# Comparing Cognitive Representations of Test Developers and Students on a Mathematics Test With Bloom's Taxonomy

**MARK J. GIERL**
**University of Illinois at Urbana–Champaign**

ABSTRACT **An examination was conducted to determine whether the Taxonomy of Educational Objectives: Cognitive Domain (Bloom, Englehart, Furst, Hill, & Krathwohl, 1956) provided an accurate model to guide item writers for anticipating the cognitive processes used by students on a large-scale achievement test in mathematics. Thirty Grade 7 students were asked to think aloud as they solved problems on a mathematics achievement test. Students' cognitive processes were classified with a coding system based on Bloom's taxonomy. The overall match between the responses expected by the item writers and the responses observed from the students was 53.7%. The match score between the expected and the observed responses differed for the high and low mathematics achievers and also differed across the 2 content areas measured on the test. Agreements between the expected and the observed responses were further assessed by comparing log-linear models. The most parsimonious model contained an achievement group, cognitive level, and content area main effect, and, most important, a cognitive level by content area interaction. This finding indicated that the 2 dimensions assumed to be independent in the table of specifications, cognitive level and content area, were, in fact, dependent. The results of this study suggest that Bloom's taxonomy does not provide an accurate model for guiding item writers to anticipate the cognitive processes used by students. Implications for test design are discussed.**

The objectives of schooling are numerous, and many of these objectives include changes in students' cognitive skills. But assessing cognition with achievement tests is difficult. Test developers try to overcome this difficulty by considering the curricular, cognitive, and predictive features of the achievement test (Millman & Greene, 1989). Often, however, the emphasis during test construction is on curricular features such as content coverage (Emmerich, 1989) and on predictive features such as student classification (Embretson, 1985). Cognitive features, such as strategy selection and higher order thinking, are often poorly evaluated because item writers are not trained to identify the cognitive processes required to solve test items. In most cases, item writers are content specialists working from test specifications that have no formal relation to contemporary psychological theory (Embretson, 1985; Snow & Peterson, 1985).

Currently, the most widely used model for identifying the cognitive processes used by examinees to solve test items is the Taxonomy of Educational Objectives: Cognitive Domain (Bloom, Englehart, Furst, Hill, & Krathwohl, 1956). The taxonomy provides a systematic outline of six different levels of thinking that were proposed by Bloom et al. as goals of classroom instruction. The taxonomy begins with the simplest level, knowledge (i.e., recall of specific information), and ends with the most complex level, evaluation (i.e., the ability to judge the value of materials and methods for given purposes). The impact of Bloom's taxonomy in test design is most apparent in the table of specifications. The table contains an outline of the achievement domain and provides a guideline for obtaining a representative sample of test items. The information in the table may also be used to interpret scores for a cluster of items with a common cognitive level that may help the user evaluate differential test performance (i.e., how well students or groups of students perform in relation to different cognitive levels in the taxonomy). Although the structure of the table can vary, the most common procedure is to create a two-way matrix in which one dimension of the matrix specifies the content coverage and the second dimension specifies the cognitive objectives (Ebel & Frisbie, 1986; Gronlund, 1991; Osterlind, 1989; Smith, 1984). Items for each cell in the table of specifications are created by writers who try to anticipate the cognitive processes that examinees will use to answer the questions correctly (Millman & Greene, 1989). Consequently, the table of specification provides a test developer's representation of cognition. However, Bloom's taxonomy is a model of cognitive *intentions*, and it may not be an accurate description of the cognitive processes that students use when solving achievement test items.

*Address correspondence to Mark J. Gierl, Alberta Education, Devonian Building, West Tower, 11160 Jasper Avenue, Edmonton, Alberta, Canada, T5K OL2.*

My purpose in this study was to determine whether Bloom's taxonomy provides item writers with an accurate model for anticipating the cognitive processes used by elementary school students to solve items in a large-scale achievement test in mathematics. Three issues were addressed. First, items for each cognitive level in the table of specifications were developed by item writers who inferred the cognitive processes required by the examinee to answer the questions correctly. The validity of this technique was assessed by addressing the question: Do students use the cognitive processes identified by item writers in the same proportions as outlined in the table of specifications? For example, if three items designed to measure knowledge processes are solved by 30 students, would 90 knowledge responses be identified? Second, the content areas and cognitive levels were represented as mutually exclusive variables. Therefore, I assumed that each item was associated with only one content area and one cognitive level (i.e., an item can be in only one cell in the two-way table of specifications matrix). The validity of this assumption was evaluated by addressing the question: Are content areas and cognitive levels independent dimensions? Third, examinees were not differentiated in the table of specifications. Consequently, high and low mathematics achievers were assumed to use the same cognitive processes to solve test items. This assumption was examined by addressing the question: Do high and low mathematics achievers use the cognitive processes identified by item writers to solve test items? I examined classification differences between Bloom's cognitive levels and student protocol reports across two achievement levels in mathematics and across two content areas to address these questions.

## Method

### *Participants*

The sample contained 30 Grade 7 students (16 boys, 14 girls) in a Catholic school system. Median age (in years and months) was 12 years 7 months (range, 12 years 1 month—14 years 0 months). Students were chosen from three Grade 7 mathematics classrooms in which the teachers agreed to help with the study. Of the 82 permission forms distributed in the three classes, 43 were returned (52.4%). Thirty-seven students agreed to participate and 6 declined. All students who took part in the study had parental consent. Seven students who agreed to participate were not tested because of absenteeism, prior commitments (e.g., class exams), or school events (e.g., dance, special events day).

### *Materials*

*Mathematics Achievement Subtest (MAS).* This subtest is an 18-item multiple-choice examination. The dimensions in the table of specifications and the test items for the MAS were selected from previous achievement tests in mathematics used in a large provincial testing program in Canada. In 1991, over 30,000 students took the provincial mathematics achievement test at the end of Grade 6. Three cognitive levels across two content areas were measured with the MAS. The MAS was used to assess knowledge, comprehension, and application—the same cognitive skills measured with the 1991 provincial mathematics achievement test. The definitions used by the provincial test developers to describe these cognitive skills closely resembled the first three levels in Bloom's taxonomy. Consequently, Bloom's taxonomy was used to code students' think-aloud protocols.

*Knowledge* was defined by the test developers as recognizing or recalling mathematics facts, definitions, rules, procedures, and performing routine mathematics manipulations. Knowledge was defined by Bloom et al. (1956) as the "remembering, either by recognition or recall, of ideas, material, or phenomena" (p. 62). Because both definitions focused on recall and recognition, knowledge was operationally defined as recalling the mathematics solution.

The test developers defined *comprehension* as understanding mathematical principles and concepts and being able to demonstrate this understanding. Translating information into different representations, such as from numbers to words, was also included in this definition. Comprehension was defined by Bloom et al. (1956) as an "understanding of the literal message contained in a communication" (p. 89) that could be demonstrated by manipulating, interpreting, and explaining concepts and ideas. Both definitions required students to demonstrate their understanding of a concept. As a result, comprehension was operationally defined as performing the mathematics operation required in the question and generating a solution.

The test developers defined *application* as solving mathematical problems by using previously learned skills and knowledge. Application was defined by Bloom et al. (1956) as the ability to "apply the appropriate abstraction without having to be prompted as to which abstraction is correct or without having to be shown how to use it in that situation" (p. 120). In other words, application requires the use of previously learned materials in new situations. Both definitions emphasized using previously learned skills to solve problems. Therefore, application was operationally defined as performing the mathematics operations required in the question, generating an intermediate solution, and then applying the intermediate solution to reach the final answer.

The content areas on the MAS were *numeration* and *operations and properties*. These two areas represented 52.7% of the content coverage on the 1991 provincial mathematics achievement test. Numeration included concepts such as recognizing and manipulating mathematical patterns, place values, numbers (whole, decimal, fractions), and numerical relationships (comparing, ordering, rounding). Operations and properties included concepts such as applying number properties (commutative, associative, distributive) as well as adding, subtracting, multiplying, and dividing whole numbers and decimals.

Combining the three cognitive levels with the two content areas produced the table of specifications used for the MAS in this study. The table contained equal numbers of items in each cell, with a total of six items for each of the three cognitive levels and nine items for both content areas. To ensure that the MAS items would not be confounded by either the mathematics concepts within each content area or the item difficulty (based on item analysis data from the 1991 administration of the provincial achievement test in mathematics), I selected items in each cell so that they measured different mathematics concepts and had a range of item difficulties (varying from .35 to .65).

Items for the MAS were ordered unsystematically with the constraints that the three cognitive levels and the two content areas were divided evenly among the first half and second half of the test. The order of the items in the first form was reversed to create a second form. Each form was administered to 15 students.

*Procedure*

Students within each of the three classes were rank ordered into two achievement categories according to teacher-assigned mathematics grades used for the first report card. Of the 30 students who agreed to participate, 15 students (9 boys, 6 girls) from the top half and 15 students (7 boys, 8 girls) from the bottom half of each class were tested. Across the three classes, the mean mathematics grades for the low scorers in the high-achievement group, computed from teacher-assigned unit mathematics examinations, were 9% to 14% greater than the mean mathematics grades for high scorers in the low-achievement group.

Participants were individually tested in an empty classroom 1 month into the school year. Students were asked to think aloud as they solved each test item and to say all of the thoughts and strategies that came to mind as they formulated their solution. After the students selected one of four possible multiple-choice options, I asked them to explain why they chose that option (Ericsson & Simon, 1993). All responses were tape recorded. Three practice items were completed prior to beginning the MAS. Each session typically lasted 20 min.

## Results

*Response and Rater Consistency*

Two measures of consistency were calculated: Cronbach's alpha coefficient and the percentage of interrater agreement. The alpha coefficient for scores on the MAS was .81, indicating a reasonably high degree of internal consistency for this sample of students. To assess interrater agreement for the coded cognitive responses, a second rater trained to make consistent judgments coded the think-aloud protocols of 5 randomly selected students (16.7% of the sample). Of the 90 responses coded, 75 agreements occurred (83.3%), indicating that the students' cognitive processes were consistently coded. Disagreements were evenly distributed across items in the three cognitive levels. Five disagreements occurred with knowledge items, 6 with comprehension items, and 4 with application items.

*Response Frequencies*

Of the 468 cognitive responses reported by the students,[1] 251 (53.6%) matched the cognitive levels anticipated by the item writers. When the response frequencies were examined, the expected responses for the three cognitive levels differed from the observed values, indicating that students used knowledge, comprehension, and application processes to solve items on the MAS, but not in the same proportions as identified by item writers (see summary matrix in the bottom right corner of Table 1). When cognitive level was partitioned across the high and low mathematics achievers, the proportion of cognitive responses expected by the item writers and observed from students differed for the high mathematics achievers: only 139 out of 248 (56%) responses matched. Similarly, the expected and observed responses differed for the low mathematics achievers; only 112 out of 220 (50.9%) responses matched. Across the two content areas, a comparable result occurred as the cognitive responses expected by item writers failed to match the cognitive responses observed from students in numeration, with only 126 out of 247 (51%) response matches; and in operations and properties, with only 125 out of 221 (56.6%) response matches. These results suggested that students, in general, used the processes described in Bloom's taxonomy to solve MAS items, but not in the same proportions as identified by item writers when summarized overall or when summarized as a function of the two achievement groups in mathematics or the two content areas.

*Factors Influencing Response Ratings*

To evaluate the assumption that the table of specification contained independent dimensions, I conducted a log-linear analysis on the cell frequencies in Table 1, using match scores as the dependent variable.[2] The goal in creating the log-linear models was to select a combination of parameters that appeared to describe the observed cell frequencies. In this context, I evaluated the likelihood-ratio statistic ($L^2$) to determine if the chosen parameters yielded frequencies that provided an acceptable fit to the contingency table data. Four models were tested. In the first model, the main effects of achievement group, content area, and cognitive level were fitted to the data. Acceptance of this model would indicate that the three factors were mutually independent. In the second model, the three main effects and the interaction between content area and cognitive level were fitted to the data. Acceptance of this model would indicate that content area and cognitive level interacted with one another and, thus, had a dependent relationship. In the third model, the

**Table 1.—Response Frequencies Across Content Area and Achievement Level as a Function of Knowledge, Comprehension, and Application Cognitive Levels**

| | Achievement group K | C | A |
|---|---|---|---|
| *High* | | | |
| Numeration | | | |
| Knowledge | 12 | 30 | 0 |
| Comprehension | 0 | 39 | 0 |
| Application | 0 | 33 | 6 |
| Operations/properties | | | |
| Knowledge | 19 | 14 | 0 |
| Comprehension | 4 | 33 | 0 |
| Application | 5 | 22 | 3 |
| Marginal sums | | | |
| Knowledge | 31 | 44 | 0 |
| Comprehension | 4 | 72 | 0 |
| Application | 5 | 55 | 9 |
| *Low* | | | |
| Numeration | | | |
| Knowledge | 10 | 31 | 1 |
| Comprehension | 0 | 43 | 0 |
| Application | 0 | 25 | 16 |
| Operations/properties | | | |
| Knowledge | 28 | 12 | 0 |
| Comprehension | 5 | 35 | 0 |
| Application | 6 | 28 | 7 |
| Marginal sums | | | |
| Knowledge | 38 | 44 | 1 |
| Comprehension | 5 | 78 | 0 |
| Application | 6 | 53 | 23 |
| *Marginal sum* | | | |
| Numeration | | | |
| Knowledge | 22 | 62 | 1 |
| Comprehension | 0 | 82 | 0 |
| Application | 0 | 58 | 22 |
| Operations/properties | | | |
| Knowledge | 47 | 26 | 0 |
| Comprehension | 9 | 68 | 0 |
| Application | 11 | 50 | 10 |
| Marginal sums | | | |
| Knowledge | 69 | 88 | 1 |
| Comprehension | 9 | 150 | 0 |
| Application | 11 | 108 | 32 |

*Note.* The responses expected by the test developers that matched the responses observed from students are along the main diagonal for each content area by achievement group matrix. The off-diagonals of each matrix represent the mismatches.

main effects and the interaction between achievement group and cognitive level were fitted to the data. Acceptance of this model would indicate that achievement group and cognitive level had a dependent relationship. In the final model, the three main effects and the two interaction terms were fitted to the data. Acceptance of this model would indicate that both interaction terms were needed to describe the data. The fourth model was also expected to produce the best fit because it contained the most parameters. Consequently, it was used to evaluate the other three models.

The results of the log-linear analysis revealed that Model 2 provided the best fit to the data (see Table 2). Model 2 contained an achievement group, content area, and cognitive level main effect, and a content area by cognitive level interaction, $L^2(5, N = 30) = 5.62$, $p = .35$, which implied that content area and cognitive level were *not* independent dimensions, as specified in the table of specification. Rather, content area and cognitive level formed a conditional relationship with one another when match score was used as the dependent variable. Models 2 and 4 were not statistically significant from one another: That finding also indicated that Model 2 adequately described the data because it produced a similar result when compared with a more complex model (Fienberg, 1980; Kennedy & Tam, 1994). The results of the log-linear analysis were used to further investigate the misclassified responses in Table 1.

The diagonal elements in Table 1 contain matches between the cognitive responses expected by the test developers and observed from students. The off-diagonals contain mismatches. When the off-diagonal elements in Table 1 were summed, two large elements that represented discrepancies between the responses expected by the test developers and the responses observed from students were noted—88 responses were expected to be knowledge but were observed as comprehension, and 108 responses were expected to be application but were observed as comprehension. To understand the misclassified responses, I used chi-square tests to evaluate the cell frequencies from the content area by cognitive-level interaction, as suggested from the results of the log-linear analysis. For the 88 responses that the test developers expected to be classified as knowledge, but were observed as comprehension, there was a difference between the content areas (the expected cell frequencies for the marginal sums of the rows in Table 1 are 44, 44 [obtained from 88/2] versus

**Table 2.—Goodness-of-Fit Associated With Four Log-Linear Models**

| Model | Parameters | $L^2$ | *df* | *p* |
|---|---|---|---|---|
| 1 | A, B, C, | 20.81 | 7 | .00 |
| 2 | A, B, C, BC | 5.62 | 5 | .35 |
| 3 | A, B, C, AC | 16.43 | 5 | .01 |
| 4 | A, B, C, BC, AC | 1.24 | 3 | .74 |
| Model comparisons | | $c^2$ | *df* | |
| Model 1 versus Model 2 | | 15.19* | 2 | |
| Model 2 versus Model 4 | | 4.38 | 2 | |

*Note.* A is the achievement group, B is the content area, and C is the cognitive level.
*$p < .05$.

the observed cell frequencies of 62, 26, $\chi^2[1, N = 30] = 14.73, p < .01$), as the cognitive processes required to solve the numeration items were misclassified more frequently than those for the operations and properties items. For the 108 responses that the test developers expected to be classified as application but were observed as comprehension, there was no difference between the two content areas (the expected cell frequencies for the marginal sums of the rows in Table 1 are 54, 54 [obtained from 108/2], vs. the observed cell frequencies of 58, 50 $\chi^2[1, N = 30] = .59, p = .44$). Differences between the high and low achievers were not statistically significant; both groups contributed equally to the 88 and the 108 misclassified responses.

The general findings that emerged from these analyses were as follows: The table of specifications treated content area and cognitive-level as mutually exclusive (i.e., one item per cognitive level and content area). However, when log-linear models were fitted to the data, a model containing a content area by cognitive-level interaction produced the most parsimonious and interpretable result. This finding indicated that items were solved with a variety of the cognitive processes listed in the table of specifications and that content area and cognitive level influenced item classification. In addition, comprehension had the highest mean match score, suggesting that it was the cognitive process most easily anticipated by item writers. Bloom et al. (1956) foreshadowed this finding when they speculated that comprehension processes were "the largest general class of intellectual abilities and skills emphasized in school" (p. 89). Conversely, knowledge and application processes were poorly anticipated by the item writers. Finally, the two achievement groups tended to use similar cognitive processes to solve test items. However, when differences occurred, the high achievers had more matches between the expected and observed responses than the low achievers did. This finding indicated that item writers were more accurate at anticipating the cognitive processes used by high mathematics achievers than by low mathematics achievers.

### *Student Protocols*

I also evaluated student protocols to understand why several items expected to elicit knowledge processes were solved with comprehension processes. A similar analysis not reported here was conducted with items expected to elicit application processes but solved with comprehension skills. Of the six knowledge items in the MAS, the items in the Appendix contained 70 of the 88 misclassified responses.

Both the provincial test developers and Bloom et al. (1956) differentiated knowledge and comprehension. Knowledge was defined as remembering, either by recognition or recall, mathematical ideas and materials. For this study, knowledge was operationally defined as recalling the mathematical solution. Comprehension was described as understanding mathematical principles and concepts and being able to demonstrate this understanding. Comprehension was operationally defined as performing the mathematical operation required in the question and generating a solution.

For Item 5 (see Appendix), there was little response variability; 29 students solved the problem directly. The numbers in parentheses were multiplied together, and the products added to generate a solution. The salient cognitive processes were multiplication and addition, and the calculations were often performed in the test booklets. This strategy demonstrated an understanding of standard notation and goes beyond recognizing or recalling a solution.

Item 7 elicited a variety of strategies. Nine students counted the total number of pieces in the pie and then counted the number of shaded pieces and formed a fraction of shaded to unshaded pieces. The students, using the lowest common denominator, then reduced the fraction and selected the appropriate solution. The calculations were often performed in the test booklet. A second strategy, used by 5 students, was to create the fraction by counting, but to explain that if two pieces of pie were counted as one the correct solution would be found. No student recalled the solution directly.

Item 15 produced the most response variability; seven strategies were identified. The most common approach, used by 3 students, was to identify the counting ratio, locate some marker points, and count. One examinee explained, "You can count in 2s. 50% is 10, 100% is 20, so 90% is 18." Two students recalled 90% was .9, then computed in the test booklet that $20 \times 0.9 = 18$. Another strategy, used by 2 students, was to present the problem as $90/100 = x/20$ and identify a common divisor through trial and error. Five was often used and applied to both the numerator and denominator so that $100/5 = 20$, and $90/5 = 18$. Knowledge was another strategy used to solve Item 15; 3 students explained that "18 out of 20 is 90% because I've got 90% on tests before." The knowledge response clearly demonstrated recalling a solution. In contrast, the comprehension responses involved calculating the solution.

An examination of student protocols on three misclassified items yielded two findings: Response variability was prevalent as students used different strategies to solve an item, and many knowledge items were solved with processes and strategies that corresponded to comprehension, as defined in this study.

## Discussion

The results of this study indicate that the cognitive domain in the Taxonomy of Educational Objectives (Bloom et al., 1956) does *not* provide an accurate model to guide item writers for anticipating the cognitive processes used by students to solve items on an achievement test in mathematics. The model failed in four important ways.

1. The cognitive processes expected by item writers matched the processes used by students in only 54% of the cases outlined in the table of specifications. This finding

demonstrated that Bloom's taxonomy enabled item writers to correctly anticipate the cognitive processes used by students about half of the time.

2. The table of specifications treated the content areas and cognitive levels as mutually exclusive. Yet when the match between expected and observed responses was examined, it was shown that the content areas and cognitive levels interacted. This finding indicated that the dimensions in the table of specifications were not mutually exclusive, and that content area and cognitive level had a dependent relationship with one another.

3. Item writers were able to anticipate the processes used by high mathematics achievers more readily than by low mathematics achievers. Differences between the two groups were not pronounced, but when differences occurred they favored the high mathematics achievers. Consequently, the cognition section in the table of specifications provided a more accurate guide of the mental processes used by high-achieving mathematics students.

4. Within each level of the taxonomy there was response variability. For example, Item 15 (see Appendix) on the MAS was solved with seven different strategies. This result demonstrated that the levels in Bloom's taxonomy, as used in test construction, concealed response variability, and that much of the cognitive complexity was lost by coding responses under general categories such as knowledge, comprehension, or application. Furthermore, when think-aloud protocols were used to evaluate students' cognitive processes, some items that were expected to be solved with knowledge processes were solved instead with comprehension processes.

If, by using Bloom's taxonomy, item writers are unable to accurately anticipate students' cognitive processes, what can be done to improve this aspect of test design? The following suggestions are provided. The cognition section in the table of specifications should not be used for test interpretation because it does not accurately identify the cognitive processes used by students to solve test items. Moreover, there is a need to adopt the concepts and methods of cognitive psychology in test design. Fortunately, cognitive psychology is making tremendous inroads into the field of educational measurement. This progression seems inevitable because most educational tests are based on cognitive problem-solving tasks. Psychometricians working with cognitively diagnostic assessments are currently leading this movement to integrate the principles of cognitive psychology into educational measurement (e.g., Frederiksen, Glaser, Lesgold, & Shafto, 1990; Frederiksen, Mislevy, & Bejar, 1993; Nichols, 1994; Nichols, Chipman, & Brennan, 1995; Snow & Lohman, 1989). The work of Tatsuoka (1993, 1995) and her associates (Birenbaum, Kelly, & Tatsuoka, 1991) with the rule-space model is one example of how data from psychometric modeling can be used to make inferences about students' cognitive processes when they use their item response patterns. Finally, student protocol data should be collected and analyzed during the pilot or item field-testing phase of test development. When Bloom's taxonomy was developed, the editorial staff collected a large list of educational objectives, identified the intended behaviors, and created groups of similar behaviors; six cognitive levels were identified. During the development of the standardized achievement test in this study, items for each cognitive level in the table of specifications were created by writers who tried to anticipate the cognitive processes that examinees would use to answer the items correctly. In both cases, the cognitive processes of students were inferred rather than measured directly. If test developers hope to assess students' cognitive processes successfully, researchers should use think-aloud protocols to evaluate directly the problem-solving strategies that *students* use to solve achievement test items. This approach would provide test developers with a better understanding of how students solve items in the achievement domain.

## NOTES

This research was supported with funds from the Social Sciences and Humanities Research Council of Canada. I would like to thank Tom Maguire for his comments throughout this project. I would also like to thank an anonymous reviewer for his or her patience and thoughtful suggestions.

1. Students whose responses had no cognitive justification (e.g., guess) were excluded from the analyses.
2. The reader should note that the assumption of response independence was likely violated to some degree in this analysis because the achievement group and content area variables contained the same students. As a result, the test statistic may be positively biased and the results from the log-linear analysis must therefore be interpreted with caution (Kennedy & Tam, 1994). Despite the potential violation of the independence assumption, log-linear models were fit to these data because the dependent variable was a categorical response and alternative procedures, such as analysis of variance, were judged to be more problematic.

## REFERENCES

Birenbaum, M., Kelly, A. E., & Tatsuoka, K. K. (1991). Diagnosing knowledge states in algebra using the rule-space model. *Journal for Research in Mathematics Education, 24*, 442–459.

Bloom, B. S., Englehart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). *A taxonomy of educational objectives: Handbook I Cognitive Domain.* New York: David McKay.

Ebel, R. L., & Frisbie, D. A. (1986). *Essentials of educational measurement* (4th. ed.). Englewood Cliffs, NJ: Prentice-Hall.

Embretson, S. E. (1985). Introduction for the problem of test design. In S. E. Embretson (Ed.), *Test design: Developments in psychology and psychometrics* (pp. 3–17). New York: Academic Press.

Emmerich, W. (1989). *Appraising the cognitive features of subject tests* (Research Rep. No. RR-89-53). Princeton, NJ: Educational Testing Service.

Ericcson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data* (Rev. ed.). Cambridge, MA: MIT Press.

Fienberg, S. E. (1980). *The analysis of cross-classified categorical data* (2nd ed.). Cambridge, MA: MIT Press.

Frederiksen, N., Glaser, R. L., Lesgold, A. M., & Shafto, M. G. (1990). *Diagnostic monitoring of skills and knowledge acquisition.* Hillsdale, NJ: Erlbaum.

Frederiksen, N., Mislevy, R. J., Bejar, I. I. (1993). *Test theory for a new generation of tests.* Hillsdale, NJ: Erlbaum.

Gronlund, N. E. (1991). *How to write and use educational objectives* (4th ed.). New York: Macmillan.

Kennedy, J. J., & Tam, E. P. (1994). Log-linear models. In T. Husen & T. N. Postlethwaite (Eds.), *International Encyclopedia of Education* (2nd ed., pp. 3499–3508). New York: Pergamon.

Millman, J., & Greene, J. (1989). The specification and development of tests of achievement and ability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 335–366). New York: American Council of Education: Macmillan.
Nichols, P. (1994). A framework of developing cognitively diagnostic assessments. *Review of Educational Research, 64*, 575–603.
Nichols, P. D., Chipman, S. F., Brennan, R. L. (1995). *Cognitively diagnostic assessment.* Hillsdale, NJ: Erlbaum.
Osterlind, S. J. (1989). *Constructing test items.* Boston, MA: Kluwer.
Smith, F. M. (1984). *Constructing and using achievement tests in the classroom: A competency based text.* New York: Peter Lang.
Snow, R. E., & Lohman, D. F. (1989). Implications of cognitive psychology for educational measurement. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 263–331). New York: American Council of Education: Macmillan.
Snow, R. E., & Peterson, P. L. (1985). Cognitive analyses of tests: Implications for redesign. In S. E. Embretson (Ed.), *Test design: Developments in psychology and psychometrics* (pp. 149–166). New York: Academic Press.
Tatsuoka, K. K. (1993). Item construction and psychometric models appropriate for constructed responses. In R. E. Bennett & W. C. Ward (Eds.), *Construction versus choice in cognitive measurement* (pp. 107–134). Hillsdale, NJ: Erlbaum.
Tatsuoka, K. K. (1995). Architecture of knowledge structures and cognitive diagnosis: A statistical pattern recognition and classification approach. In P. D. Nichols, S. F. Chipman, & R. L. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 327–359). Hillsdale, NJ: Erlbaum.

## APPENDIX
### Student Protocols in the Mathematics Achievement Test

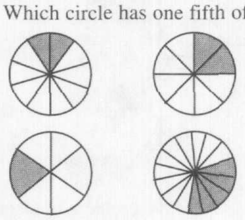
**The Three Knowledge Items Most Frequently Solved With Comprehension Strategies on the Mathematics Achievement Subtest**

Item 5. $(5 \times 1{,}000) + (2 \times 10) + (1 \times 1) + (9 \times 0.1) + (6 \times 0.01)$ written in standard notation is

a. 502.196
b. 521.096
c. 5,021.96*
d. 5,219.6

Item 7. Which circle has one fifth of its area subshaded?

A.* B.



C. D.

Item 15. 90% of 20 is

a. 9
b. 17
c. 18*
d. 19

*Correct solution.