Construct Validation of Mathematics Achievement: Evidence from Interview Procedures

CSE Technical Report 493

Haggai Kupermintz, Vi-Nhuan Le, and Richard E. Snow
CRESST/Stanford University

January 1999

Center for the Study of Evaluation
National Center for Research on Evaluation,
Standards, and Student Testing
Graduate School of Education & Information Studies
University of California, Los Angeles
Los Angeles, CA 90095-1522
(310) 206-1532

Project 3.1 Construct Validity: Understanding Cognitive Processes—Psychometric and Cognitive Modeling Richard J. Shavelson, Project Director, CRESST/Stanford University

The work reported herein was supported under the Educational Research and Development Centers Program, PR/Award Number R305B60002, as administered by the Office of Educational Research and Improvement, U.S. Department of Education. This material is also based upon work supported by the National Science Foundation under Grant No. REC- 9628293.

Any opinions, findings, and conclusions or recommendations expressed in this report are those of the authors and do not reflect the positions or policies of the National Institute on Student Achievement, Curriculum, and Assessment, the Office of Educational Research and Improvement, the U.S. Department of Education, or the National Science Foundation.

CONSTRUCT VALIDATION OF MATHEMATICS ACHIEVEMENT: EVIDENCE FROM INTERVIEW PROCEDURES[1]

Haggai Kupermintz, Vi-Nhuan Le , and Richard E. Snow[2]
CRESST/Stanford University

Abstract

This study investigated the validity of measures derived from a large-scale multiple-choice achievement test in mathematics, using evidence from introspective think-aloud protocols of students as they attempted test items. In a small-scale study of 21 local high school students, we sought to identify and describe cognitive processes underlying performance on test items, and to examine their utility in supporting validity claims about the achievement dimensions tapped by the test. We examined differences and similarities in solution strategies and sources of knowledge used to solve items representing five achievement dimensions. The results provided further evidence for the plausibility of interpretations of the dimensions derived from a large-scale factor analysis and support the conclusion that 12th-grade mathematics achievement is factorially and cognitively complex. Test scores that do not capture such complexity may mask important achievement information.

The study reported here examined the validity of measures derived from a large-scale multiple-choice achievement test in mathematics. Messick (1989) defines validity as "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment" (p. 13). It follows that an important goal of construct validation of test scores should be to identify the forms of reasoning processes involved and the extent to which they can account for test performance. In order to claim validity of test scores for a particular measurement purpose, one must consider such evidence in an effort to refute rival hypotheses concerning what these scores actually measure.

[1] We wish to thank Yuko Butler for assistance with the development of the coding system and coding student transcripts, Larry Gallagher for assistance with interviewing, and Rich Shavelson for helpful comments on an earlier draft of this report.

[2] Richard Snow, the original project director for this research, passed away in December 1997.

The investigation of construct-relevant and irrelevant sources of variance is especially important for test items that may require students to apply a variety of knowledge and skills because it is often difficult to infer from a low score where a deficiency lies (Messick, 1995). Furthermore, the complex interaction between task demands and student characteristics affords different students the opportunity to use different strategies in response to the same task. This necessarily complicates validity arguments in favor of molar constructs such as "level of mathematical ability reached" as the only interpretation one can attach to total scores from conventional achievement tests. Kupermintz & Snow (1997) have recently demonstrated the utility of using more refined distinctions in measuring mathematics achievement in a large-scale multiple-choice test.

As suggested above, efforts to support score interpretations need to take into account evidence pertaining to the nature of the cognitive activities students engaged in while taking the test. However, despite recent calls for considering the cognitive psychology of performance in a domain alongside the psychometrics of achievement assessment (see, e.g., Frederiksen, Glaser, Lesgold, & Shafto, 1990; Frederiksen, Mislevy, & Bejar, 1993; Snow & Lohman, 1989), there has been as yet little construct validation research that combines the psychometric and cognitive approaches. Both multiple-choice and constructed-response items are routinely submitted to rigorous psychometric evaluation, and, increasingly, cognitive psychologists are studying learning and problem-solving tasks that are closely related to tasks used in educational tests (Snow & Lohman, 1989). But integrated, or even concomitant, psychometric and cognitive psychological analysis of task performance remains rare.

Several methods have been used for process analysis to advance the investigation of cognitive aspects of test performance. This report follows Messick's (1989) recommendations for analyses of think-aloud protocols, retrospective reasons for answering in a certain way, and errors made by examinees. Strong relationships in a large-scale psychometric analysis may suggest that certain items require a certain type of knowledge or reasoning, but evidence that students actually use that knowledge or process when they solve an item can best be obtained by observing students working on the items and asking them to introspect about their performance. Think-aloud protocols collected while students perform a task have been used in several recent test validation studies. For example, Hamilton, Nussbaum, and Snow (1997) have identified differential patterns of solution

strategies in multiple-choice items representing three science achievement dimensions (see also Magone, Cai, Silver, & Wang, 1994).

The current study extends the psychometric analysis presented by Kupermintz and Snow (1997) that extracted five dimensions in a 12th-grade multiple-choice test in mathematics. A small-scale study was conducted to obtain evidence from student think-aloud protocols to identify and describe cognitive processes underlying performance on test items, and to examine their utility in supporting validity claims about the achievement dimensions tapped by the test.

Method

The study pursued a cognitive analysis of multiple-choice test items administered to a national representative sample of high school students. The National Educational Longitudinal Study of 1988 (NELS: 88), sponsored by the National Center for Education Statistics (NCES), is the most recent in a series of large-scale surveys designed to provide longitudinal data about critical educational experiences of students as they leave elementary school and progress through high school and into postsecondary tracks. NELS:88 followed a national probability sample of about 25,000 8th graders into the 10th and 12th grades using a series of cognitive tests as well as questionnaires completed by students, parents, teachers, and school administrators.

In a recent validation study, the NELS:88 multiple-choice test in mathematics was factor analyzed to reveal several interpretable achievement dimensions (Kupermintz & Snow, 1997). On Form H—a 40-item test that was assigned to high-ability 12th-grade students—five dimensions emerged from a full information factor analysis. The original factor interpretations were as follows: *Compound Mathematical Reasoning (CMR)* involves complex tasks with advanced content that require students to set up abstract representations (algebraic expressions or equations), and to consider multiple perspectives of the problem; *Concrete Mathematical Reasoning (NMR)* requires mainly inferential reasoning, with a demand for logical argument rather than direct computations; *Applied Algebra Knowledge (AAK)* consists mainly of problems introducing algebraic expressions, using variables and functional forms in which students are required to apply algebra knowledge in a straightforward computation or manipulation, such as numerically solving an equation or simplifying an algebraic expression; *Spatial Visualization (SV)* taps visual-spatial ability by items that call for operations such as mental rotation and folding of two-

dimensional shapes; *Algebra Systems Comprehension (ASC)* calls for conceptual understanding of algebraic expressions where no direct calculation is required but understanding of the algebraic system is essential.

Kupermintz and Snow (1997) further hypothesized that such complex factorial structure arose because advanced problems demanded application of various specialized knowledge and problem-solving skills and strategies, allowing students to exhibit sophisticated mixes of complex abilities. In a series of regression analyses, scores on these dimensions were shown to have differential patterns of relationships with student, program, and instructional variables that were not captured by using only total scores.

Small-Scale Interview Study

The current study used interview and think-aloud procedures to gain evidence concerning the processes underlying task performance on items representing each of the hypothesized five mathematics achievement dimensions. Twenty-one (8 males and 13 females) local high school students (10 seniors, 9 juniors, and 2 sophomores) participated in the study. All of the students had taken trigonometry, geometry, and algebra I and II courses; 12 students were enrolled in calculus courses. Students were asked to verbalize their thought processes and to identify sources of knowledge they utilized as they attempted to solve 15 items from the original NELS:88 test.

All interviews were conducted at Stanford University by the investigators. After a brief introduction to the study, interviewers asked students to think aloud while completing test items and did not intervene except to remind students to think aloud if a specified period of silence passed (prompting with "Can you tell me what you're thinking about now?"). Students were encouraged to use the test booklets for scratch work. After students completed each item, they responded to two interview questions: "Can you tell me how you decided which answer to select?" and "Where have you learned to solve such problems?" The combination of spontaneous think-aloud protocols and structured interview prompts was designed to allow students to respond to items without intervention and at the same time to obtain information not volunteered in the unstructured think-aloud format.

Interviews were audiotaped and transcribed, and interviewers used a structured observation sheet to record events that could not be captured on

audiotape, such as the use of gestures. All written notes made by interviewers were added to the session transcripts.

Test items were taken from the NELS:88 mathematics test and were selected from the pool of 40 Form H items to represent the five achievement dimensions. Items showing strong loadings in the factor analysis were considered good candidates for inclusion in the small-scale study. Additional criteria for inclusion were the need for a range of item difficulties and the likelihood that they would elicit rich verbalization. Table 1 describes the items representing each dimension, together with proportion of correct responses (*p*-value) on each item and, over items, on each dimension. Because of test security concerns only brief descriptions and not actual items are presented.

Table 1

Proportion of Correct Responses

Dimension/Item		*p*-value
Compound mathematical reasoning (CMR)	Total	0.24
Radius of a cylinder within box		0.29
Algebraic expression for relations of triangle sides		0.14
Overall average based on group averages		0.29
Concrete mathematical reasoning (NMR)	Total	0.89
Length of side of figure given area		0.90
Word problem involving area and dimensions		0.81
Distance between points on a line		0.95
Applied algebra knowledge (AAK)	Total	0.89
Equation involving function notation		1.00
X-intersection points of a function		0.71
Equation with function notation and exponents		0.95
Spatial visualization (SV)	Total	0.60
Match pattern of unfolded box		0.81
Shape of rotated line		0.57
Length of side-parallel within a triangle		0.43
Algebra system comprehension (ASC)	Total	0.76
RHS multiplication in a function		0.76
Multiplicative functional form		0.81
Functional form of a graph		0.71

Coding System

A coding scheme was developed to reflect the entire range of student responses (see Table 2), and two raters simultaneously coded all the items, using the same coding system and reaching agreement through discussion. Each item was coded for all the solution strategies and sources of knowledge that were indicated in the students' work. Sources of knowledge included courses in general math, algebra, algebra II, geometry, trigonometry, calculus, and SAT preparation.

Table 2

Solution Strategies

Strategy	Description
Trial and error	This approach involved working backwards from the multiple-choice options, substituting each of the response alternatives into an equation usually specified in the item stem.
Manipulation of algebraic expressions	Involved simplifying or other manipulations on an algebraic expression or equation.
Application of advanced algebraic properties	Necessitated consideration of advanced concepts and procedures, for example conceptual understanding of a given equation and its graphical representation on the Cartesian coordinate system.
Application of basic algebraic properties	Involved attention to basic concepts and procedures. For example, setting up a single equation or calculating the distance between points on a line were classified in this category.
Application of geometric properties	Involved geometry concepts, such as the Pythagorean theorem and similar triangles.
Application of trigonometric properties	Involved trigonometry concepts, such as sine and cosine expressions.
Abstract/analytic approach	Characterized by abstract reasoning without assigning numerical values or using direct calculations.
One-step formula	Involved a single application of a formula, such as the area of a rectangle or a linear equation.
Multiple-step formula	Characterized by the application of more than one formula. A response involved multiple operations; for example, using both volume and area formulas.
Substitution	This strategy involved the replacement of a variable with one numerical value that was specified in the item stem. A solution was subsequently determined via direct computation.
Value assignment	Involved investigating the association underlying two variables via the assignment of numerical values to the variables. This strategy was used not to obtain a direct solution per se, but to allow generalizations about the underlying relationship between the two variables.

Table 2 (continued)

Strategy	Description
Elimination of responses	Students who employed this strategy typically had some knowledge of the item, which they used to evaluate the merit of each option, rejected certain options, and subsequently chose an answer from the remaining alternatives.
Guess	Characterized by no knowledge of the item. This strategy was used by students who did not know how to approach an item. They often justified their selection via intuition or statements such as "it just sounded right."
Visualization	Involved using mental representation to visualize hypothetical shapes referred to in the item.
Gestures	In this strategy students used hand movements to facilitate their visualization efforts.
Picture/graph	Involved drawing graphical representations of information specified in the item.
No Answer	This category was assigned when students chose not to select a multiple-choice alternative.

Results

In this section we present findings on the relative prominence of the various solution strategies and sources of knowledge observed across the five achievement dimensions. Table entries are the proportion of responses in a particular category across the three items representing each dimension given by the 21 students (that is, the proportion from the 3x21 = 63 responses representing each dimension). Results are presented for total responses (across correct and incorrect answers), and separately for correct and incorrect solutions. For example, 15 correct answers (15/63 = .24) were given for CMR items. Overall, 8 out of 63 responses to CMR items employed a multiple-step formula (15/63=.13; the entry in the "Total" column). Out of the 15 correct answers, 6 used multiple-step formula (6/15 = .40; the entry in the "Correct" column), whereas only 2 of the incorrect answers used multiple-step formula (2/48 = .04; the entry in the "Incorrect" column). The comparison of distributions of strategies and sources of knowledge in correct versus incorrect solutions allowed us to consider their efficiency for answering the items. To reduce clutter, only categories that were observed in more than 10% of the responses (total, correct, or incorrect) are presented.

Compound Mathematical Reasoning (CMR)

Not surprisingly, CMR items, the most difficult items on Form H, elicited the highest proportion of "guess" or "no answer" responses (see Table 3). In fact, many students expressed uncertainty about how to start CMR problems. Correct solutions were likely to apply compound procedures (multiple-step formula) and advanced algebraic properties, whereas incorrect solutions employed basic algebraic properties.

One item, for example, required the application of both volume and area formulas. Students who answered correctly accurately remembered and used both formulas, whereas unsuccessful students did not recognize the complex nature of the problem and instead tried to find a one-step formula that would yield a direct solution: "I'd probably need the formula for it . . . I'm not sure I ever memorized it." Another item required students to take into account the effect of the different group sizes on the overall average. Unsuccessful students used a basic algebraic approach

Table 3

Solution Strategies and Sources of Knowledge Used in Compound Mathematical Reasoning (CMR) Items

Strategy/Source of knowledge	Total	Correct	Incorrect
Algebraic manipulation	.11	.13	.10
Advanced algebraic properties	.13	.33	.06
Basic algebraic properties	.22	.00	.29
Geometric properties	.22	.20	.23
One-step formula	.08	.00	.10
Multiple-step formula	.13	.40	.04
Value assignment	.06	.13	.04
Guess	.25	.13	.29
Picture/graph	.25	.27	.25
No answer	.22	.00	.29
General math	.24	.33	.21
Algebra	.30	.20	.33
Geometry	.46	.33	.50
SAT/SAT prep	.10	.20	.06

Note. Total: 63 responses; Correct: 15 responses; Incorrect: 48 responses.

that resulted in an unweighted average. In contrast, successful students explicitly indicated the need to take into account the multiple conditions of the problem.

An interesting and somewhat surprising result was that both correct and incorrect solutions were equally likely to use graphical representations and employ geometric properties. Closer inspection revealed that the unsuccessful students were more inclined to draw two-dimensional pictures, whereas successful students' pictures were three-dimensional representations that depicted the problem more accurately.

Geometry, algebra, and general math were the most common knowledge sources, which was consistent with the fact that students were most likely to approach such items by applying both geometric and algebraic properties. Students who answered correctly referred more often to general math while students who answered incorrectly mentioned more frequently algebra and geometry.

These findings are consistent with the hypothesis that CMR task demands called for a complex strategy rather than simple application of domain knowledge and support the original dimension interpretation based on the factor analysis. That is, an efficient solution strategy required attention to the complex nature of the problem, but placed relatively small computational or algorithmic demands in processing its constituent components. It is also interesting to note that successful students were likely to encounter similar problems while preparing for the SAT.

Concrete Mathematical Reasoning (NMR)

Items on the NMR dimension were fairly easy, and most students correctly responded to each question (see Table 4). Application of a one-step formula was the dominant strategy, and correct responses typically used basic algebraic properties. The few conceptual errors, leading to incorrect responses, usually involved procedural knowledge: "I don't really see how I can find the area of the square if I don't even have one side of it." Incorrect solutions were more likely to result from unnecessary graphical representations for items that asked students to calculate areas. Incorrect solutions were also more likely to use an analytic approach.

These findings suggest that our initial interpretation of this dimension (partially based on different test forms taken by low- and medium-ability students) was not supported. Although the low computational demand was correctly identified, this dimension appears to be characterized by a straightforward application of domain knowledge, declarative and procedural, as opposed to more

Table 4

Solution Strategies and Sources of Knowledge Used in Concrete Mathematical Reasoning (NMR) items

Strategy/Source of knowledge	Total	Correct	Incorrect
Basic algebraic properties	.33	.36	.14
Analytic approach	.05	.02	.29
One-step formula	.51	.50	.57
Elimination	.02	.00	.14
Guess	.03	.00	.29
Guess part	.02	.00	.14
Picture/graph	.17	.13	.57
No answer	.02	.00	.14
General math	.29	.32	.00
Algebra	.33	.32	.43
Algebra II	.08	.07	.14
Geometry	.32	.32	.29
SAT/SAT prep	.10	.11	.00
Can't remember	.03	.02	.14

Note. Total: 63 responses; Correct: 56 responses; Incorrect: 7 responses.

general reasoning abilities. To answer NMR items, students reported drawing upon general math, algebra, and geometry almost equally. As in CMR, algebra was mentioned more frequently for incorrect answers. It is also worthwhile noticing here that in the large-scale regression analysis (Kupermintz & Snow, 1997) a strong predictor of performance on NMR was the number of units taken in geometry. It is, therefore, reasonable to describe this dimension as tapping the application of basic concepts and procedures from geometry.

Applied Algebra Knowledge (AAK)

AAK items were also relatively easy, involving substitution and direct computation as dominant strategies, evident in nearly three quarters of the correct answers (see Table 5). Erroneous responses (due mainly to a somewhat more sophisticated item asking students to find the roots of a function) were characterized by unnecessary manipulation of an algebraic expression or an attempt to graph a function in order to determine its X-intersection points. Some students even

Table 5

Solution Strategies and Sources of Knowledge Used in Applied Algebra Knowledge (AAK) Items

Strategy/Source of knowledge	Total	Correct	Incorrect
Trial and error	.06	.04	.29
Algebraic manipulation	.08	.04	.43
Advanced algebraic properties	.25	.27	.14
Substitution	.70	.73	.43
Elimination	.02	.00	.14
Guess	.05	.00	.43
Picture/graph	.03	.00	.29
No answer	.05	.00	.43
General math	.14	.16	.00
Algebra	.41	.38	.71
Algebra II	.37	.39	.14
Geometry	.08	.07	.14
Trigonometry	.27	.27	.29
Calculus	.11	.13	.00

Note. Total: 63 responses; Correct: 56 responses; Incorrect: 7 responses.

expressed a need for a graphing calculator to provide visual representation: "So what I would probably do if I had a [TI82] calculator is plug this equation in and then look and see where y intersects the x axis." Correct responses, on the other hand, identified the relationship between the algebraic expression and the Cartesian coordinate system, arriving at a solution using minimal computation or manipulation. Unsuccessful strategies also employed the less direct approaches of trial and error and elimination instead of using direct computation.

Consistent with the relatively advanced algebraic material, algebra II was often drawn upon in correct solutions, while algebra was mentioned as a source of knowledge in the majority of incorrect solutions. In the large-scale regression analysis (Kupermintz & Snow, 1997), performance on AAK was positively related to the number of algebra II units, and negatively related to the number of algebra I units. Students also cited trigonometry as a knowledge source, often in conjunction with algebra. A typical statement was: "I used trigonometry for the notation, but algebra to solve the problem."

Spatial Visualization (SV)

Items representing the SV dimension were markedly distinguished from the other dimensions by the use of visualization and gestures as efficient solution strategies (see Table 6). For example, when asked to match patterns of an unfolded box, students frequently commented that they "tried to picture how it would look" and also used their hands to make folding movements. On an item that called for inferring three-dimensional shape from the rotation of a two-dimensional figure, successful students were more likely to use their hands to simulate the rotation patterns around an imaginary axis. Furthermore, successful students drew the geometric shapes that they mentally envisioned and traced out the axis of revolution on the figure. Students who did not augment their visualization with a picture often expressed confusion while considering different alternatives: "I don't understand how they want me to rotate it . . . By keeping this plane, that makes a cone. And this around like that . . . can make a sphere."

Students were likely to identify geometry as a source of knowledge, but were also likely to report that the items were unfamiliar, and that such problems were not formally taught in their classes. This is consistent with results of Hamilton, Nussbaum, and Snow (1997), who concluded that the knowledge necessary to

Table 6

Solution Strategies and Sources of Knowledge Used in Spatial Visualization (SV) Items

Strategy/Source of knowledge	Total	Correct	Incorrect
Analytic approach	.10	.11	.08
Elimination	.08	.05	.12
Guess	.13	.05	.24
Visualization	.59	.71	.40
Gestures	.21	.32	.04
Picture/graph	.21	.24	.16
No answer	.06	.00	.16
Geometry	.41	.39	.44
SAT/SAT prep	.08	.11	.04
Not familiar	.24	.24	.24

Note. Total: 63 responses; Correct: 38 responses; Incorrect: 25 responses.

answer SV items in science was not confined to classroom instruction. The large-scale regression analysis (Kupermintz & Snow, 1997) also indicated no relationship between performance on SV and number of geometry units.

Algebra Systems Comprehension (ASC)

ASC items elicited more use of multiple strategies in a single solution compared to other dimensions (see Table 7). Straightforward application of basic algebraic or geometric properties were likely to result in an incorrect solution. These items were associated with analytic reasoning and value assignment, often using both for the purpose of verifying the correctness of the solution. Incorrect responses that depended solely on an analytic approach, for example, and did not assign numerical values to the variables could not benefit from a confirmation of the hypothesized relationship.

A successful approach was characterized by statements such as: "Since x is squared, I would say because x is doubled that would raise [the other variable] to the fourth. But then I always check my answer with numbers." Unsuccessful students were more likely to assume their logic or intuition was sufficient to yield a

Table 7

Solution Strategies and Sources of Knowledge Used in Algebra Systems Comprehension (ASC) Items

Strategy/Source of knowledge	Total	Correct	Incorrect
Basic algebraic properties	.03	.00	.13
Geometric properties	.05	.00	.20
Analytic approach	.35	.33	.40
One-step formula	.17	.17	.20
Value assignment	.41	.48	.20
Elimination	.24	.19	.40
Visualization	.11	.10	.13
General math	.16	.17	.13
Algebra	.48	.46	.53
Algebra II	.17	.19	.13
Geometry	.16	.17	.13

Note. Total: 63 responses; Correct: 48 responses; Incorrect: 15 responses.

correct answer and did not pursue the problem further: "I looked at the graph, tried to figure out the slope . . . it's close enough, the only [option] here that fits the formula." Thus, successful students were not only able to reason about an underlying algebraic relationship, but also demonstrated that they could employ an efficient hypothesis testing strategy. The hypothesis testing component was not identified in the initial dimension interpretation.

Algebra was clearly the major source of knowledge for ASC items, equally represented among correct and incorrect solutions, consistent with the context in which such domain knowledge was employed.

Discussion

Scores on standardized multiple-choice tests (but also on tests using other formats) are typically taken as measures of a generalized construct using labels such as "mathematics achievement." Kupermintz and Snow (1997) have demonstrated that achievement on the NELS:88 mathematics test is not represented adequately by a unidimensional construct, and that several distinct performance dimensions can be identified and measured. The analyses presented here extend that work by exploring, through probing students' work as they attempt test items, the cognitive factors that underlie performance on different dimensions.

The combination of results from large-scale statistical analyses and small-scale interview studies sharpens the distinctions among clusters of items appearing on a single test form, and thus offers an alternative interpretation of mathematics achievement measures. As mentioned before, large-scale psychometric analysis may suggest plausible hypotheses about the type of knowledge or reasoning involved in test performance; such hypotheses are summarized in the initial interpretations of achievement dimensions and are based on reasoning about common features of items grouped together in a dimension. These hypotheses, however, should not be taken as a definitive or sufficient validity argument for the interpretation of scores based on empirically derived dimensions. More appropriately, they should be considered as a useful starting point to guide further theoretical and empirical investigation, in an iterative process that will support, refine, or challenge score interpretations. Further insight can be gained by examining the relationships of the achievement dimension with student, program, and instructional variables, as in Kupermintz and Snow (1997).

Cognitive analyses of student performance subject construct interpretation to additional inquiry that may yield further support or suggest the plausibility of rival hypotheses. For example, our think-aloud data provided clear support for the initial interpretation of the CMR dimension, but led to qualifying the NMR interpretation as application of domain knowledge rather than abstract reasoning. Further investigations can now target more focused hypotheses about content and process in test performance on specific dimensions. Examples of other relevant strategies may include using general reference construct measures (such as fluid intelligence) and experimental manipulation of item content and format.

This study highlights the need to consider the design and development of methods for item construct validation more directly and fully. The design and evaluation of assessment tools typically rely on expert judgment rather than empirical analysis to determine what they measure. However, test specification tables, routinely a part of psychometric reports, often fail to reflect important psychological distinctions that emerge from a more rigorous process analysis. A rich empirical investigation of task performance can detect cognitive and other task demands, resulting in a better understanding of the constructs being measured, and, consequently, better score interpretation and use. Clearly, such analysis, using methods reported here but also other procedures, could be useful in test development by providing evidence to support and clarify differences among items, identify sources of construct-irrelevant variance, and guide scoring.

Students bring different combinations of abilities and experiences to a testing situation, and these influence their responses to particular tasks. Thus, as demonstrated in the current study, items may elicit different response processes from different test takers (see also Haertel, 1985). This makes the validation process even more complex because it suggests that the degree to which a test is considered valid for a particular purpose may vary across examinees. It is worthwhile, therefore, to consider various student characteristics, cognitive and others, and the ways in which each relates to test validity. As pointed out by Snow (1993), affective and conative variables should be brought to bear on test performance as they interact with the repertoire of knowledge and skills that students employ in the testing situation. Motivation, for example, is likely to play an important role in determining which solution strategies are invoked and in regulating or sustaining commitment to execute the strategy. The incorporation of noncognitive factors is needed to further our appreciation of the complexity of test performance and to

advance the development of appropriate theories that aim to bridge the psychometric and cognitive psychological perspectives.

Test validation requires a series of empirical investigations and analyses. Theoretical considerations are needed to ground test performance in a broader perspective of thinking about knowledge and skill in a domain. Evidence obtained by observing students working on test items and asking them to introspect about their performance is an important method to "thicken" the necessary empirical base but is of limited use without an adequate theoretical framework. Empirical findings about solution strategies and other processes exhibited by students in task performance should ultimately be understood within the context of a coherent model linking student characteristics, task demands, and testing purposes. With the growing demands on assessment systems and the proliferation of test formats, a unifying theoretical framework of academic achievement is greatly needed to guide progress in test development, scoring, and use.

References

Frederiksen, N., Glaser, R., Lesgold, A., & Shafto, M. (Eds.). (1990). *Diagnostic monitoring of skill and knowledge acquisition*. Hillsdale, NJ: Lawrence Erlbaum Associates.

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

Haertel, E. H. (1985). Construct validity and criterion-referenced testing. *Review of Educational Research*, *55*, 23-46.

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

Kupermintz, H., & Snow, R. E. (1997). Enhancing the validity and usefulness of large-scale educational assessments: III. NELS:88 mathematics achievement to 12th grade. *American Educational Research Journal*, *34*, 124-150.

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

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: Macmillan.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performance as scientific inquiry into score meaning. *American Psychologist*, *50*, 741-749.

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

Snow, R. E., & Lohman, D. F. (1989). Implications of cognitive psychology for educational measurement. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 262-331). New York: Macmillan.