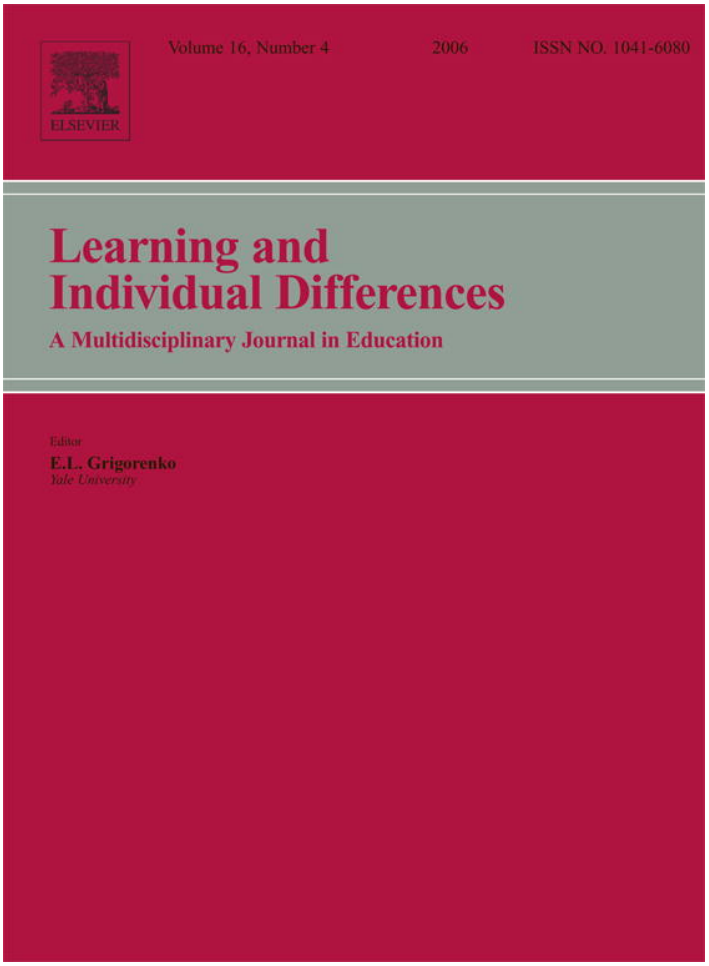

Available online at www.sciencedirect.com
ScienceDirect
Learning and Individual Differences 16 (2006) 267–276
Learning and Individual Differences
www.elsevier.com/locate/lindif

Exploring individual and item factors that affect assessment validity for diverse learners: Results from a large-scale cognitive lab ☆

Phoebe C. Winter, Rebecca J. Kopriva *, Chen-Su Chen [1], Jessica E. Emick

Center for the Study of Assessment Validity and Evaluation, Department of Measurement, Statistics, and Evaluation, University of Maryland, 1230 Benjamin Building, College Park, MD 20742

Received 30 March 2005; accepted 14 December 2006

Abstract

A cognitive lab technique ($n=156$) was used to investigate interactions between individual factors and item factors presumed to affect assessment validity for diverse students, including English language learners. Findings support the concept of *access* — an interaction between specific construct-irrelevant item features and individual characteristics that either permits or inhibits student response to the targeted measurement content of an assessment item. Access issues of 3rd and 5th grade students were explored at three stages during problem solving on mathematics items: 1) apprehension of task demands; 2) formulation of a solution; and 3) articulation of a solution. Adequacy of access at these levels appears to affect student performance. In particular, where students were able to experience increased apprehension in the first stage through the provision of item variations consistent with their individual access needs, they were more likely to formulate correct solution strategies. The implications of these findings are discussed and suggestions for future research are offered.

Within the psychological and educational arena many assessments are conceptualized with little regard for individualized differences (Messick, 4). These tests, which use targeted constructs, fail to fully consider the construct irrelevant factors (e.g., interactions between language, culture, preferred processing strategies) that are incidentally measured. For example, a mathematics problem designed to measure a student's ability to determine perimeter and area may function also (or only) as a test of a student's reading ability if that question is linguistically complex. If the problem requires that students provide their answer via a diagram or list of steps, then the ability to explain through pictures or words is being measured along with understanding of perimeter and area. To the extent that students face difficulty with the construct-irrelevant requirements of an item, the item becomes less able to provide accurate information about their achievement in the area of the targeted construct.

☆ This research was supported by Grant R305T010846-03 from the U.S. Department of Education, Institute of Education Sciences. The contents of this article reflect the views of the researchers and do not represent the policy of the Department of Education, nor is endorsement by the federal government implied.

* Corresponding author.
E-mail address: rkopriva@umd.edu (R.J. Kopriva).
[1] Chen-Su Chen is now at American Institutes for Research.

1041-6080/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.lindif.2007.01.001

The examples in the preceding paragraph help illustrate the concept of *access*, the interaction between construct-irrelevant item features and person characteristics that either permits or inhibits student response to the targeted measurement content of the item. Access issues appear to be particularly complex for English language learners, other language and cultural minority students, and those with sizeable literacy and attention-based challenges (Helwig, Rozek-Tedesco, Tindal, Heath, & Almond, 1999; Kopriva & Lara, 1998; Solano-Flores & Trumbull, 2003). Specifically, language and cultural factors have been shown to play a significant role in mathematics assessment, where difficulties with understanding context and language can impede students' ability to understand and solve problems (Abedi & Lord, 2001; August & Hakuta, 1997; Cuevas, 1984; Cummins, Kintsch, Reusser, & Weimer, 1988; Kopriva, 2000; Kopriva & Saez, 1997; LeCelle-Peterson & Rivera, 1994; Mestre, 1988; Solano-Flores & Trumbull, 2003). Access is also an issue for native English speakers who are poor readers (Clements, 1980; Helwig et al., 1999; Newman, 1977) and those who have learning disabilities (Tindal, Heath, Hollenbeck, Almond, & Harniss, 1998). Research to date has tended to focus on broad group-level differences and on the effectiveness of various test modifications, including linguistic simplification and read-aloud options, for group members (Abedi, Lord, Hoffstetter, & Baker, 2000; Helwig et al., 1999; Homan, Hewitt, & Linder, 1994). What is needed now for the improvement of test development and administration and test score interpretation is a more finely grained analysis of the interaction between individual and item factors.

We have conducted exploratory research to gain insight into how access functions, which individual student and item factors appear most salient in determining student access to item content, and whether item variations might be developed to respond to the access needs of individual students, while keeping the targeted construct constant. This article reports the procedures and results of a series of cognitive labs undertaken with elementary students, including large numbers of English language learners, to explore student-item interactions and ways measurement accuracy might be increased through the careful provision of planned item variations.

Individual student variables (see Fig. 1) were selected based upon research in the areas of test accommodations for ELLs (e.g., Abedi et al., 2000; Kopriva, 2000), cultural issues involved in assessing linguistically diverse students (e.g., August & Hakuta, 1997; Solano-Flores & Trumbull, 2003), and cognitive factors involved in mathematics problem-solving (e.g., Bransford, Brown, & Cocking, 2000; Calfee & Chambliss, 1999; Henningsen & Stein, 1997; Kilpatrick,

The student's capacity to:

- read academic material at grade level
- write in an academic context at grade level
- communicate fluently in English
- derive meaning from visual representations and graphical displays (visual, receptive)
- create diagrams and pictures to convey meaning (visual, productive)
- derive and create meaning by manipulating 2-D and 3-D mental images (spatial)
- create meaning by touching and manipulating objects and/or moving his/her body (tactile/kinesthetic)
- understand and use contextual information in a problem
- structure a problem-solving approach
- understand and solve problems that are visually or linguistically complex
- understand and follow test instructions
- focus on and attend to test questions

The student's

- experiences in the home culture that allow him or her to understand the ideas contained either implicitly or explicitly in the context of a problem or needed to convey an appropriate response
- length of time in the United States, in U.S. schools, and in schools in his or her country of origin
- literacy in home language other than English
- schooling experiences that allow him or her to understand assessment expectations, conditions, and questioning methods and appropriately convey his/her response according to local school conventions

Fig. 1. Student factors hypothesized to interact with items to support or inhibit student access to the item content.

Swafford, & Findell, 2001) as well as reviews of ELL instructional materials and recommendations, particularly in mathematics (e.g., Jarrett, 1999; Northwest Regional Educational Laboratory, 2003).

1. Method

Researchers used a cognitive laboratory retrospective probing technique that incorporated mathematics testlets followed by structured recall to investigate the relationships between student characteristics and item features. This technique enabled the exploration of access issues firsthand and in depth, as a diverse group of students (including current and former English language learners, poor readers, and special education students) attempted to solve mathematics problems and identify features of items that inhibited or promoted their problem solving. Given the resource-intensive nature of cognitive labs, research using this method is typically limited to a sample size of 10 to 50. The use of focused interviews to guide data collection enabled researchers to include a much larger sample, which in turn supported a richer, more robust exploration of the student-factor and item-feature interaction.

The impact of access was explored at three stages during the mathematical problem-solving process: apprehension of task demands, formulation of the solution, and articulation of the solution. This model is conceptually related to the four-stage model proposed by Mayer (1987) involving problem translation, problem integration, solution planning, and solution execution. In our model, the first two steps are collapsed under "apprehension of task demands." The three-stage model served as the framework for observations, coding, and analyses.

1.1. Sample

The study involved 84 third-graders and 72 fifth-graders from eight public schools in suburban Maryland. The study began with 92 third-graders and 76 fifth-graders and their teachers. Seven were removed due to absence and five to administrative error (e.g., incorrect forms or unavailable manipulatives). Participants represented a range of SES, race/ethnicity, and academic ability found in the district. Table 1 shows selected sample demographic variables.

2. Materials

2.1. Mathematics test forms

Two areas of mathematics were assessed at Grades 3 and 5: number and algebra. Multiple choice, short constructed-response, and constructed-response items were used, reflecting the item types used on the state test given in the district. For each grade level, a set of base items-released items from large-scale tests reflecting the participating district's mathematics standards-was created. Each base item was analyzed to determine its core mathematics content and construct-related features that were critical to measuring the targeted construct. Using these as a guide, *item clusters* were developed for each base item. We defined an item cluster as a related set of items, each of which measured the same core mathematics content but which varied in the degree to which they required the use of construct-irrelevant knowledge and skills. The idea behind

Table 1
Sample Demographics

Variable	Grade 3	Grade 5	Variable	Grade 3	Grade 5
	N (%)	*N* (%)		*N* (%)	*N* (%)
Free or reduced lunch			Gender		
Currently eligible	34 (40.5)	18 (25)	Female	43 (51.2)	37 (51.4)
Formerly eligible	9 (10.7)	26 (36.1)	Male	41 (48.8)	35 (48.6)
Not eligible	41 (48.8)	28 (38.9)	ESOL Status		
Race/ethnicity			Currently ESOL	32 (38.1)	19 (26.4)
American Indian	0 (0)	1 (1.4)	Not ESOL	52 (61.9)	53 (73.6)
Asian American	25 (29.8)	13 (18.1)	Special Education Status		
African American	14 (16.7)	12 (16.7)	Receiving services	13 (15.5)	13 (18.1)
Hispanic	27 (32.1)	15 (20.8)	Not receiving services	71 (84.5)	59 (81.9)
White	18 (21.4)	31 (43.1)			

an item cluster is to provide various ways students can access an assessment task so that the level of their construct-irrelevant skills does not interfere with their demonstrating their knowledge of the mathematics content measured by the item. See Kopriva and Winter (2003), for more information about cluster development procedures.

A typical item cluster might consist of an item written in English and the same one written in Spanish, and items featuring a more familiar context, clearer language, access to manipulatives, and/or a graphic. The items in each cluster were designed to differ *only* in the type and degree of construct-irrelevant factors they evoked, not in the targeted construct they measured or the relative level of difficulty. Targeted construct template specifications were used to develop item variations that minimized the effects of factors that were not being targeted for measurement, including students' cultural experience, processing approaches, reading ability, and English language acquisition, as well as other construct-irrelevant cognitive factors.

For each base item in a cluster, researchers developed an equivalent item measuring the same core mathematics item construct. Base-equivalent items used similar numbers and the same format as the base items, and included the same features. For example, if the base item was set in a story context, so was the base-equivalent item; if the base used a table of data, so did the equivalent. Each equivalent item was developed to make sure it had the same syntactic and semantic structure and complexity of the base item and was at the same level of difficulty as the base.

2.2. Teacher questionnaire

Prior to the cognitive laboratories, teachers completed a detailed questionnaire about each participating student to identify the student's use of strategies in mathematics problem solving, assessment experiences, mathematics knowledge specifically keyed to the items used, English/language arts skills, and participation in ESOL and special education services. Teacher responses were used as a basis for determining the item variations students would receive along with the four base-equivalent mathematics items during the cognitive labs.

2.3. Student observation and interview protocol

During the cognitive laboratory sessions, a researcher observed a student as he or she solved the mathematics items and then interviewed the student about factors relating to the items and the student's interactions with the items. Researchers used a protocol to structure their observations and as a source for questions to the student about how the student understood what each question was asking, selected a problem-solving strategy, and chose to communicate an answer. The protocol also included questions that asked students to compare the difficulty-related features of each pair of items in a cluster. Researchers recorded their observations and student responses on the protocol and used questions at the end of the protocol to summarize their reflections based on the entire session with the student.

2.4. Teacher interview protocol

To obtain additional information about student factors that might interact with test items, researchers conducted phone interviews with teachers after the items were administered. In particular, teachers were asked to comment on the learning strategies used by each student at three stages: when attempting to understand a mathematics question or problem, when selecting a method or approach to solve it, and when articulating the answer.

2.5. Procedures

The study involved the following steps: (1) gathering student background information through teacher questionnaires, interviews, and school records; (2) matching test items to students; (3) administering test items during cognitive labs; (4) coding cognitive laboratory student protocols; and (5) scoring item responses.

2.6. Gathering student information

Researchers collected data from teachers and schools on a number of student factors hypothesized to interact with items in a way that would either support or inhibit student access to the item content. A profile was created for each student based on teacher responses to the questionnaire and information about the student from district records.

2.7. Matching test items to students

Each student was assigned four items that were matched to their student profiles in terms of containing features that would minimize barriers to access and four base-equivalent items from the same clusters as the matched items. Matching was done via a computerized algorithm that compared vectors of student factors from student profiles to vectors of item features within each cluster. In the first stage, reading, writing, and English acquisition requirements of the item were compared to the student's reading, writing, and English acquisition levels to create a matching score in English language factors for the student/item combination.

In the second stage, all other student factors and item features were compared to create a matching score in cognitive features for the student/item combination. Thus, each item in a cluster had two matching scores for the student. First, the items with the highest matching scores in English language features were selected for consideration; the item within that set with the highest matching score on cognitive features was selected for the student. Researchers reviewed each item selected by the algorithm and replaced it if needed when a teacher's questionnaire response further explained student needs. Table 2 illustrates sample item features used to match individual student factors.

Individual student background variables were also used to determine the type of administration and response conditions the student would receive. Administration conditions included written English items, written Spanish items, side-by-side English and Spanish, oral English, and oral Spanish. Response conditions included oral response in English or the student's preferred language and modeled response. During the cognitive labs, students who were English language learners were paired with test administrators with extensive ESL backgrounds. Where possible, beginning ELL students were further paired with test administrators who spoke their native languages, including Spanish and Mandarin Chinese.

2.8. Administering test items during cognitive labs

Each student was individually administered a test with four items typical of grade-level standardized mathematics assessments (base-equivalent items) and four matched items, with the order of administration counterbalanced across the sample. During the 45-minute cognitive lab session, the student completed all the items while the researcher observed and recorded what the student did according to a standardized protocol. After the student completed the eight items, the administrator conducted an in-depth interview, according to a standardized protocol, about how the student interpreted each item, what processes he or she used to solve each item, and what features of each item helped or hindered the student's response. Questions were targeted to three stages of problem solving that are involved in access: comprehending the problem, formulating a solution, and articulating a solution. The researcher also asked the student questions comparing the features of the base-equivalent item to the matched item from the same cluster. Test items were administered by research assistants who completed standardized training and quality of

Table 2
Examples of student factors and item features

Student factors	Base-equivalent item features	Matched item features
Home culture emphasizes adult authority	Problem context requires disagreement with adult judgment	Different problem context in which adult judgment is not a factor
Low reading level	Text written at grade level	Clarified English and/or explanatory graphic
Low writing level	Written response required	Response may be provided through pictures or manipulatives
Represents knowledge concretely	Abstract representation of task, where the form of representation is not germane to targeted mathematics content	Addition of graphic organizer such as a chart, with explicit directions regarding use
Easily overwhelmed by complex items	Complex format unrelated to targeted mathematics content	Ample use of white space, simplified text format, and no extraneous information or graphics
Low English acquisition	Syntactically complex text	Clarified English and/or translation

Table 3
Correlations between process categories and item responses

	Apprehension	Strategy	Application	Score
Apprehension	–	.676	.378	.550
Strategy	.604	–	.331	.718
Application	.131	.196	–	.694
Score	.476	.742	.517	–

Unit of analysis = student/item combination. Grade 3 is above the diagonal, $n = 427$; Grade 5 is below the diagonal, $n = 436$.
All correlations significant at $p < .001$.

data collection was monitored by senior staff. Audio-tapes were collected for each session and were transcribed for additional quality control and coding.

2.9. Coding cognitive laboratory protocols

The cognitive laboratory protocols were structured to address several areas of interest regarding student interactions with items and specific item features, and results were coded so that researchers could compare the base-equivalent and matched items according to those features. In addition to standard quality control reviews (e.g., whether the student received inappropriate help from the administrator), protocols were coded to produce several variables, including:

- whether the student used specific features of the items (e.g., graphics, manipulatives);
- whether the student found specific features of the items helpful in understanding or responding to the item;
- the degree to which students apprehended the task posed by item;
- whether the student chose an appropriate solution strategy; and
- whether the student applied the selected strategy (regardless of appropriateness) correctly.

These features were coded without reference to the score the student received on the item. Five research assistants coded data under the supervision of senior staff. The assistants completed standardized training.

2.10. Scoring item responses

Multiple-choice items were scored as correct or incorrect according to an item key. Constructed-response items were scored according to cluster-specific rubrics developed by three mathematics educators and a measurement expert, in consultation with ESL specialists. All items in a cluster were scored according to the same rubric. The five scorers were mathematics experts and educators with experience teaching ELL students who were familiar with the items and the

Table 4
Strategy by apprehension

Apprehension	Solution strategy					
	All items		Base-equivalent items		Matched items	
	Inappropriate *N* (%)	Appropriate *N* (%)	Inappropriate *N* (%)	Appropriate *N* (%)	Inappropriate *N* (%)	Appropriate *N* (%)
Grade 3						
No apprehension	31 (86.1)	5 (13.9)	16 (80.0)	4 (20.0)	15 (93.8)	1 (6.3)
Partial apprehension	43 (64.2)	24 (35.8)	23 (63.9)	13 (36.1)	20 (64.5)	11 (35.5)
Full apprehension	22 (6.6)	311 (93.4)	13 (8.2)	146 (91.8)	9 (5.2)	165 (94.8)
Grade 5						
No apprehension	16 (84.2)	3 (15.8)	12 (92.3)	1 (7.7)	4 (66.7)	2 (33.3)
Partial apprehension	33 (80.5)	8 (19.5)	17 (81.0)	4 (19.0)	16 (80.0)	4 (20.0)
Full apprehension	30 (7.9)	352 (92.1)	16 (8.4)	174 (91.6)	14 (7.3)	178 (92.7)

Table 5
Application by strategy

Solution strategy	Application of strategy					
	All items		Base-equivalent items		Matched items	
	Incorrect *N* (%)	Correct *N* (%)	Incorrect *N* (%)	Correct *N* (%)	Incorrect *N* (%)	Correct *N* (%)
Grade 3						
Inappropriate	47 (48.5)	50 (51.5)	21 (41.2)	30 (58.8)	26 (56.6)	20 (43.5)
Appropriate	51 (14.3)	306 (85.7)	29 (16.4)	148 (83.6)	22 (12.2)	158 (87.8)
Grade 5						
Inappropriate	23 (29.9)	54 (70.1)	11 (25.6)	32 (74.4)	23 (35.3)	22 (64.7)
Appropriate	45 (12.3)	322 (87.7)	20 (11.0)	161 (89.0)	25 (13.4)	161 (86.6)

study. A mathematics educator who participated in rubric development provided three hours of training. Each constructed-response item was scored twice. If the scorers disagreed, a single score was determined by consensus after discussion, with consultation from other scorers and the trainer as needed.

3. Results

During the cognitive laboratories, students were asked to describe what they thought each item was asking them to do, how they solved the problem, and why they chose their response (for multiple-choice items) or why they represented their answer as they did (constructed-response items). Answers to these questions, administrator observations, student responses in their comparisons of item pairs, and student work were coded according to three process categories: the degree to which students (1) apprehended the problem task, (2) used an appropriate solution strategy, and (3) applied the

Table 6
Recursive regression of process variables on score for Grade 3

Variable entered: Application					
Adjusted R square (SE): .480 (.328)					
	Unstandardized coefficients		Standardized coefficients	*t*	Sig.
	B	S.E.	Beta		
(Constant)	.071	.034		2.067	.039
Application	.767	.039	.694	19.868	.000
Variables entered: Application, strategy					
Adjusted R square (SE): .749 (.228)					
	Unstandardized coefficients		Standardized coefficients	*t*	Sig.
	B	S.E.	Beta		
(Constant)	−.245	.028		−8.737	.000
Application	.566	.028	.512	19.912	.000
Strategy	.604	.028	.549	21.330	.000
Variables entered: Application, strategy, apprehension					
Adjusted R square (SE): .750 (.228)					
	Unstandardized coefficients		Standardized coefficients	*t*	Sig.
	B	S.E.	Beta		
(Constant)	−.228	.034		−6.784	.000
Application	.572	.029	.517	19.599	.000
Solution	.624	.036	.567	17.093	.000
Apprehension	−.021	.025	−.029	−.861	.389

strategy they chose correctly (regardless of whether the strategy was an appropriate one). Task apprehension was coded on a 0–2 scale, with 0 indicating no or incorrect apprehension of the task, 1 indicating partial apprehension, and 2 indicating full apprehension. Appropriate strategy and correct application were each scored yes/no. If the protocol and student work did not contain sufficient evidence about a category, the category was coded as missing.

The relationships between task apprehension, appropriateness of solution strategy, correctness of application of solution strategy, and accuracy of item responses (expressed as proportion of total possible points) were investigated using the student/item combination as the unit of analysis, resulting in 427 data points in Grade 3 and 436 in Grade 5. Table 3 shows the correlations between the process categories and item responses for Grade 3 above the diagonal and Grade 5 below the diagonal. As expected, the variables are all positively related.

It was hypothesized that the three process categories and the item score had a recursive chain relationship, as follows:

apprehension → strategy → application → response

Table 4 shows the probabilities of using an appropriate solution strategy given a level of apprehension, for all items and for the two item types, and Table 5 shows the probabilities of applying the strategy correctly given the appropriateness of the strategy. The probability of using an appropriate strategy increases sharply as apprehension goes from partial to full for both types of items, particularly in Grade 5. The relationship between the appropriateness of the strategy used and the accuracy of its application is stronger in Grade 3 than in Grade 5, and in Grade 5, the increase in the conditional probability of correct application is not as dramatic as the conditional probability of using an appropriate application given apprehension level.

Tables 6 and 7 show the results of a recursive regression of apprehension, strategy, and application according to the hypothesized relationship for Grades 3 and 5 for all items (because the results were similar for both item types, only the full analysis is shown). The regression results support the chain hypothesis. When strategy is added in to the regression

Table 7
Recursive regression of process variables on score for Grade 5

Variable entered: Application					
Adjusted R Square (SE): .266 (.353)					
	Unstandardized coefficients		Standardized coefficients	*t*	Sig.
	B	S.E.	Beta		
(Constant)	.238	.044		5.357	.000
Application	.605	.048	.517	12.593	.000
Variables entered: Application, strategy					
Adjusted R square (SE): 692 (.228)					
	Unstandardized coefficients		Standardized coefficients	*t*	Sig.
	B	S.E.	Beta		
(Constant)	−.237	.035		−6.825	.000
Application	.452	.032	.386	14.249	.000
Strategy	.729	.030	.666	24.547	.000
Variables entered: Application, strategy, apprehension					
Adjusted R square (SE): 692 (.228)					
	Unstandardized coefficients		Standardized coefficients	*t*	Sig.
	B	S.E.	Beta		
(Constant)	−.273	.049		−5.575	.000
Application	.452	.032	.386	14.233	.000
Strategy	.706	.037	.644	19.093	.000
Apprehension	.031	.029	.035	1.062	.289

with application, the explained variance in scores increases. However, when apprehension is added, there is no increase in explained variation. These results indicate that increasing the degree to which students apprehend a task affects the probability that they will select an appropriate solution strategy, and that the effect of apprehension on score is wholly mediated by the appropriateness of the strategy used and the correctness of the application of that strategy.

4. Discussion and conclusion

Study results indicate that performance on mathematics items is a function of a student's knowledge and skill in specific mathematics constructs and a student's adequacy in accessing the item content, and that this impact of student access varies over individual students and over items. Results show that construct-irrelevant individual characteristics and item factors interact with each other in ways that affect access, and access can be usefully modeled as mediated through apprehension of item requirements, formulation of a solution, and articulation of response.

When students were provided with item variations that improved their ability to apprehend what the test question was asking, they were much more likely to select a correct process for solving the problem, and therefore more likely to get the correct answer. Qualitative and statistical evidence that improving apprehension did not automatically lead to correct responses, that the relationship was mediated by solution processes, suggests that the item variations were not easier than the standard items in terms of the targeted mathematics, but rather, that they simply provided the necessary entrée to the content of the item. Results from this study further suggest that when students are allowed multiple methods of response (except where the ability to communicate in a certain way is what is being tested), they may be better able to show their thinking.

To improve the validity of assessments, it is essential that we understand how students with particular characteristics interact with specific items throughout the problem-solving process. Cognitive laboratories can provide a window on the extent to which items are measuring what they are intended to measure for individual students. The use of a relatively large sample ($n=156$) and carefully structured interviews allowed us to investigate a large number of potentially relevant variables and interpret results with a higher level of comfort than might have been possible given a more typically sized cognitive lab of 10 to 50 students.

Close attention to individual differences will help formulate future research questions related to student and item factors that can be investigated using experimental methods. Study results suggest, for example, that it may not be as simple as grouping ELLs and non-ELLs when looking at item functioning. Rather, it may be that there are underlying issues of access that cut across diverse groups, among them, ELLs, students with literacy issues, and students with attention problems. Within the category of ELLs, results suggest that there is a need for more complex analyses to encompass other potentially salient variables, including length of time in the country, familiarity with U.S. culture, proficiency in English and the first language, and compensatory skills that enable students to work successfully around construct-irrelevant language factors. Gathering information about student-item interactions allows us to identify how items are not working, for whom. Once this understanding is obtained, items and assessment procedures can be improved.

References

Abedi, J., Lord, C., Hoffstetter, C., & Baker, E. (2000). Impact of accommodation strategies on English language learners' test performance. *Educational Measurement: Issues and Practice*, *19*(3), 16−26.

Abedi, J., & Lord, C. (2001). The language factor in mathematics tests. *Applied Measurement in Education*, *14*(3), 219−234.

August, D., & Hakuta, K. (Eds.). (1997). *Improving schooling for language minority students: A research agenda.* Washington, DC: National Academy Press.

Bransford, J., Brown, A., & Cocking, R. (Eds.). (2000). *How people learn: Brain, mind, experience, and school.* Washington, DC: National Academies Press.

Calfee, R., & Chambliss, M. (1999). Cognitive perspectives on primers and textbooks. In D. A. Wagner, B. V. Street & R. L. Venezky (Eds.), *Literacy: An international handbook* (pp. 179−185). Boulder, CO: Westview Press.

Clements, M. A. (1980). Analyzing children's errors on written mathematics tasks. *Educational Studies in Mathematics*, *11*, 1−21.

Cuevas, G. J. (1984). Mathematics learning in English as a second language. *Journal for Research in Mathematics Education*, *15*, 134−144.

Cummins, D., Kintsch, W., Reusser, K., & Weimer, R. (1988). The role of understanding in solving word problems. *Cognitive Psychology*, *20*, 405−438.

Helwig, R., Rozek-Tedesco, M., Tindal, G., Heath, B., & Almond, P. (1999). Reading as an access to mathematics problem solving on multiple-choice tests for sixth-grade students. *Journal of Educational Research*, *93*(2), 113−125.

Henningsen, M., & Stein, M. (1997). Mathematical tasks and student cognition: Classroom-based factors that support and inhibit high-level mathematical thinking and reasoning. *Journal of Research in Mathematics Education*, *28*(5), 524−549.

Homan, S., Hewitt, M., & Linder, J. (1994). The development and validation of a formula for measuring single-sentence test item readability. *Journal of Educational Measurement*, *31*, 349−358.

Jarrett, D. (1999). *The inclusive classroom: Teaching mathematics and science to English-language learners.* Portland, OR: Northwest Regional Educational Laboratory.

Kilpatrick, J., Swafford, J., & Findell, B. (Eds.). (2001). *Adding it up: Helping children learn mathematics.* Washington, DC: National Academies Press.

Kopriva, R. (2000). *Ensuring accuracy in testing for English language learners.* Washington, DC: Council of Chief State School Officers.

Kopriva, R. J., & Lara, J. (1998). Scoring English language learners' papers more accurately. In Y. S. George & V.V. Van Horne (Eds.), *Science education reform for all* (pp. 77−82). Washington DC: American Association for the Advancement of Science.

Kopriva, R., & Saez, S. (1997). *Guide to scoring LEP student responses to open-ended mathematics items.* Washington, DC: Council of Chief State School Officers.

Kopriva, R. K., & Winter, P. C. (2003). Construct validity: What are we really measuring? *Presentation at the annual National Conference on Large-Scale Assessment, San Diego, CA*.

LeCelle-Peterson, L., & Rivera, C. (1994). Is it real for all kids? A framework for the equitable assessment policies for English language learners. *Harvard Educational Review*, *641*(1), 55−75.

Mayer, R. E. (1987). *Educational psychology: A cognitive approach.* Boston: Little, Brown.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (pp. 13−103). (3rd ed.). New York: American Council on Education and Macmillan.

Mestre, J. (1988). The role of language comprehension in mathematics and problem solving. In R. Cocking & J. Mestre (Eds.), *Linguistic and cultural influences on learning mathematics* (pp. 201−220). Hillsdale, NJ: Lawrence Erlbaum.

Newman, M. A. (1977). An analysis of sixth-grade pupils' errors on written mathematics tasks. In M. A. Clements & J. Foyster (Eds.), *Research in mathematics education in Australia*, *Vol. 2* (pp. 239-258). Melbourne: Swinburne.

Northwest Regional Educational Laboratory. (2003). *Strategies and resources for mainstream teachers of English language learners.* By Request series Portland, OR: Author.

Solano-Flores, G., & Trumbull, E. (2003). Examining language in context: The need for new research and practice paradigms in the testing of English language learners. *Educational Researcher*, *32*(2), 3−13.

Tindal, G., Heath, B., Hollenbeck, K., Almond, P., & Harniss, M. (1998). Accommodating students with disabilities on large-scale tests. An empirical study of student response and test administration demands. *Exceptional Children*, *64*, 439−450.