Behavior Research Methods
2009, 41 (2), 414-424
doi:10.3758/BRM.41.2.414

Diagnosing students' misconceptions in algebra: Results from an experimental pilot study

Michael Russell and Laura M. O'Dwyer
Boston College, Chestnut Hill, Massachusetts

and

Helena Miranda
Florida Gulf Coast University, Fort Myers, Florida

Computer-based diagnostic assessment systems hold potential to help teachers identify sources of poor performance and to connect teachers and students to learning activities designed to help advance students' conceptual understandings. The present article presents findings from a study that examined how students' performance in algebra and their overcoming of common algebraic misconceptions were affected by the use of a diagnostic assessment system that focused on important algebra concepts. This study used a four-group randomized cluster trial design in which teachers were assigned randomly to one of four groups: a "business as usual" control group, a partial intervention group that was provided with access to diagnostic tests results, a partial intervention group that was provided with access to the learning activities, and a full intervention group that was given access to the test results and learning activities. Data were collected from 905 students (6th–12th grade) nested within 44 teachers. We used hierarchical linear modeling techniques to compare the effects of full, partial, and no (control) intervention on students' algebraic ability and misconceptions. The analyses indicate that full intervention had a net positive effect on ability and misconception measures.

Student assessment is a central component of the instructional process. As Airasian (1991) and Popham (1995) describe, effective classroom instruction begins by assessing each student's current state of knowledge and understanding. To build on and advance students' knowledge, teachers develop and deliver instruction. Following initial instruction, assessment is used to make inferences about the extent to which new knowledge has developed and to inform the subsequent steps in the instructional process. Although the extent to which teachers are able to make valid assessments and successfully tailor instruction to meet the needs of their students varies widely, assessment and effective instruction are tightly intertwined (Airasian, 1991; Anderson, 2003; Pellegrino, Chudowsky, & Glaser, 2001; Popham, 1995).

The importance of assessment in the instructional process is reflected in the No Child Left Behind Act of 2001, Public L. No. 107-110, §1606, 115 Stat. 1425 (2002), which requires that states implement assessments of all students in Grades 3–8 in reading and mathematics. The assumption underpinning the establishment of standards and state-level assessment systems is that they motivate teachers and schools to improve student learning and encourage teachers to focus on specific types of learning (Shepard, 1990).

Unfortunately, research suggests that much of the information provided to teachers by externally developed tests of student achievement is of limited instructional value. The limitation stems from the lack of new information provided to teachers by most achievement tests. In a series of studies conducted since 1962, teachers have been asked to predict the performance of their students on a variety of tests. Across these studies, teachers' predictions have correlated strongly with students' actual performance (Cullen & Shaw, 2000; Demaray & Elliot, 1998; Fuller, 2000; Hoge & Coladarci, 1989; Mulholland & Berliner, 1992).

The relative scarcity of new information provided by externally developed tests stems from two shortcomings of today's tests: a single initial focus on placing students on a unidimensional scale that represents ability within a broad domain and a subsequent focus on whether a student's response to an item is correct without consideration of the cognitive processes applied to reach a given response. Despite efforts to incorporate open-ended items into some tests, most test items result in binary information about a student: Namely, did the student answer the item correctly or incorrectly? Although scoring guides for some forms of open-ended items focus on the procedures and cognitive process students use to solve problems, the scoring of these items is dependent on students' descriptions of their processes, which are often incomplete and/or inaccurate reflections of the actual process of answering questions. As a result, these items provide only indirect and crude insight into examinees' cognitive processes.

M. Russell, russelmh@bc.edu

For example, despite including multiple-item formats and requiring several hours to complete, the Massachusetts Comprehensive Assessment System (MCAS) 8th grade mathematics tests provide information to teachers and schools based on roughly 50 items. This limited set of items is used to assess a broad domain, such that only a handful of items are available to measure critical subdomains, such as algebra, geometry, measurement, data analysis, and statistics. Moreover, the information provided to teachers about students' performance within a given subdomain is limited to the percentage of items that a given student answered correctly. Although this information may allow teachers to identify students who have not developed an adequate understanding within the subdomain, the test does not attempt to probe for or provide information about *why* a student may have performed poorly within that subdomain.

Beyond the MCAS, several state-developed and commercial tests, such as the Iowa Test of Basic Skills, attempt to help teachers diagnose students' weaknesses. These tests, however, focus on specific content within a given domain and employ a series of multiple-choice items to measure student performance within the several subdomains. As a result, the diagnostic information provided to educators is limited to an indication of whether students succeed or fail on items within a particular subdomain. Although this information helps educators identify those subdomains in which students may be in need of further instruction, these diagnostic tests tend to provide little or no information about *why* students may be struggling. Rather than diagnosing the misconceptions and/or specific skill sets that interfere with students' mastery of the subdomain, most current diagnostic tests do not provide any more information than achievement or mastery tests do (Russell, 2002).

With advances in technology and almost universal access to computers in schools, computer-based diagnostic assessments hold potential to help teachers identify the origins of misconceptions that interfere with the development of students' conceptual understanding. By focusing on the sources of a student's error and capitalizing on advances in the cognitive sciences and computer-based technology, well-crafted diagnostic tests can provide teachers with an assessment tool that simultaneously measures student performance and identifies sources of misconceptions and misunderstandings (Snow & Lohman, 1989). Truly diagnostic tests may also help deepen teachers' assessments of their students' conceptual understanding of a specific topic. Computer-based diagnostic assessment systems can connect students and teachers to instructional activities and resources targeted to help students correct their misconceptions.

One example of the type of diagnostic instruments envisioned by Snow and Lohman (1989) is the Force Concept Inventory (FCI), a tool used by physics teachers to help diagnose misconceptions related to the concepts of force, power, and energy (Hestenes, Wells, & Swackhamer, 1992). Composed of a series of forced-choice items, the FCI provides diagnostic information about the six conceptual dimensions that are essential to developing a comprehensive understanding of the concept of Newtonian force. With this information, teachers can modify their instruction to address a specific misconception. However, note that, because the FCI was developed as a paper-based instrument, it does not attempt to link teachers or students to instructional materials designed to target misconceptions identified among students.

To examine the efficacy of diagnostic assessments linked to instructional materials for improving student outcomes in the area of algebra, we developed a comprehensive online assessment system designed to measure students' comprehension of specific algebraic concepts, to identify those who hold specific misconceptions related to the measured concepts, and to connect teachers and students to instructional and learning resources that aim to assist students in restructuring their conception of a specific algebraic concept.

The Diagnostic Algebra Assessment System (DAAS) is a classroom assessment and instructional tool that comprises a battery of online diagnostic algebra tests that provide estimates of students' abilities and misconceptions, immediately available performance reports that detail each student's current ability and misconception status, and lesson plans and instructional activities that target specific misconceptions.

As part of a 3-year study, a cluster-randomized trial was conducted to examine the effect that the use of the DAAS by teachers and their students had on measures of student understanding and misconceptions for three important algebraic concepts: variables, equality, and graphing. The present article provides a detailed description of the DAAS and the three algebraic concepts, describes the research design, presents findings, and discusses implications for instructional use and future research.

THE DAAS

In 2001, the National Research Council wrote that "Advances in the cognitive and measurement sciences make this an opportune time to rethink the fundamental scientific principles and philosophical assumptions serving as the foundations for current approaches to assessment" (Pellegrino et al., 2001, p. 1). This charge served as the inspiration for the DAAS, which was developed to examine the feasibility and effect of combining findings from research in the cognitive sciences with advances in computer-based testing to develop an assessment system that provides teachers with timely, individualized, and meaningful information about students' understanding of important algebraic concepts. Guided by past research on algebraic misconceptions, a set of short online tests was developed to measure students' understanding of three of many algebraic concepts: variables, equality, and graphing. Unlike traditional achievement tests that provide a unidimensional measure of student ability, the DAAS tests were designed not only to provide a measure of ability but to identify the presence of specific misconceptions. In addition, for each misconception, a set of instructional and learning activities was developed to assist those students identified with a particular misconception to refine their understanding of the targeted concept. The sections that

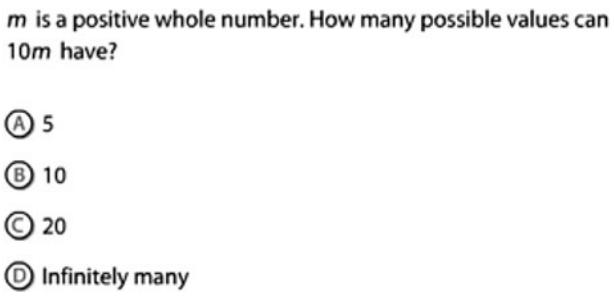


Figure 1. Example of a concept-of-a-variable item. Option D is the correct response; Option B is the misconception response.

follow provide a brief description of each component of the DAAS. Note that additional sample test items and instructional materials can be accessed at www.bc.edu/research/intasc/researchprojects/DiagnosticAlgebra/daa.shtml.

Test Items

The DAAS uses a multipurpose assessment strategy in which a set of 10–12 multiple-choice items is used to simultaneously estimate ability and diagnose each misconception. Each of the multiple-choice test items has four response options: a correct response, a misconception response, and two distractors. Each misconception response option was designed to measure a single misconception. The two distractors were designed so that they would not tap the misconception but would be obtained through other types of errors. Students who selected 35% or more of the misconception responses were classified as having the misconception, and their performance reports were flagged for teachers. Note that a validity study conducted prior to the efficacy study presented here provided evidence that students who select misconception responses more than 35% of the time demonstrate the application of the misconception across a wide array of problems that measure that concept and that the classification is consistent across multiple measures (Russell, Kay, & Miranda, 2008). Figure 1 displays a sample item for assessing misconceptions about variables.

The item in Figure 1 asks students to select the best response for the following question: "*m* is a positive number. How many values can 10*m* have?" Option D is the correct response, and Option B represents the misconception response: A student with this misconception tends to ignore the variable represented by the letter *m* and instead assigns to the letter the number (in this case, 10) that is associated with it in the expression.

Performance Reports

After students complete a DAAS test, the system automatically scores students' responses for ability and misconception and generates two reports that immediately are available to teachers. The *ability report* summarizes students' performance on the tests when items are scored for ability only (i.e., item correct/incorrect). The *misconception report* summarizes students' performance when items are scored for misconception (i.e., misconception option selected/not selected). Each report presents a student × test item matrix that indicates whether the student answered the item correctly and whether the student selected the misconception option. Figure 2 presents an example of a misconception report.

Instructional Interventions

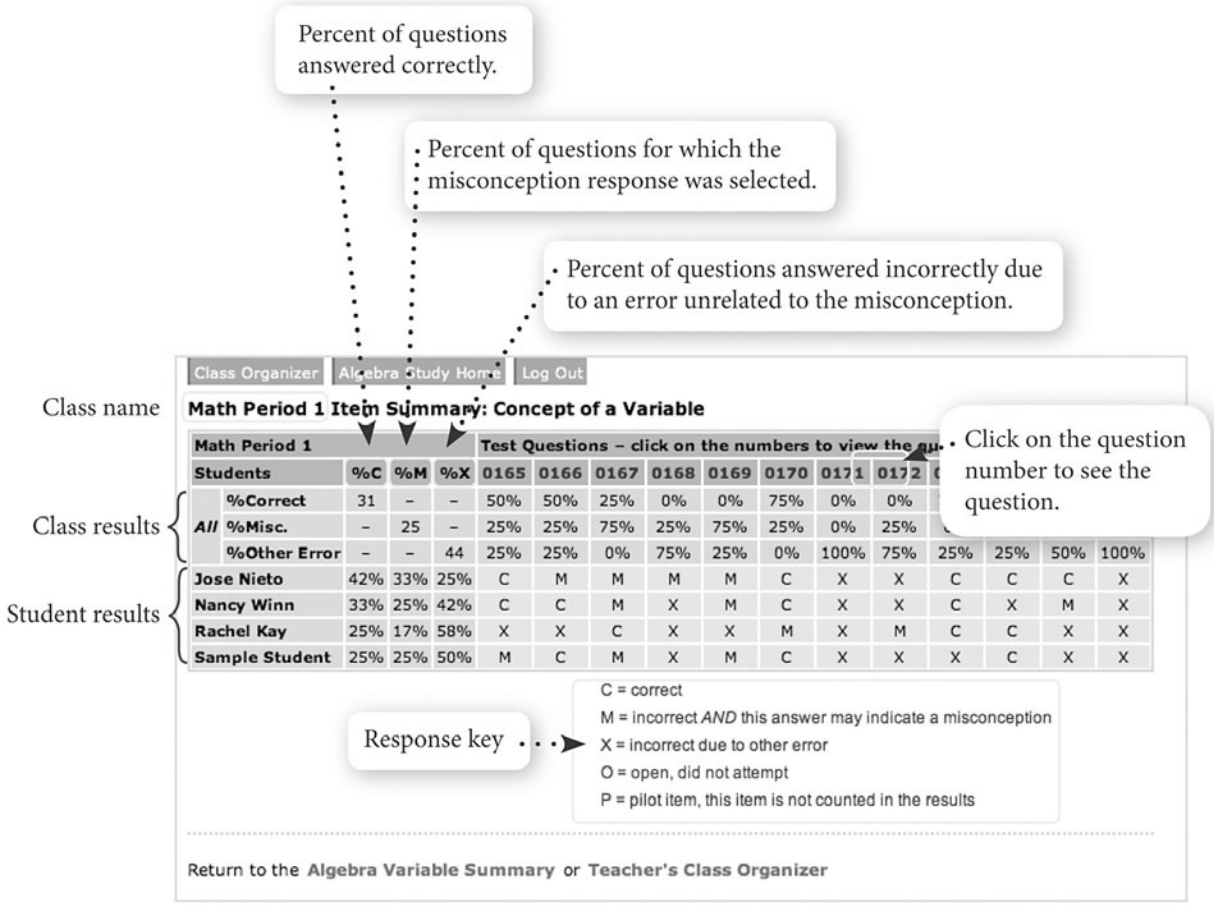
Along with the ability and misconception performance reports, the DAAS provides each teacher with two lesson plans and accompanying materials for each misconception that has been diagnosed in his/her classroom. The lesson plans and learning activities were created by curriculum developers with expertise in algebraic misconceptions and are designed to help teachers address the algebraic misconceptions identified among their students. Each lesson plan lists the objectives of the lesson related to the skills or concepts associated with the specific misconception and provides teachers with a detailed description of the misconception, definitions associated with the concepts covered in the lesson plans, and a list of the materials needed for the instructional intervention. The lesson plans also describe activities that may be used with individual students or small groups, independent practice exercises, and answer keys that identify misconception and correct response options (Russell et al., 2008).

Misconceptions Measured by the DAAS

Research suggests that algebraic misconceptions impede the acquisition of concepts crucial to algebra achievement (Birenbaum, Kelly, & Tatsuoka, 1992; Clement, 1982; Mestre, 1987; Schwartzman, 1996; Stacey & McGregor, 1997). Hiebert and Carpenter (1992) explain that misconceptions arise when students fail to link new knowledge to previous knowledge for which the brain has established cognitive networks. If new knowledge is not anchored to existing networks, to solve new problems, students rely on strategies developed through their experience with similar material.

Although an error is considered to be a random or haphazard mistake, misconceptions arise when students incorrectly apply previously learned strategies to solve new problems (Hiebert & Carpenter, 1992). Experts believe that errors that signal deeper misunderstandings about algebraic concepts are not haphazard but are systematic and derive from experience with arithmetic or from student-created theories (Clement, 1982; Davis, 1971; Herscovics, 1989; Kieran, 1992; Küchemann, 1981; Matz, 1980; Rosnick, 1981; Rosnick & Clement, 1997; Stacey & McGregor, 1997). A review of research on this topic has identified 15 algebra misconceptions (Russell et al., 2008). Although each misconception warrants the development of a separate diagnostic instrument, the number of misconceptions on which the DAAS initially focused was restricted by time and limited resources. As such, diagnostic tests were developed and validated for misconceptions with respect to variables, equality, and graphing (Booth, 1984; Clement, 1989; Davis, 1971; Kieran, 1992; Küchemann, 1978, 1981; Rosnick, 1981; Stacey & McGregor, 1997).

Concept of a variable. Understanding the concept of a variable requires students to recognize that letters have



Math Period 1					Test Questions – click on the numbers to view the qu											
Students		%C	%M	%X	0165	0166	0167	0168	0169	0170	0171	0172				
All	%Correct	31	–	–	50%	50%	25%	0%	0%	75%	0%	0%				
	%Misc.	–	25	–	25%	25%	75%	25%	75%	25%	0%	25%				
	%Other Error	–	–	44	25%	25%	0%	75%	25%	0%	100%	75%	25%	25%	50%	100%
Jose Nieto		42%	33%	25%	C	M	M	M	M	C	X	X	C	C	C	X
Nancy Winn		33%	25%	42%	C	C	M	X	M	C	X	X	C	X	M	X
Rachel Kay		25%	17%	58%	X	X	C	X	X	M	X	M	C	C	X	X
Sample Student		25%	25%	50%	M	C	M	X	M	C	X	X	X	C	X	X

Figure 2. Sample misconception report.

referents (i.e., abstractions of something rather than arbitrary entities) and to realize that a letter represents a number (Matz, 1980). According to Rosnick (1981), misconceptions about variables may be defined as a failure to understand the role of letters in equations and the tendency to interpret letters in equations as labels referring to concrete objects. Students who have this misconception fail to differentiate among ways in which letters may be used in equations. That is, they fail to distinguish between the use of letters to denote a concrete entity instead of a variable standing for an abstract number of things. For the tests designed to diagnose such misconceptions around variables, misconception responses reflected the assignment of a concrete numerical value to a letter intended to represent a variable (Kieran, 1992; Küchemann, 1978, 1981).

Equality. Students are first exposed to the equal symbol in arithmetic. In elementary school, students are taught to associate the equal symbol with a command to perform an operation (Sáenz-Ludlow & Walgamuth, 1998), instead of with an expression of a symmetric transitive relation between the expressions on the left- and right-hand sides of the symbol (Kieran, 1992). When learning algebraic concepts, students transfer their understanding of the equality symbol and interpret it as "makes" or "gives" and as a connector between the two sides of an equation (Stacey & McGregor, 1997). As such, students with the equality misconception experience difficulty with statements such as $7 = 3 + 4$ or $5 = 5$ because these do not involve a problem on the left and an answer on the right (Falkner, Levi, & Carpenter, 1999). Students with this misconception tend to experience difficulty with the idea that adding or subtracting the same amount from both sides of an equation maintains equality.

Graphing. Students of algebra learn that a graph is a representation for a function, and they learn to translate between graphs, equations, and tables of values. However, just as the translation between word problems and equations can be difficult, some students also find it challenging to interpret the graph of a real-world situation. These students tend to forget the algebraic relationships they have learned and resort to graphical misconceptions, often by treating a graph as a picture that represents a given scenario or by confusing slope with height (Clement, 1989). For example, consider a problem that requires students to draw a speed versus time graph for a cyclist riding over a

Table 1
Student Sample Composition by Grade and Ethnicity

	Ethnicity							
Grade	African American	Asian	Hispanic	Native American	White	Other	Did Not Answer	Percentage of Total by Grade
6		1						0.11
7	3	12	4		27	1	6	5.86
8	58	21	59	17	314	13	64	60.33
9	25	8	31	4	167	2	34	29.94
10	1	1	5	1	17		2	2.98
11					3	1		0.44
12	2				1			0.33
	89 (9.83%)	43 (4.75%)	99 (10.94%)	22 (2.43%)	529 (58.45%)	17 (1.88%)	106 (11.71%)	100

hill. A student who holds a misconception about graphing ignores what the problem asked of them and, instead of depicting the speed of the cyclist, draws an image of the hill. Students with the misconception do not understand that a graph represents speed as a function of time and instead conceive of the graph as a representation of the scenario.

METHOD

Research Design

The theoretical framework guiding this research assumes that some students hold a specific misconception that interferes with their understanding of a algebraic concept, that an online test that measures understanding the concept and any related misconception can identify students who are likely to hold that misconception, that the provision of a timely report for each test can help teachers identify students who hold a specific misconception, that providing access to instructional and learning materials designed to help students alter their conception of a given algebraic misconception can help teachers help students correct a given misconception, and that the above will result in higher performance on a test of a given algebraic concept and in a decreased application of the misconception.

In spring 2007, a pilot study was conducted to investigate the effect that the use of the DAAS has on students' understanding of three algebraic concepts. For this study, all participating teachers and their students were provided access to the DAAS misconception tests. In addition, teachers were assigned randomly to one of four groups, each of which was given access to different resources built into the DAAS.

Specifically, a four-group cluster-randomized controlled trial (C-RCT) was conducted. For this study, one group of teachers (Group 4) received the "full intervention," which comprised access to ability and diagnostic test scores and to the instructional materials. To disentangle the potential effects of the diagnostic test results and the instructional materials, Group 2 was provided access to the ability and diagnostic score reports but not to the instructional materials. Group 3 was provided access to the ability score reports and the instructional materials but not to the diagnostic score reports. Group 1 served as a control group and was not given access to the diagnostic score reports or instructional materials. Like all groups, the control group had access to the algebra concept tests and the ability score reports without any information about the related misconception. The specific intervention received by each group was as follows: Group 1 (control; ability reports only), Group 2 (partial intervention; ability and misconception reports only), Group 3 (partial intervention; ability reports only and instructional intervention), and Group 4 (full intervention; ability reports, misconception reports, and instructional intervention).

Varying the level of access to the misconception reports and instructional materials allowed the research team to examine whether the DAAS was effective for improving students' content knowledge in algebra and decreasing students' misconceptions and whether the varying amount of access to the instructional materials was associated with improved performance and decreased misconception scores.

Under the C-RCT design, participating teachers were assigned to conditions randomly and were treated as clusters within which students were nested. Before being provided access to the DAAS, students completed a diagnostic pretest and a background questionnaire. After pretests were completed, teachers in all groups (1–4) were given the ability reports for their students. Teachers in Group 2 were also provided with the misconception reports, in addition to which, teachers in Groups 3 and 4 were given access to the assigned components of the DAAS for 3 weeks, during which those teachers were asked to use the DAAS information and resources to inform their instructional practices. At the end of the 3-week intervention period, the students of the teachers in each group completed the diagnostic posttest.

Participants

Teachers were eligible to participate in the research provided that they were currently teaching algebra at the middle school or high school level (Grades 6–12) and that they had adequate access to computers, either in their classroom or in a laboratory setting. Teachers were recruited via e-mail sent to several math teachers' listservs, including the National Council of Supervisors of Mathematics, California, and New Hampshire listservs. Teachers who responded to the solicitation and who met the two criteria for participation were then randomly assigned to one of the four treatment conditions.

A total of 60 teachers volunteered to participate in the study and met the criteria for inclusion. Of those, 44 completed the data collection requirements within the allotted time frame. Collectively, these 44 teachers administered the diagnostic pre- and posttests to 905 students. The majority of participating teachers had been teaching for more than 5 years (75%), most held a master's degree (64%), and most were certified to teach mathematics (77%). Most were female (78%), and most were white (84%).

Because of different attrition rates across groups and different class sizes, the final sample sizes varied across groups (Group 1, 11 teachers, 227 students; Group 2, 17 teachers, 278 students; Group 3, 7 teachers, 153 students; Group 4, 9 teachers, 247 students). All of the students were enrolled in an algebra class, and 90% were in either Grade 8 or 9. Of the 905 students in the sample, 457 (50.5%) were female, 448 (49.5%) were male, 10% were African American, 11% were Hispanic, 58% were White, 5% were Asian, 2% were Native American, and 12% did not identify their ethnic background. Table 1 presents the composition of the student sample by grade and ethnicity.

Instrumentation: Diagnostic Tests

The diagnostic pre- and postintervention algebra tests each contained 34 items (approximately 10–12 multiple-choice items per misconception) and were matched with respect to content and difficulty. These tests contained three subtests, each focusing on a specific algebra misconception. The items on the posttest were isomorphic

Table 2
Diagnostic Testlet Information for Each Ability/Misconception Scale

Misconceptions	Number of Items	Range of Component Loadings	Reliability Estimates (Cronbach's alpha)	
			Ability	Misconception
Variables	12	0.30–0.70	.96	.85
Equality	10	0.30–0.90	.95	.86
Graphing				
Slope and height confusion	7	0.50–0.80	.90	.71
Height and rate confusion	5	0.40–0.80	.96	.70

equivalents of the pretest items, and on field tests were found to have psychometric properties nearly identical to those of their respective original items (Russell et al., 2008). We used principal components analysis to confirm the existence of unidimensional scales relating to each of the misconceptions. The Concept of a Variable and Equality testlets each comprised one scale; whereas the Graphing testlet comprised two, one related to slope and height confusion and the other related to height and rate confusion. Table 2 summarizes the number of items in each testlet, the range of component loadings, and the reliability estimates.

The component loadings for the 12-item Concept of a Variable scale ranged from 0.30 to 0.70, with a reliability estimate (Cronbach's alpha) of .96 for items scored for ability and .85 for items scored for misconception. The component loadings for the 10-item Equality scale ranged from 0.30 to 0.90, and the reliability estimates were .95 with respect to items scored for ability and .86 with respect to items scored for misconception. The 7-item Graphing scale related to slope and height confusion had component loadings ranging from 0.50 to 0.80, and the reliability estimate was .90 with respect to items scored for ability and .71 with respect to items scored for misconception. The 5-item Graphing scale related to height and rate confusion had component loadings ranging from 0.40 to 0.80, and the reliability estimate was .96 with respect to items scored for ability and .70 with respect to items scored for misconception.

The outcome measures for this study (i.e., scores for ability and misconception) were calculated from the diagnostic pre- and posttests. The ability score was defined as the total number of correct responses on the 34-item test; the misconception score was defined as the total number of items for which the student selected a misconception response on the same 34-item test. The *mean class score* was defined as a class's average total score on the DAAS test battery with respect to items scored for ability (i.e., the average number of correct responses on the 34-item test).

Data Analysis

To examine the treatment effect, we first looked at the standardized differences among the four groups and subsequently modeled students' posttest scores (ability and misconception) as a function of their group membership, after we controlled for students' pretest scores. The standardized effect sizes were calculated for combinations of the treatment conditions to examine (1) the difference between the mean scores for the full intervention group (Group 4) and those for the control (Group 1) and partial intervention groups (Groups 2 and 3), and (2) the difference between the mean scores for groups in which teachers had access to the lesson plans and classroom activities (Groups 3 and 4) and those for groups in which teachers did not (Groups 1 and 2).

Because the unit of assignment was at the teacher level within which students were clustered, nesting of students within teachers was accounted for in our estimation of the treatment effect. Specifically, a two-level, hierarchical linear regression model or linear mixed model was used. This analytic approach accounts for the dependence among students nested within the same teacher, and correct standard errors of the regression coefficients are estimated (Raudenbush & Bryk, 2002). In the models, student outcomes (posttest ability and misconception scores) were modeled at Level 1 as a function of their pretest ability and misconception scores and at Level 2 by their teachers' membership in one of the four treatment conditions. The hierarchical regression analyses focused on the extent to which membership in any one of the treatment groups had a differential effect on students' ability and misconception scores after we controlled for initial differences in achievement. For these analyses, students' ability and misconception scores were standardized to have a mean of 0 and a standard deviation (*SD*) of 1.

RESULTS

Before presenting analyses that address the research questions, we summarize the percentages of students who were classified as holding each misconception (variables, 14%; graphing, 12%; equality 11%) and the percentages of students identified as holding one, two, or three misconceptions (one misconception, 19%; two misconceptions, 6%; three misconceptions, approximately 1%).

Table 3 presents descriptive statistics for the pre- and posttest ability and misconception scores for each treatment group. Mean pretest ability scores ranged from 20.05 for Group 3 to 26.78 for Group 4, whereas posttest ability scores ranged from 21.29 for Group 3 to 27.66 for Group 4. Students' mean pretest misconception

Table 3
Pretest and Posttest Descriptive Statistics for Ability and Misconception Measures

Treatment Condition	Scored for Ability				Scored for Misconception*			
	Pretest		Posttest		Pretest		Posttest	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Group 1	21.91	7.49	22.24	7.60	4.98	3.70	4.50	3.45
Group 2	22.83	6.20	22.99	6.64	4.33	2.99	4.10	3.05
Group 3	20.05	7.97	21.29	8.12	5.48	3.81	4.90	3.80
Group 4	26.78	5.53	27.66	5.68	2.61	2.58	2.33	2.53
Total (N = 905)	23.21	7.10	23.79	7.33	4.22	3.40	3.85	3.31

Note—Lower misconception scores represent fewer misconceptions. *Maximum score = 34.

Table 4
Effect Size Differences for Posttest Ability Scores

Comparison Pair	*M*	*SD*	Effect Size
Full Treatment Group (4) Versus Control Group (1) and Partial Intervention Groups (2 and 3)			
Group 1 ($n = 227$)	22.24	7.60	0.81
Group 4 ($n = 247$)	27.66	5.68	
Group 2 ($n = 278$)	22.99	6.64	0.76
Group 4 ($n = 247$)	27.66	5.68	
Group 3 ($n = 153$)	21.29	8.12	0.91
Group 4 ($n = 247$)	27.66	5.68	
Groups Receiving Instructional Intervention (3 and 4) Versus Groups That Did Not (1 and 2)			
Groups 1 and 2 ($n = 505$)	22.65	7.09	0.36
Groups 3 and 4 ($n = 400$)	25.23	7.39	

scores ranged from 2.61 for Group 4 to 5.48 for Group 3, and posttest misconception scores ranged from 2.33 for Group 4 to 4.90 for Group 3.

Effect sizes were calculated to permit standardized comparisons between the groups' posttest means. We were interested in two comparisons: (1) the standardized difference between the mean for the full treatment group (Group 4) and the means for the control group (Group 1) and the partial intervention groups (Groups 2 and 3) and (2) the standardized difference between the means for the groups for which teachers had access to the instructional intervention in the form of lesson plans and classroom activities (Groups 3 and 4) and the means for those that did not have this access (Groups 1 and 2). The effect sizes were computed separately for the items scored for ability and misconception.[1]

Table 4 shows that, with respect to ability scores, the standardized difference between the means for Group 4 and Group 1 was 0.81 *SD*s, and the mean for Group 4 was 0.76 *SD*s higher than that for Group 2 and 0.91 *SD*s higher than that for Group 3. The direction of the effect size indicates that students in the group that received the full intervention had higher ability scores than did students in the groups that did not. Table 4 shows that the groups for which teachers implemented instructional intervention (Groups 3 and 4) had scores that were 0.36 *SD*s higher than did groups for which teachers did not (Groups 1 and 2).

Table 5
Effect Size Differences for Posttest Misconception Scores

Comparison Pair	*M*	*SD*	Effect Size
Full Treatment Group (4) Versus Control Group (1) and Partial Intervention Groups (2 and 3)			
Group 1 ($n = 227$)	4.50	3.45	−0.72
Group 4 ($n = 247$)	2.33	2.53	
Group 2 ($n = 278$)	4.10	3.05	−0.63
Group 4 ($n = 247$)	2.33	2.53	
Group 3 ($n = 153$)	4.90	3.79	−0.80
Group 4 ($n = 247$)	2.33	2.53	
Groups Receiving Instructional Intervention (3 and 4) Versus Groups That Did Not (1 and 2)			
Groups 1 and 2 ($n = 505$)	4.28	3.24	−0.30
Groups 3 and 4 ($n = 400$)	3.31	3.32	

Table 5 shows similar results with respect to items scored for misconceptions. The standardized mean difference between Group 4 and the control group and two partial intervention groups were −0.72, −0.63, and −0.80 for Groups 1, 2, and 3, respectively. The direction and magnitude of these effect sizes indicate that the students in Group 4 had lower misconceptions scores than did students in the partial intervention groups had and that the effect sizes were in the moderate-to-large range (according to Cohen's conventions). The bottom panel of Table 5 shows that students in the groups with which teachers implemented the instructional intervention (Groups 3 and 4) had misconception scores that were 0.30 *SD*s lower than those for students without access to the intervention (Groups 1 and 2).

Although these calculations do not take into account either the nesting of students within teachers or the pretest differences among the groups, these findings provide preliminary evidence that students who received the complete intervention had better student outcomes (i.e., higher ability scores and lower misconception scores) than did students in the groups that received only partial intervention. In addition, the results indicate that there were differences between the groups that had access to the instructional intervention and those that did not.

To account for the fact that random assignment occurred at the teacher level and the nesting of students within teachers, multilevel regression models were applied to estimate the treatment effect for ability and misconception scores. The additional benefit of these analyses is that students' pretest scores could be included as a covariate at the student level. This is particularly important for these data because, even with random assignment, there were differences observed in the pretest scores.

Two hierarchical linear regression models were formulated for ability and misconception scores. Paralleling the comparison made with the effect sizes, we were interested in two subsequent comparisons: (1) the difference between the mean outcome scores for the full treatment group (Group 4) and the means for each of the control and partial intervention groups (Groups 1–3) after we controlled for pretest differences (Model 1) and (2) the difference between the mean outcome scores for the groups for which teachers had access to the instructional intervention (Groups 3 and 4) and the mean for the groups that did not (Groups 1 and 2), after we controlled for pretest differences (Model 2). Tables 6 and 7 present the results from the hierarchical regression analyses in which posttest outcome data for 905 students nested in 44 teachers were modeled as a function of their pretest outcome measures and their teachers' membership in the treatment conditions. Recall that to facilitate the interpretation of the regression coefficients in the multilevel models, students' ability and misconception scores were standardized to have a mean of 0 and an *SD* of 1.[2]

Model 1 in Table 6 presents the results when three dummy variables were added at Level 2 to represent

Table 6
Multilevel Models for Predicting Students' Standardized Ability Scores

	β	*SE*	*t* Test
	Model 1		
Intercept	.06	0.05	$t(40) = 1.20, p = .24$
Student-Level Predictors			
Pretest ability scores	.89	0.02	$t(900) = 55.39, p < .01$
Teacher-Level Predictors			
Group 4 vs. Group 1	−.13	0.06	$t(40) = -2.08, p < .05$
Group 4 vs. Group 2	−.09	0.06	$t(40) = -1.48, p = .15$
Group 4 vs. Group 3	.01	0.02	$t(40) = 0.08, p = .94$
	Model 2		
Intercept	−.05	0.03	$t(42) = -1.73, p = .09$
Student-Level Predictors			
Pretest ability scores	.90	0.02	$t(902) = 57.93, p < .01$
Teacher-level predictors			
Groups 1 and 2 vs. Groups 3 and 4	.11	0.05	$t(42) = 2.43, p < .05$

Note—$N_{\text{Teachers}} = 44$ and $N_{\text{Students}} = 905$. In these models, students' pretest scores grand mean centered in the Level 1 model. Only the intercepts were allowed to vary across groups. The regression slopes were not allowed to vary randomly, because of the lack of power for estimating stable random coefficient or slopes-as-outcomes models. Residual variance = 0.12 (within teachers) and 0.02 (between teachers). Significance of residual between-teacher variance [Model 1, $\chi^2(40, N = 44) = 137.43$, $p < .01$; Model 2, $\chi^2(43, N = 44), = 133.30, p < .01$].

teachers' membership in Group 1, 2, or 3 (each coded 1), or Group 4 (coded 0). In these models, the coefficients represent the effect size difference between the groups being compared after we controlled for students' pretest ability scores. The regression coefficients indicate that the mean ability score for Group 1 was significantly lower than the mean ability score for Group 4 after we controlled for students' pretest ability scores; students in Group 1 had posttest ability scores that were 0.13 *SD*s lower than those for students in Group 4 after we controlled for initial pretest differences [$t(40) = -2.08, p < .05$]. The 0.09 *SD* difference between Groups 2 and 4 was not statistically significant after we controlled for initial differences [$t(40) = -1.48, p > .05$]. Similarly, the effect size difference between Groups 3 and 4 after we controlled for initial differences (0.01) was also not statistically significant [$t(40) = 0.08, p > .05$].

Model 2 in Table 6 includes a dichotomous variable at Level 2 to represent teachers' membership in either Groups 1 and 2 (coded 0) or Groups 3 and 4 (coded 1). The coefficient associated with this comparison was statistically significant [$t(40) = 2.43, p < .05$]. After we controlled for students' pretest ability scores, students in the groups whose teachers implemented the instructional in-

Table 7
Multilevel Models for Predicting Students' Standardized Misconception Scores

	β	*SE*	*t* Test
	Model 1		
Intercept	−.07	0.07	$t(40) = -1.06, p = .30$
Student-Level Predictors			
Pretest misconception scores	.75	0.02	$t(900) = 31.68, p < .01$
Teacher-Level Predictors			
Group 4 vs. Group 1	.11	0.09	$t(40) = 1.22, p = .23$
Group 4 vs. Group 2	.09	0.09	$t(40) = 0.96, p = .59$
Group 4 vs. Group 3	.06	0.11	$t(40) = 0.54, p = .59$
	Model 2		
Intercept	.03	0.04	$t(42) = 0.60, p = .55$
Student-Level Predictors			
Pretest misconception scores	.75	0.02	$t(902) = 31.77, p < .01$
Teacher-Level Predictors			
Groups 1 and 2 vs. Groups 3 and 4	−.07	0.07	$t(42) = -1.02, p < .32$

Note—$N_{\text{Teachers}} = 44$ and $N_{\text{Students}} = 905$. In these models, students' pretest scores grand mean centered in the Level 1 model. Only the intercepts were allowed to vary across groups. The regression slopes were not allowed to vary randomly, because of the lack of power for estimating stable random coefficient or slopes-as-outcomes models. Residual variance = 0.30 (within teachers), 0.04 (between teachers, Model 1), and 0.03 (between teachers, Model 2). Significance of residual between-teacher variance [Model 1, $\chi^2(40, N = 44) = 131.33, p < .01$; Model 2, $\chi^2(43, N = 44), = 133.20, p < .01$].

tervention (Groups 3 and 4) were predicted to have scores that were 0.11 *SD*s higher than those for students whose teachers did not (Groups 1 and 2).

The results in Table 7 present similar multilevel analyses with respect to items scored for misconceptions. The results for Models 1 and 2 in Table 7 show that students' pretest misconception scores are significantly and positively related to students' posttest misconception scores. Unlike in the model for predicting ability scores, the coefficients representing the comparisons between Groups 1–4 were not statistically significantly different from zero. Results for Model 2 indicate that students whose teachers implemented the instructional intervention (Groups 3 and 4) had lower misconception scores than did students in the groups for which teachers did not (Groups 1 and 2). Unlike for the ability outcome, the coefficient for this predictor was not statistically significant with respect to items scored for misconceptions.

DISCUSSION

The importance of student assessment in education has grown rapidly over the past 20 years. Although much of this attention has focused on summative assessment, emphasis on formative assessment has increased recently. For example, the Council of Chief State School Officers (2008) has formed a working group composed of leaders from several states to help make educators more aware of the importance of formative assessment. Similarly, the Federal Enhancing State Assessment program, which provides funding to states to enhance assessment practices across their schools, has provided funding to help states develop formative assessment (U.S. Department of Education, 2008).

Within the field of formative assessment, educators are increasingly recognizing the importance of diagnostic assessment. In addition to identifying concepts and skills that students struggle to master, diagnostic assessment aims to identify the underlying reasons why an individual student struggles with a specific concept or skill. As described above, the DAAS was developed to assist teachers in identifying students within their classroom who struggle with a specific algebraic topic because of a misconception specific to that topic. To assist teachers in helping students adjust their understanding so that they no longer hold a misconception, the DAAS also provides teachers with direct access to instructional interventions that focus on the specific misconception. Thus, the DAAS was developed as an online assessment and instructional tool that teachers in the middle grades can use first to diagnose algebraic misconceptions held by individual students and then to help them reconceptualize important algebra concepts.

The present study is a first attempt at establishing the effectiveness of the DAAS as an assessment and instructional tool. The present findings provide preliminary evidence that the use of the DAAS's full set of features (diagnostic tests combined with ability and misconception reports and an instructional intervention) has a statistically significant positive effect on measures of student algebraic ability. Although the effect on the presence of misconceptions was not statistically significant, the trend for both measures of student algebraic ability and presence of misconceptions suggests that use of the full set of features is more effective than is the use of a subset of its features. Similarly, the use of the DAAS was found also to be more effective than the "business as usual" practice that typically provides teachers with a summative score that represents a student's status without also providing information about possible causes of low performance. Finally, multilevel analyses that accounted for the nesting of students within teachers and adjusted for prior differences on the pretest measures indicate that the full intervention of the DAAS had a net positive effect on ability and misconception measures, as compared with the three alternative levels of the intervention.

Although the findings from this pilot study are promising, the following six factors may limit their generalizability.

1. Although larger than those for many pilot studies, the sample sizes within each treatment group were still relatively small, comprising 905 students nested within 44 teachers. The small sample size resulted in underpowered analyses for the examination of interactions among demographic variables and treatment levels.

2. Because the proportion of students within a given classroom who are expected to hold a given misconception is small, the moderate number of student participants resulted in relatively small numbers of students who were identified as holding a given misconception: Of the 905 students in the study, only 127 were identified as having a misconception regarding the concept of a variable, 109 were identified as having the graphing misconception, and 95 were identified as having the equality misconception. Approximately 73% were identified as having no misconceptions. Given that the DAAS is intended to impact those students who have a given misconception, the relatively small number of students identified with a misconception also decreases the power to detect differences among the treatment groups.

3. Although classrooms were assigned randomly to treatment groups, initial differences in student algebraic test scores existed among groups. Although estimates were adjusted to account for initial group differences, random assignment to groups would have been more effective if pretest scores were collected prior to group assignment so that a stratified or blocked random assignment procedure could be employed.

4. The techniques used to recruit participants may have attracted teachers who were familiar with diagnostic assessment or were comfortable using technology as an instructional tool. Because many teachers may be unaware of diagnostic assessment or are not accustomed to using technology during mathematics instruction, the recruitment methods may have produced a sample that is not representative of all mathematics teachers nationwide.

5. Scores on the instruments used to measure change in ability and presence of a misconception were correlated. In fact, the correlation between ability and misconception posttest score was −.78 for the concept-of-a-variable items, −.80 for the equality items, and −.74 for the graphing items. It is important to note, however, that the correla-

tions were not the result of a one-to-one relationship. That is, although a high ability score is necessarily associated with a low misconception score, a low misconception score may be associated with either a high or a low ability score. In addition, a large positive increase in ability score may or may not be associated with large decrease in misconception (e.g., a student may initially perform poorly but not hold a misconception). Similarly, a large decrease in misconception scores does not necessarily correspond with a large increase in ability scores (e.g., a student may initially hold a misconception, correct the misconception, but still make other types of errors that result in a low ability score). Thus, although we recognize that there is correlation between ability and misconception scores, the lack of a one-to-one relationship suggests that it is appropriate to use these measures to examine the simultaneous effect of the DAAS on ability and misconceptions. Nonetheless, future researchers may want to employ a measure of ability that is composed of items that are distinct and separate from those used to measure misconception.

6. The present study focused on only three of several algebraic misconceptions identified in the literature. Although it is plausible that the inclusion of additional misconceptions would increase the effectiveness of the DAAS, it is also possible that reliable and valid measures for these other misconceptions cannot be developed. If that is so, the extent to which the DAAS can be used to help students overcome a broader set of algebraic misconceptions may be limited. Clearly, additional research is needed to examine whether the positive effects detected through this pilot study are generalizable to a larger, more representative sample of American classrooms and whether these findings apply to a broader set of algebraic concepts.

Despite these limitations, the results presented here provide preliminary evidence that the use of the DAAS may be more effective than the use of the conventional methods teachers currently use to assess and develop students' algebraic understanding. Our findings provide evidence that test items can be developed to reveal specific misconceptions. The study indicates that, when teachers use diagnostic information to identify misconceptions and subsequently use instructional strategies to help students reconceptualize those misconceptions, students' algebraic ability improves. In closing, this study suggests that the use of diagnostic assessment systems, such as the DAAS, promises to enhance teaching and learning by enabling teachers to more effectively assess student understanding in a timely manner, diagnose misconceptions, and then help students develop their understanding so that a given misconception is no longer held.

AUTHOR NOTE

The present research was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305H040099 to the Technology and Assessment Study Collaborative at Boston College. The opinions expressed are those of the authors and do not represent views of the U.S. Department of Education. We thank Joan Lucariello, Rachel Kay, and Michele Tully, who made invaluable contributions during the development of the misconception items that the DAAS tests comprise. Finally, we thank the many teachers and students who assisted us in piloting test items, validating the tests, and conducting the efficacy study presented here. Address correspondence to M. Russell, Measurement, Educational Research, and Evaluation Department, Boston College, Campion Hall, Room 332C, 140 Commonwealth Ave., Chestnut Hill, MA 02467 (e-mail: russelmh@bc.edu).

REFERENCES

Airasian, P. W. (1991). *Classroom assessment*. New York: McGraw-Hill.

Anderson, L. (2003). *Classroom assessment: Enhancing the quality of teacher decision making*. Mahwah, NJ: Erlbaum.

Birenbaum, M., Kelly, A. E., & Tatsuoka, K. K. (1992). *Towards a stable diagnostic representation of students' errors in algebra*. Princeton, NJ: Educational Testing Service (ERIC Document Reproduction Service No. ED 356973).

Booth, L. R. (1984). *Algebra: Children's strategies and errors*. Windsor, U.K.: NFER-Nelson.

Clement, J. (1982). Algebra word problem solutions: Thought processes underlying a common misconception. *Journal for Research in Mathematics Education*, **13**, 16-30.

Clement, J. (1989). The concept of variation and misconceptions in Cartesian graphing. *Focus on Learning Problems in Mathematics*, **11**, 77-87.

Council of Chief State School Officers (2008). *Formative assessment for students and teachers*. Retrieved May 3, 2008, from www.ccsso.org/projects/scass/Projects/Formative%5FAssessment%5Ffor%5FStudents%5Fand%5FTeachers/.

Cullen, J., & Shaw, S. (2000). *The accuracy of teacher prediction of student test performance for students referred to special education.* Danbury, CT: Western Connecticut State University, Department of Education and Educational Psychology.

Davis, R. B. (1971). Cognitive processes involved in solving simple algebraic equations. *Journal of Children's Mathematical Behavior*, **1**, 7-35.

Demaray, M. K., & Elliot, S. N. (1998). Teachers' judgments of students' academic functioning: A comparison of actual and predicted performances. *School Psychology Quarterly*, **13**, 8-24.

Falkner, K. P., Levi, L., & Carpenter, T. P. (1999). Children's understanding of equality: A foundation for algebra. *Teaching Children Mathematics*, **6**, 232.

Fuller, M. L. (2000, April). *Teacher judgment as formative and predictive assessment of student performance on Ohio's fourth and sixth grade proficiency tests*. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans.

Herscovics, N. (1989). Cognitive obstacles encountered in the learning of algebra. In S. Wagner & C. Kieran (Eds.), *Research issues in the learning and teaching of algebra* (pp. 60-86). Reston, VA: National Council of Teachers of Mathematics.

Hestenes, D., Wells, M., & Swackhamer, G. (1992). Force concept inventory. *Physics Teacher*, **30**, 141-158.

Hiebert, J., & Carpenter, T. (1992). Learning and teaching with understanding. In D. A. Grouws (Ed.), *Handbook of research on mathematics teaching and learning* (pp. 65-100). New York: Macmillan.

Hoge, R. D., & Coladarci, T. (1989). Teacher-based judgments of academic achievement: A review of the literature. *Review of Educational Research*, **59**, 297-313.

Kieran, C. (1992). The learning and teaching of school algebra. In D. A. Grouws (Ed.), *Handbook of research on mathematics teaching and learning* (pp. 390-419). New York: Macmillan.

Küchemann, D. E. (1978). Children's understanding of numerical variables. *Mathematics in School.* **7**, 23-26.

Küchemann, D. E. (1981). Algebra. In K. Hart (Ed.), *Children's understanding of mathematics.* London: Murray.

Matz, M. (1980). Towards a computational theory of algebraic competence. *Journal of Mathematical Behavior*, **3**, 93-166.

Mestre, J. (1987). Why should mathematics and science teachers be interested in cognitive research findings? *Academic Connections* (pp. 3-5, 8-11). New York: The College Board.

Mulholland, L. A., & Berliner, D. C. (1992, April). *Teacher experience and estimation of student achievement.* Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco.

NO CHILD LEFT BEHIND ACT OF 2001, PUB. L. 107-110, §1606, 115 STAT. 1425 (2002).

PELLEGRINO, J. W., CHUDOWSKY, N., & GLASER, R. (EDS.) (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academy Press.

POPHAM, W. J. (1995). *Classroom assessment: What teachers need to know.* Needham Heights, MA: Allyn and Bacon.

RAUDENBUSH, S. W., & BRYK, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. Newbury Park, CA: Sage.

ROSNICK, P. (1981). Some misconceptions concerning the concept of variable: Are you careful about defining your variables? *Mathematics Teacher*, **74**, 418-420.

ROSNICK, P., & CLEMENT, J. (1997). Learning without understanding: The effect of tutoring strategies on algebra misconceptions. *Journal of Mathematical Behavior*, **3**, 3-27.

RUSSELL, M. (2002). How computer-based technologies can disrupt the technology of testing. In *Technology and assessment: Thinking ahead—Proceedings from a workshop* (pp. 63-78). Washington, DC: National Research Council.

RUSSELL, M., KAY, R., & MIRANDA, H. (2008). *Diagnostic algebra assessment study technical report*. Chestnut Hill, MA: Technology and Assessment Study Collaborative.

SÁENZ-LUDLOW, A., & WALGAMUTH, C. (1998). Third graders' interpretations of equality and the equal symbol. *Educational Studies in Mathematics*, **35**, 153-187.

SCHWARTZMAN, S. (1996). Some common algebraic misconceptions. *Mathematics & Computer Education*, **30**, 164-173.

SHEPARD, L. (1990). Inflating test score gains: Is the problem old norms or teaching the test? *Educational Measurement: Issues & Practice*, **9**, 15-22.

SNOW, R. E., & LOHMAN, D. F. (1989). Implications of cognitive psychology for educational measurement. In R. L. Linn (Ed.), *Educational measurement* (pp. 263-331). New York: Macmillan.

SPYBROOK, J., RAUDENBUSH, S. W., LIU, X., & CONDON, R. (2006). *Optimal design for longitudinal and multilevel research: Documentation for the "Optimal Design" software. V. 1.76.*

STACEY, K., & MCGREGOR, M. (1997). Ideas about symbolism that students bring to algebra. *Mathematics Teacher*, **90**, 110-113.

U.S. DEPARTMENT OF EDUCATION (2008). Federal Register, Vol. 73, No. 23, p. 6487-6491. Retrieved May 5, 2008, from www.ed.gov/legislation/FedRegister/announcements/2008-1/020408d.pdf.

NOTES

1. The effect size was calculated using Cohen's *d*, so a pooled *SD* was used: $d = (\bar{X}_T - \bar{X}_C)/\sigma_{\text{pooled}}$, where

$$\sigma_{\text{pooled}} \sqrt{\frac{\left(\sigma_T^2 + \sigma_C^2\right)}{2}}.$$

For the effect sizes described here, Group 4 represents the treatment group and Groups 1–3 represent the comparison groups.

2. A power analysis was conducted to estimate the statistical power for the two-level hierarchical regression analyses. For an intraclass correlation coefficient of .15, $\alpha = .05$, with 905 students nested in 44 teachers and including a pretest covariate, $R^2 = .75$. The power is 0.80, to detect an effect size of 0.26 (Spybrook, Raudenbush, Liu, & Condon, 2006).

(Manuscript received November 23, 2008;
revision accepted for publication January 29, 2009.)