

**Robust Understanding of Algebra:
A Framework for Capturing Student Learning and Instructional Practices**

Jamie Wernet
Michigan State University

Jerilynn Lepak
Michigan State University

Kimberly Seashore
University of California – Berkeley

Daniel Reinholz
University of California – Berkeley

Sihua Hu
Michigan State University

Paper presented at the 2013 annual meeting of the American Educational Research Association,
San Francisco, CA.

Purpose of the Study

This paper reports on research completed through the Algebra Teaching Study (ATS, <http://gse3.berkeley.edu/research/ats>), a larger study focused on developing tools for linking teaching practices in middle school algebra and students' robust understanding necessary for solving complex contextual algebraic problems (Ayieko, Floden, Hu, Lepak, Reinholz, & Wernet, 2012). We hypothesize that analysis of classroom practices along a limited number of key dimensions should correlate with corresponding differences in the type of gains in competencies students display in their written work. We focus specifically on contextual word problems in algebra due to their central importance in the curriculum, the fact that they have been documented as a source of difficulty for students, and their potential for uncovering a wide variety of competencies (cf. Walkington, Sherman, & Petrosino, 2012). We call such mathematically rich word problems *contextual algebraic tasks* ([CATs], Ayieko, Floden, Hu, Lepak, Reinholz, & Wernet, 2012).

In this paper we focus on one component of the larger study: operationalizing and measuring changes in student understanding. Our assessments build on algebraic tasks developed by the Mathematics Assessment Resource Service ([MARS], available at <http://www.noycefdn.org/resources.php>). While these tasks and scoring rubrics help teachers see strengths and weaknesses in student understanding, the scoring rubrics were not designed to link *specific* classroom practices to *specific* aspects of robust algebraic understanding. Thus, we designed rubrics focused on specific competencies required for robust understanding (i.e. robustness criteria) to capture particular competencies we hypothesized students might be developing and to compare profiles of developing student competencies at the class-level. This detailed analysis allows nuanced links to specific instructional practices as captured in the classroom observation scheme developed for the larger study. We address the following research questions:

- *How can robust algebraic understanding be analyzed through meaningful subcomponents, changes in which can be measured over the course of the school year?*
- *What class profiles result from this detailed analysis, and what might they reveal about the opportunities to develop robust thinking in different classrooms?*

Theoretical Framing: Robustness Criteria

To answer the questions, “what would robust understanding of algebra necessary to solve contextual algebraic tasks (CATs, henceforth) look like, and how can we measure students development of this understanding?”, we first developed a set of five *robustness criteria* (RCs) for student understanding. These RCs play a central role in our project by helping us navigate the dialectic relationship between classroom observations and task selection and analysis. Simultaneously, these criteria helped us focus on specific classroom practices that seemed likely to lead to this robust algebraic understanding. The RCs are defined and operationalized in assessments and the observation scheme as follows:

RC1: Reading and interpreting text, and understanding the contexts described in problem statements. RC1 represents the extent to which students are able to understand problem situations. Schoenfeld (2004) identified this as a core part of problem solving. In many ways RC1 corresponds to Mayer’s *problem representation* phase in which the solver “constructs

a mental representation of the situation described in the problem” (in Brenner et al., 1997, p. 665).

RC2: Identifying salient quantities in a problem and articulating relationships between them. One objective of the Algebra 1 Content Standards (NCTM, 2000) is for students to “identify essential quantitative relationships in a situation and determine the class or classes of functions that might model the relationship” (p. 665). Indeed, one conceptualization of algebra itself is as the study of relationships among quantities (Usiskin, 1988). Driscoll (1999) also identified building rules to represent functions—considering how variables are changing and how the input is related to output by well-defined rules—as an algebraic habit of mind.

RC3A: Generating representations of relationships between quantities, and RC3B: Interpreting and making connections between representations. We have operationalized the important competencies of representing quantitative relationships into two subcriteria. Representations are a key part of creating a mathematical model of a given situation to solve a given word problem (cf. Schoenfeld, 2004), and in algebraic understanding more generally (cf. Chazan, 2000). Consistent with this literature, we consider algebraic representations to include coordinate graphs, bivariate tables, diagrams or picture, and variable equations. By making connections between these various representations, students develop a deeper, more integrated understanding of algebra (e.g., Brenner et al., 1997; Driscoll, 1999; Kieran, 2007).

RC4A: Executing calculations and procedures with precision, and RC4B: Checking plausibility of results. Solving complex algebraic problems typically involves a number of procedures and calculations, which students must be able to execute accurately. Once students arrive at a result, they must be able to connect the result to the problem context and reflect critically on their work (cf. NCTM, 2000; Schoenfeld, 2004).

RC5: Explain and justify reasoning. Explanation and justification are critical not only in algebra, but in mathematics in general (CCSS-M, 2010; NCTM, 2000). In order to support their solutions, students should be able to use general reasoning as well as justifications grounded specifically in the domain of algebra (cf. Yackel, 2001). Explanation can also be seen as a precursor to mathematical proof, the crux of higher-level mathematics (cf. Graham, Cuoco, & Zimmerman, 2010).

Table 1 provides a brief summary of how the RCs were operationalized for scoring student work and observing algebra instruction.

Table 1

Conceptualization of Robustness Criteria in Student Assessments and Classroom Observation

Robustness Criterion	Examples of Operationalizing the RCs	
	Students' Written Work	Classroom Events
RC 1 – Interpreting text, understanding context	Use of appropriate units and terms (both contextual and mathematical)	Definition or rewording of nonmathematical and technical terms Elaboration on context Attention to embedded mathematical situation
RC 2a – Identifying salient quantities	Correct choice of quantities to include in representations Correct choice of quantities in subparts of tasks	Identifying quantities in givens and required solution

RC 2b – Articulating relationships between quantities	Statements about relationships between quantities in explanations (e.g. as the size number increases, the number of people increases by 4)	Considering how one variable changes with another Identifying features of families of functions
RC 3a – Generating representations	Solicited generation of representation Spontaneous generation of representation as part of solution strategy	Teacher or students generating diagrams, tables, graphs, and/or variable equations
RC 3b – Interpreting and connecting representations	Drawing information from a representation to solve a problem Use of one representation to generate another Use of representation in explanation of solution	Explication of global features of representations (e.g. slope of a curve) Making connections between representations Discussing affordances of different representations
RC 4a – Calculations and procedures with precision	Accurate execution of calculations and algebraic procedures (e.g., solving linear equations)	Accurate execution of calculations and algebraic procedures (e.g., solving linear equations)
RC 4b – Checking plausibility of results	Attention to problem context when explaining and justifying solutions	Actively checking solutions with regard to context
RC 5 – Explaining and Justifying reasoning	Responses to questions asking <i>why</i> , <i>how you know your solution is correct</i> , or <i>to explain your solution as if to a classmate</i>	Explicit guidance about what to include in an algebraic explanation Explicit requests for explanation and justification

These robustness criteria are meant to capture the proficiencies students should have to be successful in solving contextual algebraic tasks, but do not necessarily represent a linear sequence of steps. Rather, the RCs represent integrated understandings that can be drawn on when interpreting and mathematizing algebraic situations. The data and analysis below demonstrate how the RCs were used to guide data collection on algebra learning and provide a fine-grained analysis of student work.

Method

The larger study is a mixed methods development project involving assessment, scoring, and observational tools to connect gains in student understanding with observations of classroom teaching practices. The assessments consist of free-response tasks adapted from MARS and multiple-choice items released from the Massachusetts Comprehensive Assessment System (available <http://www.doe.mass.edu/mcas>). These tasks were chosen to represent more familiar and policy-important tasks from standardized tests, as well as tasks providing opportunities for open-ended problem solving. We slightly modified the open-ended tasks to increase opportunities for students to show evidence of each of the robustness criteria. In this section, we briefly describe the data collection, then describe in more detail the validity and reliability of the measures.


Data Collection

We collected data from nine 8th-grade algebra classrooms. The data consist of a pre-test administered in the fall, seven to ten classroom observations, and post-tests administered in early spring. The pre- and post-tests consisted of five multiple choice items and three open-ended MARS tasks. We scored all the tasks using the rubrics based on the robustness criteria and generated class profiles by averaging students' overall and RC scores across each class.

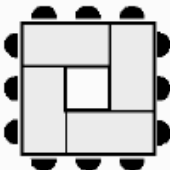
The tasks were scored in two ways: (a) an overall score based on correctness of student responses (similar to the score provided by the MARS rubrics), and (b) individual scores for each RC, providing a more fine-grained view of students' developing understandings. The RC scores represent an accumulation of students' use of particular strategies to solve each subtask, their correct answers to specific parts of the task, and resources drawn on to explain and justify their solutions. Table 2 provides a sample rubric for the linear pattern task shown in Figure 1.

Arranging Tables

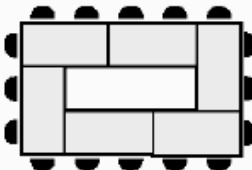
A company supplies tables for business meetings. Each table is a rectangle, and can seat one person on its short edge, and two people on its long edge, like the figures on the right.



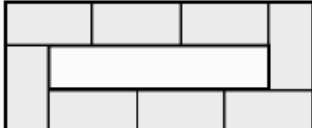
The diagrams below show how these tables can be made into arrangements of people. The different arrangements are numbered, like the figures below. No one sits inside the arrangements of tables.



Size 1
4 tables
12 people



Size 2
6 tables
16 people



Size 3

- (1) How many people can sit at a Size 3 arrangement?
- (2) How many people can sit around a Size 13 arrangement? Explain how you know your answer is correct.
- (3) Write an equation for the number of people p who can sit at a Size S arrangement. Explain how the parts of your equation relate to the table arrangements.
- (4) James is trying to seat 75 people. What size arrangement will he need? Write an explanation that would convince a classmate your answer is correct.

Figure 1. Sample of linear pattern CAT.

Table 2
Rubric for Arranging Tables

Task Subpart	Description (points available)	Associated RCs	Possible Strategies	Associated RCs
Part 1	Find next term in pattern (1)	3b		
Part 2	Find 13 th term in pattern (1)		Generates representation – table, graph, or equation	3a, 4a
			Generates representation - diagram	3a, 4b
			Generates pairs of numbers	4a
	Explains solution (1)	5	Articulates (in words) relationship between size number and number of people	2b
			Draws on representation	3b
			Draws on problem context	4b
Part 3	Write variable equation (2 – partial credit available)	2a, 3a, 3b	Generates a representation	3a
	Explanation (1)	5	Articulates (in words) relationship between size number and number of people	2b
			Draws on representation	3b
			Draws on context	4b
Part 4	Determine an error in pattern generation (1)	1, 2a	Solves equation from Part 3	4a
			Generates table or diagram	3a, 4a
			Generates pairs of numbers	4a
	Explanation	5	Draws on a representation	3b
			References relationship between size and number of people, explains remainder	2b, 4b

Validity

To be valid, our assessments needed to cover appropriate algebra content that accurately captures their competencies along the RC dimensions. Also, we needed scoring rubrics that provided appropriate information on students' robust understanding of algebra. Thus, we completed several stages of validity checks on both the tasks and the rubrics.

Content validity of tasks. To ensure content validity on the open-ended tasks, we largely kept the tasks in their original form since they represent well-established assessments of students' ability to solve non-routine problems, explain their reasoning, and show evidence of high level thinking (<http://www.noycefdn.org/svmi.php>). We made only slight adaptations from the MARS versions to provide more opportunities for students to explain their reasoning and change some language in the contexts to be more appropriate for students in the United States. Pattern tasks and real-world linear tasks were chosen based on pilot studies conducted during the first two years of the study, through which we found that students generally provided more written work on these types of tasks. We also conducted think-aloud interviews in year 1, and post-hoc interviews with randomly selected students in year 2. We found that the interviews did not afford significantly more or different information about students' understanding than their

written work on the assessment, so we considered the tests to be valid measures of how students demonstrated their algebraic understanding on these types of tasks.

Validity of scoring rubrics. To test the validity of our rubrics, which consisted of both an “overall correctness” score as well as scores for individual RCs, we compared the scores acquired using the ATS rubrics to scores obtained using existing MARS rubrics and more holistic scores for twenty student assessments from four classes. Holistic scoring involved analyzing students’ overall performance on a task and making a general judgment about their robust algebraic understanding. We considered this rating to be the “gold standard” for scoring student work because it represented the scorers’ collective understanding of what it meant to show evidence of robust understanding, and sought to compare other scoring techniques with holistic scores. To score holistically, we used a three-point scale (low, medium, high levels of algebraic understanding) to assign scores for students’ algebraic competency for the entire task as well as more fine-grained scores for each RC across two tasks. Results from the ATS and MARS rubrics for the students’ overall scores were moderately correlated ($r = 0.552$ and $r = 0.904$ for the two tasks). Results from the ATS rubrics and overall holistic scores were more highly correlated (r values ranged from 0.694 to 0.764 for two scorers over two tasks).

Following this analysis, we concluded that the ATS rubrics were closely aligned with our holistic perception of student understanding and were a reasonable way to track evidence of students’ possession of specific aspects of robust algebraic understanding. Even so, we made slight revisions to the rubrics with three goals in mind. First, we reconsidered which RCs were being assessed in each part of each task. Secondly, we ensured that the RCs being captured, strategies that may be used, and wording in the rubrics were consistent for matched pairs of tasks on the two forms of the assessments. For example, the Arranging Tables and Hexagons tasks are a matched pair, so we modified their rubrics as needed so that the same strategies were captured, matched to the same RCs, and so on. We also allowed for partial credit as necessary to better align with the MARS rubrics. Finally, we added a holistic scoring rubric on a binary scale for RC 1 at the end of each task. This was because it seemed like the most reasonable way to capture whether students had navigated the language of the task.

After the rubric revisions, we conducted a second round of validity testing similar to that described above where we compared individual RC scores provided through use of ATS rubrics to the results of holistic analysis. The correlations between individual scorers’ holistic and ATS scores were significantly high. These are summarized in Table 3 below.

Table 3
Correlations Between ATS Rubric and Holistic Scores

	Overall	RC 1	RC 2a	RC 2b	RC 3a	RC 3b	RC 4a	RC 5
Scorer 1	0.900	0.640	0.732	0.670	0.765	0.432	0.568	0.776
Scorer 2	0.858	0.712	0.671	0.790	0.702	0.536	0.686	0.851

Again, most correlations were moderately high or high. However, there were differences across the RCs. Certain RCs (e.g., RCs 3b and 4b) were particularly problematic. These problematic correlations led to further revision of the scoring rubrics and refinement of the tasks in order to better reflect these RCs in future phases of development.

Reliability

During the development of our scoring rubrics and scoring guide, we also tested for reliability. Reliability was tested in two ways. First, we selected assessments to determine how well six additional scorers performed against benchmark scores. Second, once all the assessments were scored, we double scored 10% ($n = 12$) of the assessments to see how well the scores from the two raters aligned. The results of this testing along with our validity testing provided feedback on how to revise the scoring guide accordingly. The following sections provide more detail of these processes.

Benchmark scoring and reliability among scorers. For the benchmark scoring, two expert scorers scored and reached full agreement on the scores of twelve tests. These scores served as the benchmark scores used as comparison for the other six scorers' results. Once the assessments were scored, we compared scores using the Intraclass Correlation Coefficient to measure the absolute agreement between each rater and the benchmark (See Table 4). We chose to use the Intraclass Correlation Coefficient because unlike the Pearson coefficient, it accounts for agreements between every score and not just the linear relationship between scores. In most cases, the agreement between each scorer and the benchmarks was acceptable, but RC1 and RC5 had varied results among the scorers, with some correlations below 0.7. These results allowed for further discussions on the clarity of RC1 and RC5. Some modifications were made to the scoring guidelines specific to these robustness criteria.

Table 4

Results of Intraclass Correlation Between Six Scorers and Benchmark Scores

	Intra-class Correlations								
	Overall Score	RC 1	RC 2a	RC 2b	RC 3a	RC 3b	RC 4a	RC 4b	RC 5
Scorer 1	0.986	0.681	0.971	0.938	0.987	0.987	0.959	0.687	0
Scorer 2	0.933	0.501	0.928	0.946	0.891	0.798	0.893	0.927	0.889
Scorer 3	0.966	-0.667	0.967	0.888	0.915	0.976	0.931	0.796	0.894
Scorer 4	0.922	0.727	0.932	0.783	0.790	0.907	0.739	0.934	0.615
Scorer 5	0.953	0.793	0.902	0.879	0.921	0.924	0.906	0.807	0.634
Scorer 6	0.979	1	0.957	0.903	0.956	0.979	0.947	0.941	-0.296

Note. Values above 0.72 are considered acceptable.

Double-scoring results. Eighteen tests (nine from each form of the assessment) were randomly selected and assigned to eight raters. Two scorers independently scored each test. Then, we randomly assigned the two scores for each test as the scores of Rater 1 and Rater 2 score respectively, and calculated the Pearson correlation between the two representative scorers (shown in Table 5) for the correctness score, each RC score, and the overall score. All the Pearson Correlation Coefficients are calculated at one-tail to achieve higher precision on the basis that we have a priori knowledge that the scores would be positively correlated.

Table 5

Correlations Between Scorer 1 and Scorer 2

	Total Correct	RC1	RC2a	RC2b	RC3a	RC3b	RC4a	RC4b	RC5	Overall
Form A	0.997	0.701	0.975	0.972	0.608	0.970	0.909	0.955	0.877	0.978

Form B	0.924	0.783	0.885	0.944	0.891	0.920	0.800	0.496	0.326	0.822
--------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------

The correlations for RC 3a on Form A and RCs 4b and 5 on Form B were moderately low, which indicate inconsistency between raters. The low correlation coefficients for both RC 4b (0.496) and RC 5 (0.326) on Form B were mainly caused by large discrepancies between two raters' scoring of a single students' test. Overall, we concluded that the scoring rubrics are acceptably reliable and that these issues with the problematic test were minor. First, this kind of discrepancy only occurred for one of the sample tests. We decided to annotate the problematic test as a boundary example for future reference. Second, we assigned the scorers randomly, evenly distributing the scoring of all the tests from each class among the scorers. Also, we analyzed results based on class-level averages, not by selecting individual assessments for analysis like we did in this reliability checking. Thus, the reliability of the results will be substantially higher than in the reliability tests. For instance, correlation of the overall scores between the two raters on both forms, as seen in Table 5, indicates high reliability for overall assessment scores. For this report, we share illustrative results from four representative classrooms.

Results

In this study, we sought to identify components of robust algebraic understanding (RCs), describe the tools we used to capture students' possession of these competencies and measure growth, and consider class profiles resulting from analysis of student learning associated with each RC. It is not our intent to present conclusive evidence of student learning in participating classrooms; relatively low n -values prevent statistical significance of these results. Rather, in this section we provide results from students' pre- and post-assessments, providing summative results as well as how classrooms performed by each RC. Our purpose is to illustrate the kind of results possible with these instruments. In particular, we emphasize how students can make greater gains in some RCs than others, which will ultimately allow for more detailed and specific links to classroom instruction.

Summative Scores by Class

Summative results include changes in students' overall scores and multiple choice and open-ended task subtotals. Table 6 provides class average scores on the assessments for each measure. All classes showed increases in multiple choice scores, but differences exist in patterns of change. Students in Class A improved most from their pre-assessment scores on the multiple choice items. This trend did not carry through to the summative scores on the open-ended items, however, where the average student score showed only slight improvement. Conversely, students in Class C had the greatest improvement (10% increase in scores) on open-ended items, while students in Class B showed a slight decline. Figure 2 highlights these changes. To find change in percents, we took the pre-test and post-test averages, then found the difference.

Table 6
Class Average Overall and RC Scores

	Class A		Class B		Class C		Class D	
	Pre	Post	Pre	Post	Pre	Post	Pre	Post
Overall Correct	0.228	0.263	0.324	0.336	0.236	0.332	0.224	0.273

Mult Choice	0.336	0.509	0.541	0.588	0.482	0.565	0.369	0.446
Open-ended	0.136	0.164	0.239	0.235	0.140	0.239	0.168	0.205
RC 1	0.439	0.485	0.490	0.569	0.392	0.549	0.462	0.500
RC 2a	0.148	0.195	0.231	0.295	0.125	0.242	0.163	0.162
RC 2b	0.166	0.190	0.175	0.195	0.144	0.247	0.148	0.223
RC 3a	0.084	0.134	0.146	0.178	0.137	0.216	0.083	0.129
RC 3b	0.220	0.252	0.286	0.274	0.195	0.289	0.196	0.246
RC 4a	0.075	0.101	0.147	0.165	0.137	0.186	0.115	0.132
RC 4b	0.107	0.120	0.174	0.136	0.102	0.179	0.112	0.143
RC 5	0.115	0.130	0.199	0.196	0.171	0.236	0.115	0.220

These scores provide some insights into student growth (or decline) in performance on contextual algebraic tasks. Yet, we desired a more fine-grained picture of student understanding: did students show greater gains in certain RCs, which would allow for links to and focused analysis of instruction in their classrooms?

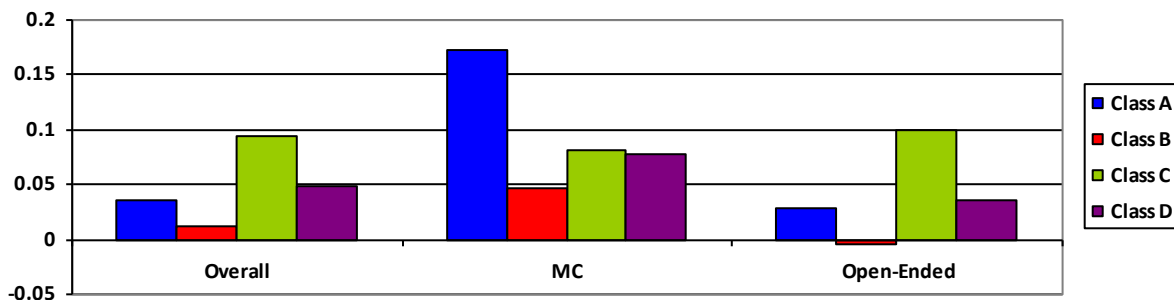


Figure 2. Comparison of the change in percent of summative scores for each class.

RC Scores by Class

To better understand changes in summative scores and provide more fine-grained class-level profiles of student understanding, we analyzed how scores changed for each RC by calculating the differences between pre- and post-test scores along each RC dimension (see Figure 3).

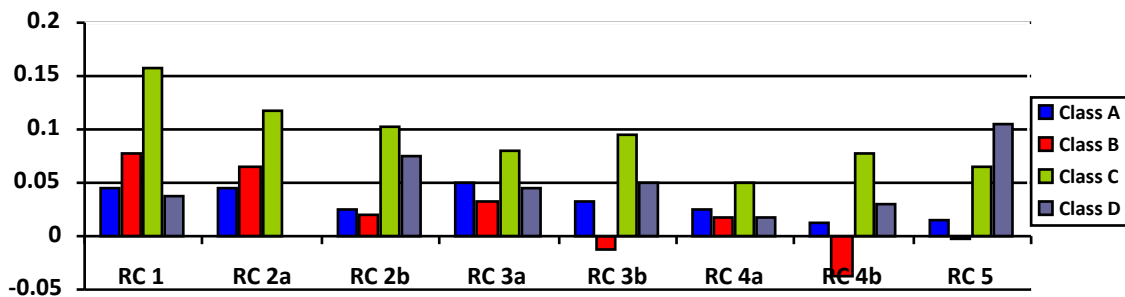


Figure 3. Comparison of the change in percent of RC scores by class.

This level of analysis provided a more detailed look at student learning, though in this sample gains in individual RCs were small. We can see, for example, that the increase in open-ended task scores for Class C is distributed across many of the RCs. In particular, students’ performance in identifying relevant quantities and the relationship between them (RC2),

construct, interpret and connect algebraic models (RC3), and check solutions in the context of the problem (RC 4b) showed substantially growth than in other classes.

Although it is useful to separate the RCs for analytical purposes, the RCs likely develop simultaneously in practice. In particular, the 15% increase in RC1 is likely related to the growth in RCs 2 and 3. This is because building situation models of problems requires interpreting language (RC1), identifying the relationships between quantities (RC2), and representing them mathematically (RC3).

Determining overall changes in student performance on the assessments, especially in open-ended tasks, provides some information about strengths and weaknesses in student understanding. Nevertheless, examining areas of growth at a finer level provides a detailed profile of students' robust understanding at the classroom level. The development of our rubrics allowed us to gather this detailed information efficiently and to identify what resources students are drawing from to solve the tasks. In turn, this may allow us to make stronger links to instructional moves that promote students' proficiency at solving open-ended contextual tasks.

Scholarly Significance

This method of analysis, in the context of a larger study connecting classroom practices with student performance, allows us to unpack the development of student understanding at a relatively fine grain size. Frequently, large-scale studies measure changes in student understanding by computing the average number of correct responses. Our work measures changes in students' strategies that demonstrate underlying algebraic competencies. By disaggregating student work by RCs, rather than a more typical focus on problem type (e.g. systems of equations in two variables), this analysis provides class-level profiles of change in students' use of algebraic representations, their interpretation of problems, and the quality of mathematical justification in their explanations. This approach will indicate which aspects of student understanding are developing, rather than which types of problems they can solve, and can point to instructional practices that may foster students' development of those underlying competencies.

Our analyses may have a larger implication for research. For instance, as many states adopt the Common Core State Standards in Mathematics, student success will be measured by students' demonstration of mathematical practices in addition to their content-specific skills. Research on the teaching practices that support the development of these practices, closely aligned with the underlying proficiencies represented by the robustness criteria, is essential to support teachers in enabling all students to meet these standards (Shaughnessy, 2011). New assessments will include open-ended tasks similar to the tasks used in this study (see, e.g., Smarter Balanced Assessment, 2012). Our technique for analysis of student performance on these tasks using the robustness criteria provides detailed information on what aspects of algebraic understanding students are developing in problem solving and where they continue to struggle.

References

- Ayieko, R., Floden, R. E., Hu, S., Lepak, J., Reinholz, D. L., & Wernet, J. (2012). *Transitioning from executing procedures to robust understanding of algebra*. Paper presented at the annual meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education, Kalamazoo, MI.
- Brenner, M. E., Mayer, R. E., Moseley, B., Brar, T., Duran, R., Reed, B. S., et al. (1997). Learning by understanding: The role of multiple representations in learning algebra. *American Educational Research Journal*, 34(4), 663-689.
- Chazan, D. (2000). *Beyond formulas in mathematics and teaching: Dynamics of the high school algebra classroom*. New York, NY: Teachers College Press.
- Common Core State Standards for Mathematics. (2010). Retrieved November 10, 2010 from http://www.corestandards.org/assets/CCSSI_Math%20Standards.pdf
- Driscoll, M. (1999). *Fostering algebraic thinking: A guide for teachers grades 6-10*. Portsmouth, NH: Heinemann.
- Graham, K., Cuoco, A., & Zimmerman, G. (2010). *Focus in high school mathematics: Reasoning and sense making in algebra*. Reston, VA: NCTM.
- Kieran, C. (2007). Learning and teaching of algebra at the middle school through college levels: Building meanings for symbols and their manipulations. In J. F. K. Lester (Ed.), *Second Handbook of Research on Mathematics Teaching and Learning* (pp. 707-762). Charlotte, NC: Information Age Publishing.
- NCTM (2000). *Principles and Standards for School Mathematics* (Vol. 2000). Reston, VA: National Council of Teachers of Mathematics.
- Schoenfeld, A. H. (2004). Beyond the purely cognitive: belief systems, social cognitions, and metacognitions as driving forces in intellectual performance *Classics in Mathematics Education Research* (pp. 329-363). Reston, VA: National Council of Teachers of Mathematics.
- Shaughnessy, M. (2011). *CCSSM and Curriculum and Assessment: NOT Business as Usual*. NCTM Summing up. Retrived from <http://www.nctm.org/about/content.aspx?id=30009>, July 2011.
- Smarter Balanced Assessment Consortium (2012). Content Specifications for the Summative assessment of the *Common Core State Standards for Mathematics*. See <http://www.smarterbalanced.org/smarter-balanced-assessments/>.
- Usiskin, Z. (1988). Conceptions of school algebra and uses of variables. In A. Coxford & A. Shulte (Eds.) *The Ideas of Algebra, K-12. 1988 Yearbook* (pp. 8-19). Reston, VA: NCTM.
- Yackel, E. (2001). Explanation, justification, and argumentation in mathematics. Presented at the PME 25, Utrecht, Holland, 9-24.
- Walkington, C., Sherman, M., & Petrosino, A. (2012). 'Playing the game' of story problems: Coordinating situation-based reasoning with algebraic representation. *Journal of Mathematical Behavior*, 31, 174-105.