B) $2.78
C) $2.47
D) $2.41
E) $1.66

Number Sense F1-6. Fractions and Ratios

These items require understanding or skill in the use of fractions, ratios, and proportions. Items involving problem solving and comparing fractions are included. Problems involving time (hours and minutes) are included in this domain because they require equating number of minutes to a fraction of an hour.

Examples

1. If $\frac{2}{25} = \frac{n}{500}$, then n =

A) 10
B) 20
C) 30
D) 40
E) 50

6. The weight of an object on the Moon is 1/6 the weight of that object on the Earth. An object that weighs 30 pounds on Earth would weigh how many pounds on the Moon?

Answer: ____________________

19. Of the following, which is closest in value to 0.52?

A) 1/50
B) 1/5
C) 1/4
D) 1/3
E) 1/2

22. If 1 1/3 cups of flour are needed for a batch of cookies, how many cups of flour will be needed for 3 batches?

A) 4 1/3
B) 4
C) 3
D) 2 2/3

12. If $\frac{10.3}{5.62} = \frac{n}{4.78}$, then, of the following, which is closest to n?

A) 2.61
B) 3.83
C) 8.76
D) 8.82
E) 12.11

Number Sense F1-7 Rates and Percents

These items involve the solution of problems using the percents and rates.

Examples

7. Kate bought a book for $14.95, a record for $5.85, and a tape for $9.70. If the sales tax on these items is 6 percent and all 3 items are taxable, what is the total amount she must pay for the 3 items, including tax?

A) $32.33

B) $32.06

C) $30.56

D) $30.50

E) $ 1.83

10. The Zandalia Zoo uses 214,964 kilograms of meat per year. If the meat costs $2.53 per kilogram, how much does the meat cost per week?

Answer: ____________________

11. If the price of a can of beans is raised from 50 cents to 60 cents, what is the percent increase in the price?

A) 83.3%

B) 20%

C) 18.2%

D) 16.7%

E) 10%

26. Ken bought a used car for $5,375. He had to pay an additional 15 percent of the purchase price to cover both sales tax and extra fees. Of the following, which is closest to the total amount Ken paid?

A) $806

B) $5,510

C) $5,760

D) $5,940

E) $6,180

39. Of the following, which is the closest approximation of a 15 percent tip on a restaurant check of $24.99?

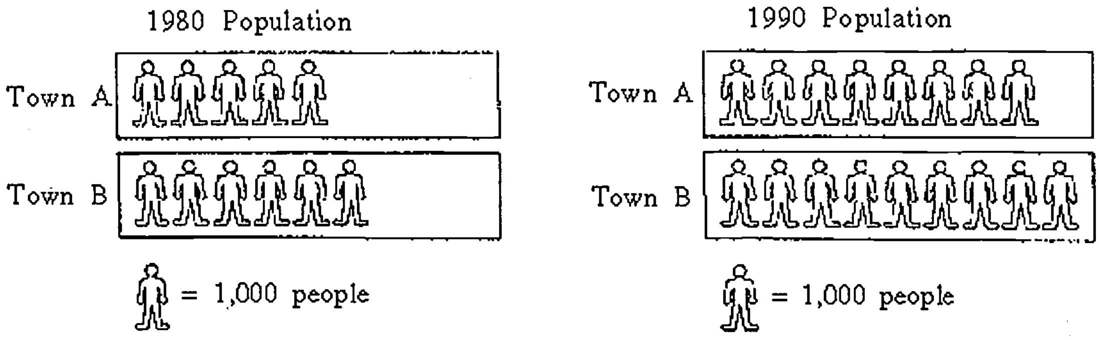
A) $2.50

B) $3.00

C) $3.75

D) $4.50

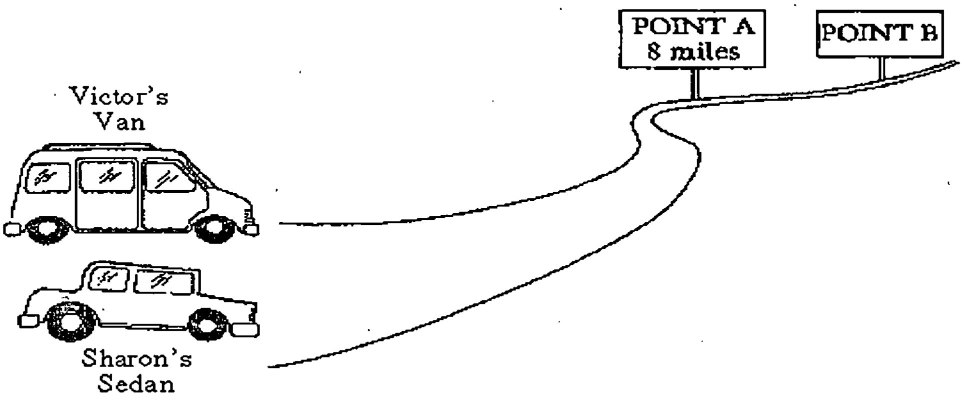
E) $5.00



43. In 1980, the populations of Town A and Town B were 5,000 and 6,000, respectively. The 1990 populations of Town A and Town B were 8,000 and 9,000, respectively.

Brian claims that from 1980 to 1990 the populations of the two towns grew by the same amount. Use mathematics to explain how Brian might have justified his claim.

Darlene claims that from 1980 to 1990 the population of Town A had grown more. Use mathematics to explain how Darlene might have justified her claim.



44. Victor's van travels at a rate of 8 miles every 10 minutes. Sharon's sedan travels at a rate of 20 miles every 25 minutes.

If both cars start at the same time, will Sharon's sedan reach point A, 8 miles away, before, at the same time, or after Victor's van?

Explain your reasoning.

If both cars start at the same time, will Sharon's sedan reach point B (at a distance further down the road) before, at the same time, or after Victor's van?

Explain your reasoning.

Number Sense F1-8. Number Theory

These items involve working knowledge of elementary number theory concepts, including odd/even numbers, primes, multiples, integers, and consecutive integers. It also includes involving reasoning with numbers in the context of largest/smallest.

Examples

3. The least common multiple of 8, 12, and a third number is 120. Which of the following could be the third number?

A) 15
B) 16
C) 24
D) 32
E) 48

8. If 12 divides a whole number n without a remainder, list all whole numbers greater than 1 and less than 12 that must also divide n without a remainder.

Answers: ____________________

25. What is the difference between the smallest positive 3-digit number and the largest positive 2-digit number?

A) 1
B) 9
C) 10
D) 90
E) 900

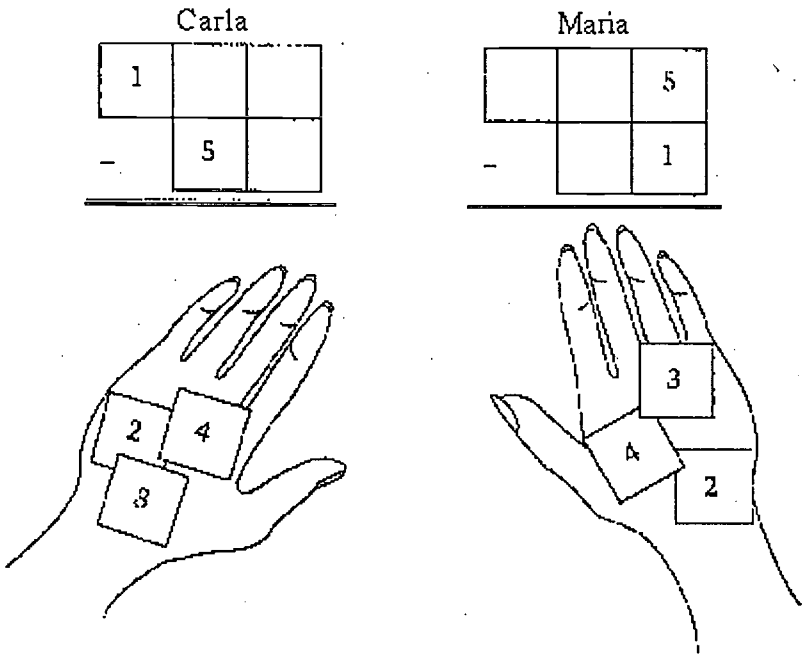
27. Which of the following is both a multiple of 3 and a multiple of 7?

A) 7,007
B) 8,192
C) 21,567
D) 22,287
E) 40,040

28. Tracy said, "I can multiply 6 by another number and get an answer that is smaller than 6." Pat said, "No, you can't. Multiplying 6 by another number always makes the answer 6 or larger." Who is correct? Give a reason for your answer.

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

40. In a game, Carla and Maria are making subtraction problems using tiles numbered 1 to 5. The player whose subtraction problem the largest answer wins the game. Look at where each girl placed two of her tiles.



Who will win the game?_____________________

Explain how you know this person will win.

42. If each of the counting numbers from 1 through 10 is multiplied by 13, how many of the resulting numbers will be even?

A) One
B) Four
C) Five
D) Six
E) Ten

Number Sense F1-9 Scientific Notion

These items involve working with numbers in scientific notation, and recognizing them in equivalent forms.

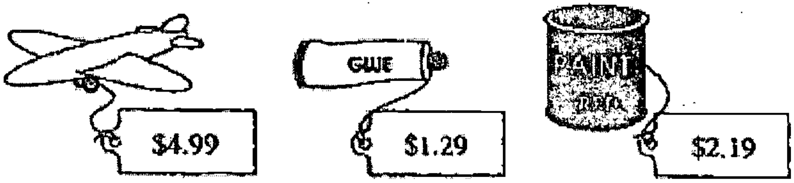
Examples

35. A certain reference file contains approximately one billion facts. About how many millions is that?

A) 1,000,000
B) 100,000
C) 10,000
D) 1,000
E) 100

Unclassified Items

Item 16: Might be placed in NS-F1-2 because it involves rounding, but the correct answer does not depend on the result of rounding, but rather, knowing when to estimate—knowing when an estimate will be sufficient.



16. Chen had $10 to buy a model plane, glue, and paint as shown above. At which of the following times could an estimate have been used instead of exact numbers?

A) When Chen tried to decide whether or not he had enough money to buy the plane, glue, and paint
B) When the clerk entered each amount into the cash register
C) When the clerk told Chen how much he owed
D) When Chen counted his change

Item 24: This item requires the computational skill in the Number Sense F1-2 domain (the pattern involves adding or subtracting a constant), but the most difficult skill required by this problem probably fits better into Algebra F1-2 domain—perceiving logical or algebraic relationships in simple patterns.

42, 51, 49, 58, 56, . . .

24. If the pattern in the list above continues, what will be the next number after 56?

A) 54
B) 63
C) 64
D)`65
E) 67

Item 30: Computationally, the skill required by this item is described in Number Sense F1-2 domain. But the most difficult aspect of this item may be the amount of reading and the amount of information that has to be sorted out to solve the problem. Perhaps the most difficult skill needed to solve this problem is that of taking notes—writing down the information in order to sort it out.

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

30. Treena won a 7-day scholarship worth $1,000 to the Pro Shot Basketball Camp. Round-trip travel expenses to the camp are $335 by air or $125 by train. At the camp she must choose between a week of individual instruction at $60 per day or a week of group instruction at $40 per day. Treena's food and other expenses are fixed at $45 per day. If she does not plan to spend any money other than the scholarship, what are all choices of travel and instruction plans that she could afford to make? Explain your reasoning.

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, *or*, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:

Address:

Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:

Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

ERIC CLEARINGHOUSE ON ASSESSMENT AND EVALUATION
UNIVERSITY OF MARYLAND
1129 SHRIVER LAB
COLLEGE PARK, MD 20742-5701
ATTN: ACQUISITIONS

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
4483-A Forbes Boulevard
Lanham, Maryland 20706

Telephone: 301-552-4200
Toll Free: 800-799-3742
FAX: 301-552-4700
e-mail: ericfac@inet.ed.gov
WWW: http://ericfacility.org