PROJECT AAIMS: ALGEBRA ASSESSMENT AND INSTRUCTION – MEETING STANDARDS

Reliability and Criterion Validity of
Two Algebra Measures:
Translations and Content Analysis-Multiple Choice

Technical Report #6

Anne Foegen, Ph.D.
Jeannette Olson, M.S.

Iowa State University

August 2005

Project AAIMS is funded by the U.S. Department of Education,
Office of Special Education Programs, Grant # H324C03006

PROJECT AAIMS: ALGEBRA ASSESSMENT AND INSTRUCTION – MEETING STANDARDS

Reliability and Criterion Validity of Two Potential Algebra Measures: Translations and Content Analysis-Multiple Choice

Technical Report 6

Abstract

This technical report summarizes the results of a study in which we examined the technical adequacy of two potential measures for algebra progress monitoring. Eighty-seven students (11 of whom were receiving special education services) completed two forms of a *Translations* measure and two forms of a *Content Analysis-Multiple Choice* measure during each of two data collection sessions. In addition, we gathered data on criterion variables including grades, overall grade point average, teacher ratings of student proficiency, and scores on district-administered standardized tests, as well as a measure of algebra aptitude. We examined both test-retest and alternate form reliability for both single probe scores and aggregated scores (computed by averaging two individual scores). Criterion validity was examined by computing correlations between students' single and aggregated scores on the probes with their scores on other indicators of proficiency in algebra.

The results of this study suggest that the *Translations* measure is more promising than the *Content Analysis-Multiple Choice* measure in terms of both reliability and criterion validity. The strength of the relations obtained in this study were in the low to moderate range and were not as strong as the relations obtained with a different sample in this district using three other algebra measures (see Project AAIMS Technical Report 2 for details of the earlier study).

Both measures produced acceptable distributions that were free from floor and ceiling effects. Students had roughly similar means and standard deviations on both measures. Reliability estimates for both measures fell short of expected levels for both single probes and aggregated scores. The *Translations* measure produced stronger correlations than the *Content Analysis-Multiple Choice* measure, but did not demonstrate a level of reliability that would be acceptable for instructional decision making. The majority of the criterion validity relations were in the low to moderate range. Aggregated scores produced improvements in the criterion validity estimates for the *Translations* measure, but not for the *Content Analysis-Multiple Choice* measure. The strongest relations were identified between the *Translations* measure and eighth graders' performance on the district's math achievement test, as well as between the *Translations* measure and all students' performance on the algebra aptitude test. These two relations were in the moderate to strong range; relations between the *Translations* measure and the remaining criterion variables were in the low range.

Full Report

Introduction

Algebra often functions in the role of a 'gatekeeper,' with proficiency in algebra having significant influence on individuals' access to higher education and professional career paths. If students with disabilities are to have access to these opportunities, it is critical that they develop proficiency in algebra. Robert Moses, a mathematics educator and civil rights advocate, sees algebra as the 'civil right' of the 21st century. He argues that algebra proficiency provides the same access to economic and social equity that the right to vote represented during the Civil Rights movement of the 1960s (Moses & Cobb, 2002). Project AAIMS (Algebra Assessment and Instruction—Meeting Standards) strives to improve student learning in algebra for all students, including those with and without disabilities. Project AAIMS has two primary objectives. First, we will examine the alignment between algebra curriculum, instruction, and assessment for students with and without disabilities. Second, we will develop and validate progress monitoring tools to support teachers' instructional decision making relative to student learning in algebra. In Technical Report 2 (Foegen & Lind, 2004), we reported the reliability and criterion validity of three measures developed as potential indicators of student proficiency in algebra. In this report, we describe a study in the same district with two additional potential measures of algebra proficiency.

Purpose

The purpose of this study was to examine the technical adequacy of two newly developed algebra progress monitoring measures. In particular, we planned to address the following research questions:

- To what extent do the distributions, means, and standard deviations produced by the measures reflect a normal distribution of scores and an absence of floor and ceiling effects?
- What levels of test-retest and alternate form reliability do these measures demonstrate? Does aggregating students' scores increase the level of reliability?
- What levels of criterion validity do the measures demonstrate? Are there variations across different types of criterion measures? Do the criterion validity levels improve if students' scores are aggregated across multiple probes?
- To what extent do the measures differentiate across different student performance groups?

Method

The study described in this report was conducted in October 2004 in District A. This district serves four small towns as well as the rural agricultural areas between the towns. Approximately 7,000 residents reside in the school district. The junior/senior high school has an enrollment of approximately 600 students; about 12 percent of these students receive special education services. Approximately 13 percent of the district's students are eligible for free and reduced lunch; three percent are of diverse backgrounds in terms of race, culture and ethnicity. Data for the study were gathered on three consecutive Tuesdays in October 2004. The first two weeks, students completed. The algebra aptitude measure was administered on the third

Tuesday. All data collection activities involving students were completed during regular class time. Project AAIMS staff administered all measures.

Participants

Eighty-seven students in District A participated in the study. Written parental/guardian consent and written student assent were obtained for all of these students using procedures approved by Iowa State University's Human Subjects Review Committee. A description of the participating students is provided in Table 1.

Table 1. Demographic Characteristics of Student Participants by Grade Level

	Total	Grade 8	Grade 9	Grade 10	Grade 11
N	87	10	60	15	2
Gender					
Male	44	4	29	10	1
Female	43	6	31	5	1
Ethnicity					
White	85	10	58	15	2
Black	1	0	1	0	0
Hispanic	1	0	1	0	0
Lunch					
Free/Red	8	0	5	3	0
Disability					
IEP	11	0	8	2	1

As the data in Table 1 indicate, the vast majority of the participants (98%) were white and 69% were in ninth grade, the traditional grade in which students in District A complete algebra. Nine percent participated in federal free and reduced lunch programs and 12.6% were students with disabilities who were receiving special education services. Ten of the students were advanced eighth graders who were enrolled in a high school level Algebra 1 course that included high school students in grades nine and ten. Four students (all of whom had disabilities) were enrolled in a pre-algebra course taught by a special education teacher. Of the remaining students, 60 were participating in traditional Algebra I courses, and 14 were enrolled in Pre-Algebra (a course in which the first half of traditional Algebra 1 content is taught over the course of an entire academic year).

Additional Information on Students with Disabilities. Because the applicability of the algebra probes to students with disabilities is an important part of Project AAIMS, additional information about the eleven students with disabilities participating in the project is provided in Table 2.

Table 2. Descriptive Information on the Programs of Students with Disabilities (N=11)

Characteristic	Quantification
Disability category	100% Entitled Individual (EI)
% time in general education	Range = 47 –95%; Mean = 73% 55% of students spend more than 75% of their instructional time in general education
# of students with math goals	8
# of students receiving math instruction in general education classes	7
# of students receiving math instruction in a special education setting	4
# of students receiving English instruction in a special education setting	2
# of students receiving social studies instruction in a special education setting	3
# of students receiving science instruction in a special education setting	2
# of students receiving health instruction in a special education setting	2
# of students with one period of resource study hall daily	10
# of students with two periods of resource study hall daily	1
# of students with goal code D2: *Is responsible for self*	2
# of students with goal code F2C: *Comprehension*	6
# of students with goal code F2F*: Fluency*	1
# of students with goal code F3A*: Applied math*	6
# of students with goal code F3C*: Computation*	3
# of students with goal code F4M: *Mechanics of writing: punctuation, grammar, spelling*	4

Students with disabilities earned a mean GPA for the Fall 2004 semester of 2.20 (range 1.47 – 3.67). In algebra, students with disabilities earned mean grades of 2.33 [C+] (range 1.0 [D] to 4.0 [A]). Standardized test data for these students reflect their challenges with academic content; in District A, the Iowa Tests of Educational Development are used as a district-wide assessment. On average, students with disabilities obtained national percentile rank scores of 23 and 30 in Concepts/Problem Solving, and Computation, respectively. They obtained a mean percentile rank of 30 on the Reading Total scale.

Measures

Two groups of measures were used in this study. The first group consisted of the curriculum-based measures of algebra performance developed by the Project AAIMS research team. The second group consisted of the measures that served as criterion indicators of students' proficiency in algebra. Each group of measures is described below.

Algebra Progress Monitoring Measures. Two algebra measures were examined in this study; sample copies of each are provided in the appendices. The first, which we refer to as the ***Translations*** probe, was designed to assess the students' proficiency in recognizing translations between multiple representations of the relationships between two sets of numbers. In creating this probe, we drew from curriculum materials for teaching algebra concepts at the middle school level created as part of the *Connected Mathematics* curriculum materials (Lappan, Fey, Fitzgerald, Friel, Phillips, 2004). In this curriculum, students explore the connections between numerical relationships in multiple formats. For example, they might examine how changing elements of an equation (i.e., changing $y = 2x$ to $y = 2x + 3$) influences the graphic representation of the equation. Likewise, they examine relationships between data tables, graphs, and equations. Contextualized problems representing real life situations are also used as a basis for exploring algebraic relationships. In our *Translations* probe, we assessed whether students could recognize the same relationship between two sets of numbers presented in four different formats. At the top of the first page, students were given four 'base' graphs (on the second page, equations were used as the stimulus and on the third, data tables). Below these four prompts (labeled A through D), students were presented with rows of alternative representations of the same relationships. One row contained equations, another data tables, and a third, story scenarios. The students' task was to identify matches between the four prompts at the top of the page and the same relationships represented in another format in each of the following three rows. Copies of the two *Translations* probes are presented in Appendix A.

The *Translations* probe was created in response to feedback from the Project AAIMS Advisory Committee during a review of the initial three algebra probes. The Advisory Committee noted that the initial three probes focused heavily on algebraic manipulations and procedures, and urged the AAIMS research staff to pursue the development of a task that allowed students to demonstrate conceptual understandings of algebraic topics without requiring procedural accuracy with manipulations of algebraic symbols. In order to fit with the design constraints for progress monitoring tasks (i.e., brief, easy to administer and score), we selected a multiple choice format for the task. We created two parallel forms of the *Translations* probes. Each probe consisted of 43 items; we scored the probes by counting the number of correct and incorrect responses. Because of the multiple choice format, we were concerned that scores might be artificially inflated by guessing. Previous work by Foegen (2000) has demonstrated that applying a correction formula for guessing increases the reliability and criterion validity of the scores. We incorporated alternative scoring procedures into our research design.

The second algebra progress monitoring measure that we developed was the ***Content Analysis-Multiple Choice*** measure. This measure was a variation of the *Content Analysis* probe examined in the initial study. The original *Content Analysis* probe (which we now refer to as the *Content Analysis-Constructed Response* probe), was created by analyzing the content taught in the algebra textbook. Because all three districts participating in Project AAIMS are using the same textbook series, we wanted to investigate a measure that was directly derived from the instructional materials. We developed the items by sampling from the chapter tests and reviews. We sought to identify items that represented core concepts/problem types in each chapter. Based on teacher feedback, we sampled chapters in the middle portion of the text at a higher rate (two questions per chapter) than the chapters at the beginning (review) and end (advanced concepts/skills) sections of the text. We anticipated that this probe might provide a more direct reflection of the extent to which students had learned the content of instruction than would the other probes, which represented more general indicators of algebra proficiency.

The original *Content Analysis- Constructed Response* probes consisted of 16 items, each worth from one to six points, depending on the complexity of the problem. Students worked on the probe for ten minutes. The probes were scored by awarding points corresponding to any of the steps on the key that they completed correctly in their responses. In the directions for this probe, we encouraged students to show their work if necessary to obtain 'partial credit' even if they weren't able to solve the entire problem. We also informed them that if they were able to complete the problems without showing all the steps, they would be awarded the full number of points possible for the correct solution. We opted to use this practice in order to reinforce/reward students who were so proficient that it would be tedious for them to record each step of the problem.

For this study, we revised to original *Content Analysis-Constructed Response* probe by creating four multiple-choice alternatives for each problem. Our rationale for going to a multiple choice option was that this format would improve scoring efficiency (and potentially interscorer agreement), that it might reduce the difficulty of the task (on the open ended version of the probe, we obtained significant floor effects, even when the probe was administered at the end of a year of instruction), and that the multiple choice format was one with which students needed to be proficient for district-administered assessments. We reduced the amount of time available for students to work on the probe from 10 minutes to 7; in the first study, we found that many students had stopped working on the task within 5 minutes. Students were encouraged to show their work in order to earn partial credit even if they were not able to completely solve a problem. In addition, students were advised NOT to make wild guesses, as these would result in deductions from their total scores. The two *Content Analysis-Multiple Choice* probes used in the study are presented in Appendix B.

Scoring for the *Content Analysis-Multiple Choice* probes was done by comparing student responses to a rubric-based key created by the research staff. Each of the 16 problems was worth up to three points. Students earned full credit (three points) by circling the correct answer from among the four alternatives. If students circled an incorrect response and did not show any work, their answer was considered a 'guess' and counted as part of the final score assigned to each probe. In cases where students showed work, the scorer compared the student's work to the rubric-based key, and determined whether the student had earned 0, 1, or 2 points of partial credit. A student's final score on the probe consisted of the number of points earned across all 16 problems. The number of guesses was also recorded and entered in the data files.

<u>Criterion Measures</u>. In order to evaluate the criterion validity of the algebra progress monitoring measures, we gathered data on a variety of other indicators of students' proficiency in algebra. Some of these measures were based on students' performance in class (and in school more generally) and their teachers' evaluation of their proficiency. Other measures reflected students' performance on standardized assessment instruments.

The classroom-based measures included grade-based measures and teacher ratings. Each student's *algebra grade*, the grade s/he earned in algebra during the fall semester of the 2004-05 school year, was recorded using a four-point scale (i.e., A = 4.0, B = 3.0). *GPA* represented students' overall grade point average for the 2004 fall semester year and was recorded using the same four-point scale, with scores rounded to the nearest hundredth. We also wanted to include the teachers' evaluations of students' proficiency in algebra. To accomplish this, we asked each teacher to complete a *teacher rating* form for all the students to whom s/he taught algebra. Student names were alphabetized across classes to minimize any biases that might be associated with particular class sections. Teachers used a 5-point Likert scale (1=low proficiency, 5= high proficiency) to rate each student's proficiency in algebra in comparison to same-grade peers. A copy of the teacher rating form is presented in Appendix C.

Student performance on standardized, norm-referenced assessments was evaluated using school records and with an algebra instrument administered as part of the project. In District A, 8th grade students complete the *Iowa Tests of Basic Skills* (ITBS) each spring. Students in grades 9 through 11 complete the *Iowa Tests of Educational Development* (ITED), also in the spring. District records were used to access students' scores on these instruments; national percentile ranks were used for the analyses. For the ITBS, the following scores were recorded: Problems/Data, Concepts/Estimation, Computation, Math Total, Reading Total. For the ITED, we recorded the Concepts/Problems score (which was identical to the Math Total score), the Computation score, and the Reading Total score. Because these tests were completed in the spring, we were able to evaluate the predictive validity of the algebra probes, which were administered in the fall.

Neither of the district-administered measures provided a direct assessment of algebra, so we also administered the *Iowa Algebra Aptitude Test* (IAAT). This norm-referenced instrument is typically used to evaluate the potential of 7th grade students for successful study of algebra in 8th grade. Although we recognized the limitations of using this aptitude measure, we were unable to identify a norm-referenced test of algebra achievement. We had some concerns that there might be ceiling effects when using this measure, but these concerns proved to be unwarranted.

Procedures

The algebra probes were administered in a single 45-minute class period. During each class, students completed two parallel forms of the *Translations* probe and two parallel forms of the *Algebra Concepts-Multiple Choice* probe. The order in which the two types of probes were administered was counterbalanced across classes, as was the order of each of the parallel forms. Students completed the tasks in the same order both weeks. A copy of the standardized directions used for each administration session is provided in Appendix D. Table 3 depicts the order in which the probes were administered during each of the two testing sessions.

Table 3. Administration Schedule for Probe Forms by Period

Session	Algebra 1 (Per. 2)	Algebra 1 (Per. 3)	Algebra 1 (Per. 5)	Pre-Algebra (Per. 6)	Algebra1 (Per. 7)	SpEd Pre-Algebra
1 and 2	D1	E1	D2	E2	E1	D1
	D2	E2	D1	E1	E2	D2
	E2	D2	E1	D2	D1	E1
	E1	D1	E2	D1	D2	E2

D1, D2 = *Translations* probes 1 and 2
E1, E2 = *Algebra Concepts-Multiple Choice* probes 1 and 2

Results

Scoring Reliability

Scoring accuracy was evaluated by re-scoring approximately one-third of the probes. For each probe, an answer-by-answer comparison was conducted and an interscorer reliability estimate was calculated by dividing the number of agreements by the total number of answers scored. These individual probe agreement percentages were then averaged across all the selected probes of a common type to determine an overall average.

We selected the probes to be re-scored by drawing from each of the class periods across the two administration periods. The special education class was omitted because of small student numbers (4 students in the class). Each form of the probes was rescored for 2 of the 6 class periods (33%). The number of student papers rescored and the average agreement for each form of the probe are reported in Table 4.

Table 4. Interscorer Agreement Rates and Student Papers Rescored

Probe	# Papers Rescored	Range of Agreement	Mean % Agreement
Translations, Form 1	107	83 – 100%	98.9%
Translations, Form 2	68	74 – 100%	98.3%
Content Analysis-Multiple Choice, Form 1	92	60 – 100%	94.3%
Content Analysis-Multiple Choice, Form 2	85	50 – 100%	91.3%

The *Translations* probes were scored with high levels of accuracy. The *Content Analysis-Multiple Choice* probes were clearly more difficult to score consistently, although the scoring accuracy for both forms exceeded the minimum level for acceptable agreement that we had established at 90%. Although the levels of interscorer agreement for the Multiple Choice format of the *Content Analysis* probe are comparable or higher than the Constructed Response agreement levels reported in Technical Report 2 (88% - 91%), we plan to continue to refine our scoring rubrics to pursue higher levels of agreement. In reviewing individual papers where the agreement level was less than 75%, virtually all cases involved student papers in which a small number of problems were completed. In these situations, a single error is magnified (i.e., 1 disagreement on a paper with 3 responses produces an agreement estimate of 67%, while 1 disagreement on a paper with 10 responses produces an agreement estimate of 90%). We will continue to strive to increase the reliability with which multiple individuals can score the *Content Analysis-Multiple Choice* probes.

<u>Descriptive Data on Score Ranges and Distributions</u>

Table 5 lists the ranges, means, and standard deviations for each of the probes. On the *Translations* probe, the Correct score represents the number of correct matches, while the Incorrect score represents the number of incorrect responses. The total possible for the *Translations* probe was 43 points. On the *Content Analysis-Multiple Choice* probes, the Correct score represents the number of points earned on the probe (each of the 16 problems was worth up to 3 points, for a maximum score of 48) and the Incorrect score represents the number of incorrect responses.

Results for the *Translations* probes reveal that students' Incorrect scores exceeded those of their Correct scores on both forms of the probes during both weeks. In addition, standard deviations for the Incorrect scores were more than double the standard deviation for the Correct scores in each instance. These findings raise great concern about the extent to which students understood the task and completed it to the best of their ability, rather than making random

Table 5. Descriptive Data for Algebra Probes Across Administration Sessions – Raw Scores

Measure	Session/ Week	N	Score	Range	Mean	Standard Deviation
Translations Form 1	1	77	Correct	0 – 21	8.82	4.53
		77	Incorrect	0 – 34	10.32	9.86
	2	77	Correct	0 – 26	10.43	4.85
		77	Incorrect	0 – 39	15.47	12.65
Translations Form 2	1	74	Correct	0 – 24	8.92	4.48
		74	Incorrect	0 – 36	10.12	10.24
	2	77	Correct	2 – 26	10.84	5.10
		77	Incorrect	0 – 37	14.17	12.66
Content Analysis-Multiple Choice Form 1	1	81	Correct	0 – 26	14.32	5.23
		81	Incorrect	0 – 15	4.04	3.17
	2	80	Correct	3 – 26	14.30	4.65
		80	Incorrect	0 – 13	4.36	3.96
Content Analysis-Multiple Choice Form 2	1	80	Correct	1 – 27	12.39	5.50
		80	Incorrect	0 – 14	3.35	3.47
	2	80	Correct	3 – 24	11.33	4.42
		80	Incorrect	0 – 14	3.68	4.52

guesses for their responses. It should be noted that students were not explicitly instructed NOT to guess on this task, so many students might have opted to provide a response for all problems. In subsequent sections, we examine ways in which corrections for guessing might be applied to counter this issue.

Results for the *Content Analysis-Multiple Choice* probe indicate that the probe has a reasonable level of difficulty (serious floor effects were not evident, even though students were only approximately 8 to 10 weeks into the academic year). The average number of incorrect problems was less than five, indicating guessing was less of an issue on this probe than for the same students on the *Translations* probe. We found it encouraging that the floor effect issues identified with the *Content Analysis-Constructed Response* probes in Technical Report 2 were not evident in these data.

Reliability of Individual Probe Scores

The reliability of individual probes was evaluated by examining alternate form reliability (the Pearson product moment correlation between the two forms of each type of probe given during the same data collection session) and test-retest reliability (the Pearson correlation between the same form of each probe given across the two data collection sessions). We compared the effects of three different scoring procedures on the reliability of students' scores on the probes. The first scoring method involved using the total points earned on the probe (i.e., the values listed in Table 5 as 'Correct'). Findings for this scoring method are listed under the column titled Correct in Table 6. The second method (listed in the column titled C – I in Table 6) involved subtracting the number of incorrect problems (the 'Incorrect Value' in Table 5) from each student's total Correct points. The third method (labeled '1/3' in Table 6) involved subtracting one third of the number of incorrect problems from the total points earned on each probe. This procedure to correct for guessing has been used in previous research involving multiple choice mathematics probes and was found to be effective in increasing the reliability and validity of the scores (Foegen, 2000). In circumstances where the scoring procedure produced a negative value, the student's score was set to 0. This occurred more frequently with the *Translations* probes than with the *Content Analysis-Multiple Choice* probes.

Table 6: Reliability results for single probes

Probe Type		Alternate Forms				Test-Retest		
Translations		Correct	C - I	1/3		Correct	C - I	1/3
First session		.53	.60	.51	Form 1	.50	.72	.67
Second session		.48	.62	.62	Form 2	.38	.58	.41
Content Analysis-Multiple Choice								
First session		.62	.45	.56	Form 1	.42	.50	.42
Second session		ns	.42	ns	Form 2	.24	ns	.ns

Note: All correlations significant at $p < .05$.

The results in Table 6 indicate that the scoring method that produced the most reliable *Translations* scores was the Correct minus Incorrect process. In all four instances, these correlations matched or exceeded those for the other two scoring methods. For the *Content Analysis-Multiple Choice* probes, the results were mixed. No single method consistently outperformed the others, and for two of the three methods, the correlations were non-significant. This result was surprising to us, as we anticipated that the higher levels of student guessing associated with the *Translations* probe would result in reliability estimates lower than those for the *Content Analysis-Multiple Choice* probes.

Neither probe consistently met the desired level of .80 that is traditionally used as a benchmark for reliability for screening measures. In the Discussion section, we consider possible reasons for the unreliability of student scores and offer suggestions for modifying the probes to increase the reliability levels.

Reliability of Aggregated Probe Scores

Because students completed two forms of each probe during each data collection session, it was also possible to examine the effects of aggregating scores from two probes on the resulting reliability levels. Previous research in other areas of mathematics (Foegen, 2000; Fuchs, Deno, & Marston, 1983) has determined that for some types of mathematics skills and concepts, multiple probes need to be aggregated to obtain reliable scores for individual students. Table 7 presents the results for the aggregated scores on probes. The alternate form coefficients were computed by correlating the average of the scores from the two administrations of Form 1with the average of the scores obtained in the two administrations for Form 2. The test-retest coefficients were computed by averaging scores from the two forms of each probe administered on the first data collection day, and then correlating these scores with the averaged scores for the same probes from the second data collection day.

Table 7. Reliability for Aggregated *Translations* and *Content Analysis-Multiple Choice* Probes

Probe	Alternate Form Reliability	Test-Retest Reliability
Translations		
Correct	.63	.57
C - I	.71	.77
1/3	.70	.69
Content Analysis-Multiple Choice		
Correct	.56	.49
C - I	.57	.51
1/3	.52	.45

Note: All correlations significant at $p < .05$.

The results in Table 7 indicate that aggregation of two probe scores did produce substantial improvements in the reliability of the *Translations* probe for all three scoring procedures. Unfortunately, the reliability levels still fall short of conventional expectations for assessment tools. Aggregation did not increase the reliability of scores on the *Content Analysis-Multiple Choice* probes. Future research is needed to explore changes in task format and presentation to increase the reliability of both measures.

Criterion Validity for Single Probes

The criterion validity of the measures was examined by correlating scores on the probes with the criterion measures that served as additional indicators of students' proficiency in algebra. The indicators we used included students' overall grade point average (GPA) and grades in algebra; teachers' evaluations of student proficiency; scores from standardized tests in mathematics administered by the district; and scores obtained from a norm-referenced test of algebra aptitude, the Iowa Algebra Aptitude Test (IAAT). In the following section, the correlation coefficients between scores on the algebra measures and each of these criterion variables are presented and discussed. Correlation coefficients are presented in Table 8, with results included for each of the three scoring methods. Because four correlation coefficients were produced in the analyses (scores from each of two forms of probe were available for each of the two administration days), mean correlations are reported. The range of obtained correlations is included in parentheses. If at least two of the four correlations were significant, the mean correlation is reported.

Table 8. Criterion Validity Results for Single Probes: Mean Correlation Coefficients and Ranges

Criterion Measure	*Translations*			*Content Analysis-Multiple Choice*		
	Correct	C - I	1/3	Correct	C - I	1/3
Overall GPA	.28 (2 NS[a], .27 - .29)	.43 (.38 - .48)	.41 (.28 - .46)	.35 (2 NS, .30 - .40)	.39 (.29 - .50)	.43 (2 NS, .37 - .48)
Grade in Algebra	.28 (2 NS, .27 - .28)	.38 (.34 - .41)	.37 (.28 - .44)	.36 (2 NS, .28 - .44)	.38 (.30 - .53)	.36 (1 NS, .26 - .50)
Teacher Rating	.34 (2 NS, .33 - .37)	.48 (.39 - .57)	.45 (.38 - .54)	NS (3 NS; .34)	.38 (.32 - .48)	.31 (1 NS, .25 - .42)
ITBS Scores[b]						
Math Total	NS (3 NS; .81)	NS (3 NS; .77)	NS (3 NS; .74)	NS	NS	NS
Prob/Data	NS (3 NS; .80)	NS (3 NS; .78)	NS (3 NS; .75)	.73 (2 NS, .75 - .72)	.71 (2 NS, .70 - .71)	.73 (2 NS, .72 - .73)
Concepts/Est	NS (3 NS; .79)	NS (3 NS; .72)	NS (3 NS; .69)	NS	NS	NS
Computation	NS (3 NS; .69)	NS	NS	NS	NS	NS
Reading Total	NS	NS (3 NS; .68)	NS	NS	NS (3 NS; .69)	NS
ITED Scores						
Con/Prob (aka Math Total)	.30 (2 NS, (.23 - .26)	.44 (.30 - .51)	.47 (1 NS, .47 - .48)	.28 (1 NS, .24 - .30)	.35 (.27 - .40)	.33 (1 NS, .29 - .38)
Computation	.31 (2 NS, .24 - .38)	.42 (.37 - .48)	.43 (1 NS, .41 - .46)	.26 (2 NS, .26 - .26)	.34 (.29 - .36)	.31 (1 NS, .28 - .34)
Reading Total	.27 (1 NS, .25 - .30)	.37 (.33 - .46)	.37 (1 NS, .35 - .39)	NS (3 NS; .25)	NS (3 NS; .36)	NS (3 NS; .33)
IAAT Scores						
Total	.36 (2 NS, .30 - .41)	.56 (.49 - .61)	.51 (.32 - .60)	.29 (1 NS, .27 - .32)	.43 (.29 - .55)	.38 (1 NS, .35 - .44)
Part A	NS (3 NS; .33)	.44 (.41 - .45)	.43 (1 NS, .38 - 51)	NS (3 NS; .26)	.35 (.26 - .42)	.30 (1 NS, .26 - .33)
Part B	NS (3 NS; .41)	.51 (.46 - .54)	.46 (.30 - .60)	NS	.39 (1 NS, .28 - .46)	.32 (2 NS, .31 - .32)
Part C	.31 (1 NS, .23 - .36)	.51 (.42 - .59)	.48 (.30 - .58)	.32 (2 NS, .28 - .35)	.36 (.27 - .47)	.38 (2 NS, .34 - .42)
Part D	.31 (2 NS, .29 - .33)	.51 (.43 - .57)	.47 (.34 - .52)	.34 (1 NS, .33 - .34)	.47 (33 - .58)	.43 (1 NS, .39 - .47)

[a] NS = nonsignificant

[b] Only 8th grade students completed the ITBS; all other students completed the ITED. Therefore, ITBS scores are based on Ns of 14 to 15

Correlations with the *grade-based measures* revealed relatively weak relations between the measures and students' performance on the algebra probes. In general, the correlations were in the .3 to .4 range, with similar coefficients for the overall GPA and for the fall algebra grade. Where differences existed, the stronger coefficients tended to be for the overall GPA. This is not surprising, given that the overall GPA represents a composite of academic performance. The obtained correlations are similar to other findings in the CBM literature base for mathematics, in which correlations between progress monitoring measures and grade-based measures are often low at best (often in the .3 to .4 range) and frequently non-significant, in part because grades include much more than isolated academic achievement. Students' work habits, motivation, and attitude also influence the grade a teacher assigns.

Scores obtained from the *teacher rating* of algebra proficiency revealed correlations in the low range, with the *Translations* probe having higher coefficients than the *Content Analysis-Multiple Choice* probe. On the *Translations* probe, the two corrected scoring procedures produced higher coefficients than did the total points correct scores approach. On the *Content Analysis-Multiple Choice* probe, the Correct minus Incorrect procedure produced the highest relative coefficients, but these were in the low range.

Two types of *standardized achievement test* data were included in the analysis: Iowa Tests of Basic Skills (ITBS) and Iowa Tests of Educational Development (ITED). Readers should note that students completed these tests in the spring of the academic year, so the correlational analyses involving the test scores address the extent to which students' scores on the probes predicted future performance on the achievement tests. Eighth grade students completed the ITBS, so the data in the table's ITBS section reflect only the ten eighth grade students in Algebra 1 classes. None of the scoring methods for the *Translations* probe produced significant results, although this is not surprising with such a small sample. Students' scores on the *Content Analysis-Multiple Choice* probe were strongly correlated (.71 - .73) to their score on the Problems/Data subtest of the ITBS, regardless of which type of scoring procedure was used. Students' performance on the reading portion of the ITBS was not related to their performance on either of the two types of probes investigated in this study.

The remainder of the students in the sample (in grades 9 to 11) completed the ITED as their district-wide achievement measure. Scores were available for two mathematics subtests: Concepts/Problems and Computation. In the district records, a Total Math score was also listed. Because this score was identical to the Concepts/Problems score in all cases, it was not included in the analyses. Reading scores were also included in the analyses to determine the extent to which reading proficiency might be associated with performance on the algebra probes. Relations between the *Translations* probe and the ITED scores were low, generally in the .3 to .4 range. We had anticipated that students' scores on this measure would reflect stronger correlations with the Concepts/Problems subtest than the Computation subtest. While the majority of the differences that occurred were in this direction, the size of the differences was small. Relations between students' scores on the *Content Analysis-Multiple Choice* measure and the ITED subtests were even smaller, with coefficients in the .2 to .3 range. While no significant relations between ITED reading performance and the *Content Analysis-Multiple Choice* probe were identified, small (but statistically significant) relations were found between ITED reading and the *Translations* probe. The size of the obtained correlations lead us to believe that neither of the measures are not likely to be especially helpful in predicting future performance on district achievement measures. This result is not surprising because neither the ITBS nor the ITED includes much attention to algebra.

The *algebra aptitude* measure consisted of four subscale scores and a total score from the IAAT. The subscales included Part A: Interpreting Mathematical Information, Part B: Translating to Symbols, Part C: Finding Relationships; and Part D: Using Symbols. Correlations between the IAAT subtest and total test scores were in the .3 to .5 range for the *Translations* probe. For the *Content Analysis-Multiple Choice* probe, coefficients were in the .2 to .4 range. The two corrected scoring procedures produced the highest coefficients for the *Translations* probe, while the Correct minus Incorrect procedure produced the highest coefficients for the *Content Analysis-Multiple Choice* probe. One interesting pattern in the results was that the IAAT Total score produced the highest correlations with the *Translations* probe, while the Using Symbols subtest (Part D) was most strongly related with the *Content Analysis-Multiple Choice* probe.

Summary of Criterion Validity Correlation Coefficients for Individual Probes

In general, relations between single scores on the *Translations* and *Content Analysis-Multiple Choice* probes were weak, with coefficients in the .2 to .4 range. The *Translations* measure produced slightly higher coefficients than did the *Content Analysis-Multiple Choice* measure for most variables. One notable exception to this pattern was for the 8th grade sample, for which the *Content Analysis-Multiple Choice* measure produced strong relations (r = .71 - .73) with the Problems and Data subtest of the ITBS. Another exception was the relation between the two corrected scores on *Translations* measure and the IAAT total score (.51 - .56). The two correction procedures produced stronger relations than did the raw scores, regardless of probe type. For the *Translations* probe, the Correct – Incorrect procedure produced similar, if not larger coefficients. For the *Content Analysis-Multiple Choice* probe, there was not a clear pattern favoring one correction method over the other.

Criterion Validity for Aggregated Probe Scores

In our earlier analyses, we found that only limited gains in reliability were obtained when the scores from two forms of an algebra probe were aggregated. In Table 9, we report the criterion validity coefficients using aggregated scores for each of the probes. To aggregate, we first averaged the two scores of a probe type that were administered on the same day. This produced two scores for the *Translations* and *Content Analysis-Multiple Choice* probes (Day 1 aggregate, Day 2 aggregate). We also aggregated scores from a single form across data collection sessions (Form 1 aggregate, Form 2 aggregate). To report the results of correlations involving aggregated probe scores, we considered the four coefficients produced for each probe and summarized these results in Table 9 using the same reporting conventions used in Table 8.

With only a few minor exceptions, aggregating students' scores across multiple probes produced stronger relations with the criterion variables. This was especially notable for the 8th grade sample with the ITBS data, where correlations were very strong (.7 to .8) with the *Translations* probe. Moreover, because the students completed the ITBS in the spring of the academic year, these correlations represent a measure of predictive validity, rather than the concurrent validity evaluated by the other criterion measures. As with the coefficients for single probes, the relations with the criterion measures were stronger for the *Translations* probe than for the *Content Analysis-Multiple Choice* probe. In addition, the two corrected scoring procedures produced stronger relations than did the raw "number correct" scores. Using the aggregated scores resulted in more definitive results for the correction procedures for the *Content*

Table 9. Criterion Validity Results for Aggregated Probes: Mean Correlation Coefficients and Ranges

Criterion Measure	*Translations*			*Content Analysis-Multiple Choice*		
	Correct	C - I	1/3	Correct	C - I	1/3
Overall GPA	.32 (2 NS, .21 - .32)	. 46 (.45 - .47)	. 47 (.42 - .50)	.31 (1 NS, .22 - .38)	.44 (.36- .52)	.37 (.26 - .47)
Grade in Algebra	.27 (1 NS, .21 - .31)	. 40 (.38 - .42)	. 42 (.35 - .45)	.40 (2 NS, .39 - .40)	.43 (.31 - .46)	.38 (.27 - .48)
Teacher Rating	.37 (1 NS, .34 - .41)	.52 (.48 - .56)	.52 (.46 - .55)	.35 (2 NS, .31 - .38)	.44 (.37 - .52)	.35 (.27 - .47)
ITBS Scores[b]						
Math Total	.80 (2 NS, .79 - .80)	.77 (2 NS, .75 - .79)	.78 (2 NS, .77 - .79)	NS (3 NS; .73)	NS (3 NS; .69)	NS (3 NS; .72)
Prob/Data	.76 (2 NS, .75 - .77)	.77 (2 NS, .74 - .79)	.78 (2 NS, .76 - .79)	.76 (2 NS, .66 - .85)	.74 (2 NS, .65 - .82)	.75 (2 NS, .66 -.85)
Concepts/Est	.78 (2 NS, .76 - .80)	.75 (2 NS, .69 - .81)	.75 (2 NS, .72 - .78)	NS	NS	4 NS
Computation	.74 (2 NS, .73 - .74)	NS (3 NS; .69)	NS (3 NS; .68)	NS	NS	4 NS
Reading Total	NS (3 NS, . 40)	NS	NS	NS	NS	4 NS
ITED Scores						
Con/Prob (aka Math Total)	.34 (2 NS, .33 - .34)	.49 (.46 - .51)	.46 (.39 - .53)	.32 (1 NS, .26 - .39)	.40 (.36 - .46)	.36 (.23 - .48)
Computation	.36 (2 NS, .35 - .37)	.48 (.47 - .49)	. 46 (.41 - .50)	.30 (2 NS, .28 - .31)	.39 (.34 - .47)	. 33 (.25 - .42)
Reading Total	.33 (2 NS, .32 - .34)	.42 (.39 - .43)	. 36 (.29 - .41)	NS (3 NS; .23)	.32 (2 NS, .29 - .34)	NS (3 NS; .32)
IAAT Scores						
Total	.33 (1 NS, .23 - .42)	.60 (.58 - .62)	.58 (.52 - .64)	.32 (1 NS, .25 - .39)	.49 (.47 - .58)	.39 (.26 - .52)
Part A	.28 (2 NS, .25 - .30)	.46 (.44 - .48)	.42 (.35 - .49)	.26 (2 NS, .25 - .26)	.40 (.38 - .42)	.30 (.26 - .36)
Part B	.32 (2 NS, .28 - .36)	.55 (.52 - .58)	.52 (.44 - .58)	NS (3 NS; .23)	.39 (.24 - .52)	.35 (2 NS, .29 - .40)
Part C	.33 (1 NS, .24 - .42)	.55 (.51 - .57)	.53 (.49 - .77)	.29 (1 NS, .24 -.34)	.40 (.35 - .46)	.34 (.26 - .42)
Part D	.29 (.24 - .36)	.56 (.55 - .58)	.54 (.50 - .57)	.34 (.22 - .46)	.54 (.40 - .64)	.45 (.30 - .59)

[a] NS = nonsignificant

[b] Only 8th grade students completed the ITBS; all other students completed the ITED. Therefore, ITBS scores are based on Ns of 14 to 15

Analysis-Multiple Choice probe, favoring the Correct – Incorrect over the Correct minus 1/3 (Incorrect).

Relations between the criterion measures and aggregated scores from the probes were in the low to moderate range. Grade-based measures produced coefficients in the .3 to .4 range, while teacher ratings were in the .3 to .4 range for the *Content Analysis-Multiple Choice* measure, but in the .3 to .5 range for the *Translations* measure. Relations with standardized test scores were much stronger for the *Translations* probe than for the *Content Analysis-Multiple Choice* probe, which had only a weak relation with the ITED subtests and (with the exception of Problems/Data) no significant relation with ITBS scores. The *Translations* measure also produced moderate (.4 to .6) correlations with the IAAT Total and subtest scores, while the *Content Analysis-Multiple Choice* measure demonstrated weaker relations (most in the .3 to .4 range).

Discrimination Between Groups

As a second means of investigating the validity of the measures, we examined whether the scores of students in different algebra options differed at a level that was statistically significant. To conduct this analysis, we labeled each participating student as belonging to one of four groups: advanced (8th grade student taking high school Algebra I), typical (Algebra I), and 'slower pace' (Pre-Algebra), and special education (Special Education Algebra or Pre-Algebra). Because the students enrolled in Pre-Algebra were completing the first half of the content of Algebra 1 across a full academic year, we selected the label of 'slower pace' for this group. We opted to aggregate scores from similar probes collected on the same day to minimize the number of tests required. Means and standard deviations for each group on each probe are reported in Table 10. Data on the means for each of the measures are depicted in graphic form in Figure 1; the first column of data points for each probe represents the total Correct points score, while the second column represents the Correct minus Incorrect scores and the third represents the Correct minus 1/3 Incorrect scores.

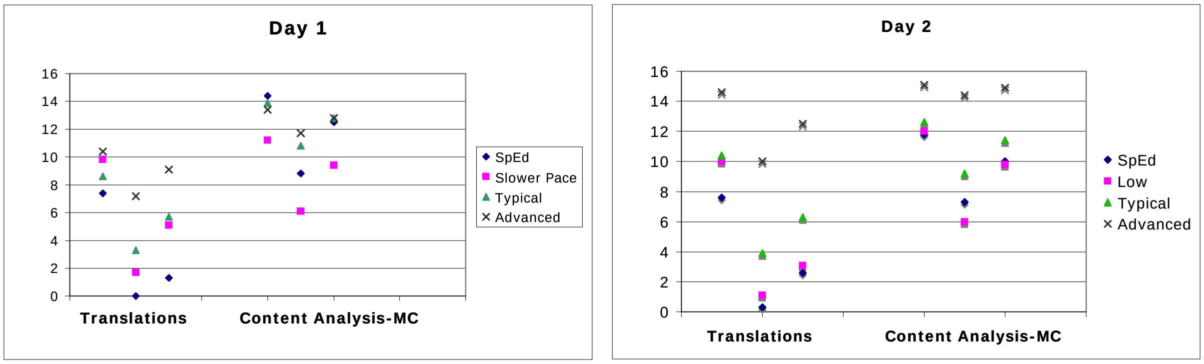
We would expect that the eighth grade students taking algebra (Advanced) group would have the highest scores, followed by the Typical group, then students in the Slower Pace group. Students receiving algebra instruction in a special education setting were expected to have the lowest scores. As the data presented in Table 10 and Figure 1 indicate, this expected pattern of results was obtained for the two sets of corrected scores for the probes on both days. The clearest distinctions between the four groups were obtained when the Correct minus Incorrect scoring procedure was used.

We next conducted analyses of variance to pursue the statistical significance of the differences between the groups' scores. The final column in Table 10 indicates that the differences on the *Translations* probe for both of the corrected scoring procedures were statistically significant for each administration. On the *Content Analysis-Multiple Choice* probes, only the Correct minus Incorrect scoring procedure resulted in consistent differences between groups. We then used Scheffe post-hoc multiple comparison tests to identify where significant differences between the groups were found. The results of the post hoc comparisons are presented in Table 11. These data reflect the limited ability of the three different scoring methods to differentiate between students in the four different performance groups. When significant differences were obtained, they typically differentiated between the Advanced (8th

Table 10. Means and Standard Deviations on Three Probes by Group Type

Probe and Day	Advanced Mean	Advanced SD	Typical Mean	Typical SD	Slower Pace Mean	Slower Pace SD	Special Ed. Mean	Special Ed. SD	ANOVA Results significant
Translations – Day 1									
Correct	10.35	3.61	8.55	4.06	9.79	4.72	7.38	0.75	
C – I	7.20	5.44	3.31	3.82	1.71	1.83	0.0	0.00	*
1/3	9.13	4.40	5.74	4.09	5.08	3.66	1.33	1.24	*
Translations – Day 2									
Correct	14.56	4.99	10.43	3.89	10.00	4.72	7.63	3.42	*
C – I	10.00	6.85	3.94	5.06	1.08	2.37	0.25	0.00	*
1/3	12.54	6.09	6.27	5.27	3.10	3.11	2.62	2.88	*
Content Analysis-Multiple Choice Day 1									
Correct	13.35	3.93	13.90	4.80	11.23	5.67	14.38	6.29	
C – I	11.65	4.12	10.76	4.60	6.12	4.51	8.75	3.30	*
1/3	12.78	3.98	12.78	4.74	9.37	5.19	12.5	5.30	
Content Analysis-Multiple Choice Day 2									
Correct	15.06	3.96	12.60	3.14	11.96	3.73	11.75	6.25	
C – I	14.44	4.01	9.21	4.48	6.04	2.71	7.25	4.21	*
1/3	14.85	3.97	11.37	3.44	9.76	3.19	9.96	5.30	*

Figure 1. Mean probe performance by group status.



grade algebra) students and other students Readers should note that the Special Education group included only four students.

Table 11. Post Hoc Comparisons by Group

Probe and Day	Post Hoc Analyses of Significant Difference Between Groups
Translations – Day 1	
Correct	no significant difference between groups
C – I	Advanced > Typical, Slower Pace, SpEd
1/3	Advanced > SpEd
Translations – Day 2	
Correct	Advanced > SpEd
C – I	Advanced > Typical, Slower Pace, SpEd
1/3	Advanced > Typical, Slower Pace, SpEd
Content Analysis-Multiple Choice Day 1	
Correct	no significant difference between groups
C – I	Advanced, Typical > Slower Pace
1/3	no significant difference between groups
Content Analysis-Multiple Choice Day 2	
Correct	no significant difference between groups
C – I	Advanced > Typical, Slower Pace, SpEd
1/3	Advanced > Typical, Slower Pace

Summary and Considerations for Future Research

The purpose of this study was to examine the technical adequacy of two additional potential measures of algebra proficiency. Eighty-seven students in grades eight to eleven participated in the study; 10 of these students were receiving special education services. The data were gathered in October, when students had completed only six to eight weeks of instruction in a year-long algebra course. On two occasions, students completed two forms of a *Translations* probe and two forms of a *Content Analysis-Multiple Choice* probe. The testing sessions were one week apart and were followed one week later by the administration of the Iowa Algebra Aptitude Test. Data collected on additional criterion variables included students' grades in school and in algebra, teachers' ratings of students' proficiency in algebra, and scores on standardized achievement tests. For the eighth graders in the sample, data were available for the math tests of the Iowa Tests of Basic Skills; for the ninth through eleventh grade students, the standardized achievement test data were drawn from the Iowa Tests of Educational

Development. This summary reviews the major findings with respect to score distributions, reliability, criterion validity, barriers encountered, and issues for future research.

<u>Distributions</u>

Mean scores (problems correct) on the *Translations* probe ranged from 8 to 11, with standard deviations of 4 to 5 points. A troubling finding for this measure was that the mean number of incorrect responses exceeded the number correct in all instances, suggesting a high rate of guessing. On the *Content Analysis-Multiple Choice* probe, mean scores (points earned) ranged from 11 to 14, with standard deviations of 4 to 5. Neither probe produced an over-abundance of 0 scores and both have sufficient room for students to improve their performance as the class continues and their proficiency increases.

<u>Reliability</u>

The reliability of individual probe scores was very low, with coefficients in the .2 to .6 range and several coefficients failing to be statistically significant, even with a sample size of approximately 80 students. Both scoring methods used to correct the raw scores for guessing produced improvements in the reliability of the scores for the *Translations* probe. For the *Content Analysis-Multiple Choice* probe, improvements were obtained for test-retest reliability, but not for alternate form reliability when the correction procedures were applied. When two scores were aggregated to increase the stability of the estimate of student performance, the reliability of the *Translations* measure increased, with coefficients in the .5 to .7 range. Aggregation did not increase the reliability of scores for the *Content Analysis-Multiple Choice* measure over the levels obtained for single probes.

<u>Validity</u>

Criterion validity coefficients for single probes were in the low range, from .2 to .4. The *Translations* measure produced slightly higher validity coefficients than did the *Content Analysis-Multiple Choice* measure. The two strongest obtained relations were (1) between the eighth grade students' scores on the ITBS Problems/Data subtest and the *Content Analysis-Multiple Choice* measure, which ranged from .71 to .73 and (2) between students' scores on the *Translations* measure and their total score on the IAAT, which ranged from .51 to .56 for corrected scores.

Aggregating scores produced improvements in the criterion validity coefficients for both measures. We obtained very strong predictive validity for the *Translations* measure with the eighth grade students' ITBS Problems/Data, Concepts/Estimation, and Math Total scores (coefficients in the .7 to .8 range). Relations between the *Translations* probe and the IAAT Total score were also in the moderate range, with coefficients for corrected scores ranging from .58 to .60. With the exception of the eighth grade students' scores on the Problems/Data subtest of the ITBS, aggregated scores on the *Content Analysis-Multiple Choice* measure had low correlations with the criterion measures (most in the .3 to .4 range) that were similar to those obtained for single probes.

<u>Considerations for Future Research</u>

Several issues arose during this study that should be addressed in future research. First, we observed several instances of student behavior that led us to conclude that some students were not putting forth their best effort on the probes. During the consent process, students were informed that their performance on the tasks would not influence their grades in class. In addition, project staff scored all student papers and only data on student performance (not actual papers) were returned to teachers. Both of these factors may have increased the level of student apathy regarding the probes.

We also observed that students' scores on the first administration of a task were often substantially lower than their scores on subsequent tasks. In the future, it may be useful to incorporate a practice task that allows students to become familiar with the format of the problems and thereby reduces the 'learning curve' effect we observed between the first and subsequent administrations.

Regarding the *Translations* measure, we were surprised by the strength of the criterion validity coefficients relative to the *Content Analysis-Multiple Choice* measure. Our impression during data collection was that students were guessing at extremely high rates on the *Translations* probe. Based on our observations in the classrooms and our discussions with teachers, the types of problems on the *Translations* measure were very unfamiliar to students, which may have increased the likelihood of guessing. In addition, students were not given a clear directive not to guess on the task, so those with good test taking skills may have opted to respond to a large number of items with random guesses. Given the relative strength of the criterion validity correlations relative to the *Content Analysis-Multiple Choice* measure, it will be important to investigate this task further.

We were surprised by the dismal performance of the *Content Analysis-Multiple Choice* measure relative to the *Translations* measure. Given the close connection between the instructional materials and teachers' expectations for student learning, we had expected stronger technical adequacy data for this probe. As we reviewed our preliminary data analyses with the teachers, we identified one potential problem in the design of the task. The probe was developed by generating problems associated with one to three key concepts or skills from each chapter of the textbook. On the probe, each chapter was represented by one or two questions. In situations where there were more key concepts than questions, the specific skill or concept sampled varied from one form of a probe to another. This design characteristic may have introduced additional variance to students' scores. Another concern expressed by one teacher was the fact that data were gathered in October, when students had only completed 6 to 8 weeks of instruction. As a result, students had completed only the two chapters in the text, which addressed primarily review material. This may have contributed to a sense of frustration on the part of the students.

Future research involving the algebra progress monitoring measures should examine the following issues:

- Exploring potential strategies to increase student motivation to put forth their best work when completing the probes
- Incorporating a 'practice probe' each time a new probe format is introduced to allow students to become familiar with the format and content of the measure
- Revising the design template for the *Content Analysis-Multiple Choice* measure so that alternate forms are assessing parallel content
- Refining scoring rubrics for the *Content Analysis-Multiple Choice* measure to further increase interscorer agreement

References

Foegen, A. (2000). Technical adequacy of general outcome measures for middle school mathematics. *Diagnostique, 25*, 175-203.

Foegen, A., & Lind, L. S. (2004). *Reliability and criterion validity for three potential algebra measures* (Tech. Rep. No. 2). Ames, IA: Iowa State University, Project AAIMS.

Fuchs, L. S., Deno, S. L., & Marston, D. (1983). Improving the reliability of curriculum-based measures of academic skills for psychoeducational decision making. *Diagnostique, 8*, 135-149.

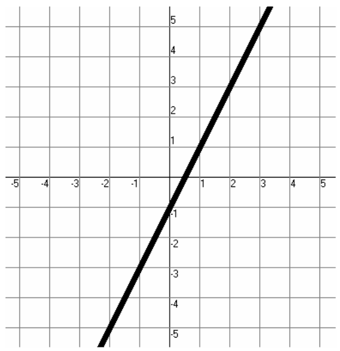
Lappan, G., Fey, J. T., Fitzgerald, W. M, Friel, S. N., & Phillips, E. D. (2004). *Prentice Hall Connected Mathematics*. Upper Saddle River, NJ: Pearson Prentice Hall.

Moses, R. P., & Cobb, C. E. (2002). *Radical equations: Civil rights from Mississippi to the Algebra Project*. Boston, MA: Beacon.

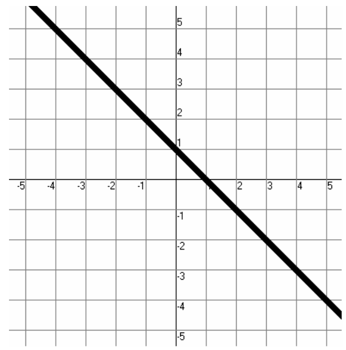
APPENDICES

Appendix A:	*Translations* – Form 1
	Translations – Form 2
Appendix B:	*Content Analysis-Multiple Choice*– Form 1
	Content Analysis-Multiple Choice – Form 2
Appendix C:	Teacher Rating Form
Appendix D:	Standardized Administration Directions

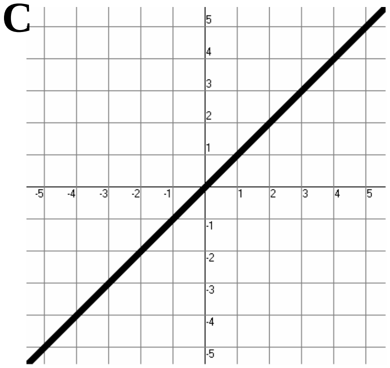
A



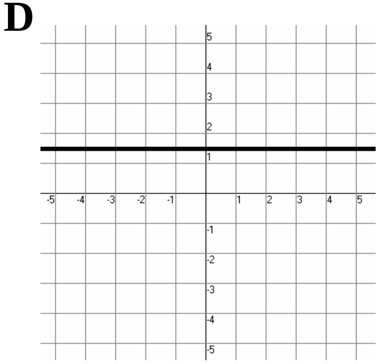
B



C



D



$y = x$ ____	$y = 2x - 1$ ____	$2y = 4x - 2$ ____	$y = 1.5$ ____	$y = -x + 1$ ____

$y = x$ ____

x	y
2	1.5
1	1.5
0	1.5
-1	1.5
-2	1.5

____

x	y
2	-1
1	0
0	1
-1	2
-2	3

____

x	y
2	3
1	1
0	-1
-1	-3
-2	-5

____

x	y
4	4
2	2
0	0
-2	-2
-4	-4

____

x	y
4	-3
2	-1
0	1
-2	3
-4	5

____

Mark needs to find half the width of pieces of pipe he is cutting to make a soccer goal. The width of the pipe is 3 inches. He made this graph to show the relationship between the length and the width of the pieces he will cut.	____
Every day that Cindy waters the garden, she earns a dollar. She made this graph to show the relationship between the number of days she waters the garden and the number of dollars she will earn.	____
Joe has one dollar in his wallet. He made this graph to show the relationship between the number of dollars he borrows from his friends for lunch and the total amount of money he has or owes.	____
The class earns $2 for each magazine subscription sold in the fund-raiser. A $1 fee per student is charged for a processing fee. Cindy made this graph to show the relationship between the number of magazines sold and the profit.	____
The flood waters are receding at a rate of 1 foot per day. The river is currently at 1 foot above flood stage. Tom made this graph to show the relationship between the number of days and the height of the river compared to flood stage.	____

A	**B**	**C**	**D**
$y = 2x + 1$	$y = 2^x$	$y = x - 1$	$y = x^2$

x	y
2	1
1	0
0	-1
-1	-2
-2	-3

____

x	y
4	16
2	4
0	0
-2	4
-4	16

____

x	y
2	4
1	2
0	1
-1	$\frac{1}{2}$
-2	$\frac{1}{4}$

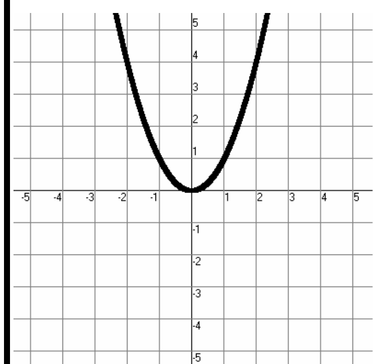
____

x	y
2	5
1	3
0	1
-1	-1
-2	-3

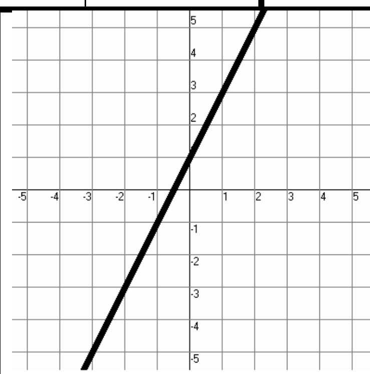
____

x	y
4	3
2	1
0	-1
-2	-3
-4	-5

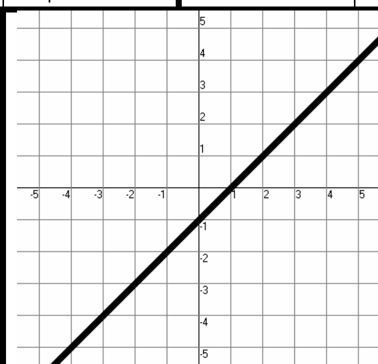
____



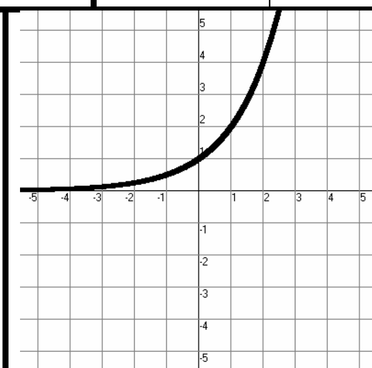
____



____



____



____

Mr. Jones is going to give a true/false test. He wrote this equation to show the number of possible answer combinations his students can give on the test. ____

Sue wrote this equation to figure out how many inches of wire she needs for a bracelet. Each bracelet uses two strands and she needs to add an extra inch to make a hood to fasten the bracelet. ____

Sam's allowance changes every year. Each month his mom pays him a dollar for each year he has lived, multiplied by his age. Sam wrote this equation to figure out his allowance. ____

Every time Hans delivers newspapers, he keeps one for his family. Hans wrote this equation to show how many newspapers he delivers to families on his route. ____

Tim's washing machine 'eats' socks. The first time he lost one sock in the wash. Now, every time he washes a load of clothes, he loses two socks. Tim wrote this equation to figure out how many socks he is losing. ____

A

x	y
2	-3
1	-1
0	1
-1	3
-2	5

B

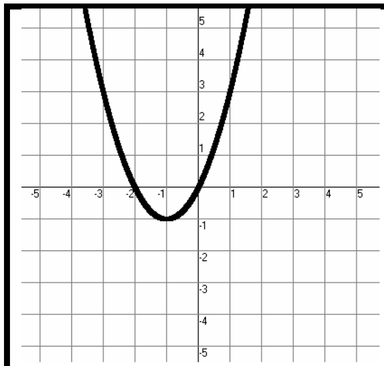
x	y
4	2
2	1
0	0
-2	-1
-4	-2

C

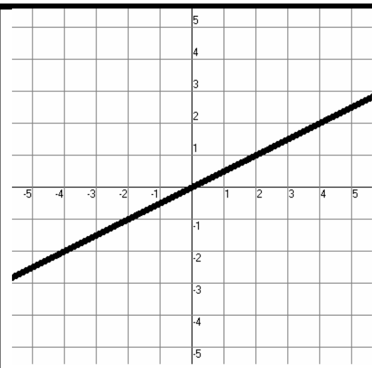
x	y
2	9
1	3
0	1
-1	$\frac{1}{3}$
-2	$\frac{1}{9}$

D

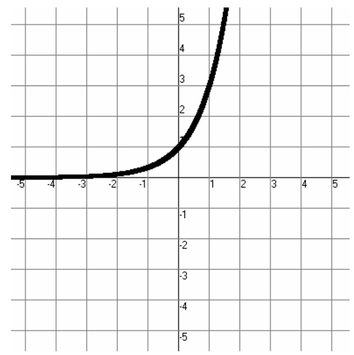
x	y
2	8
1	3
0	0
-1	-1
-2	0



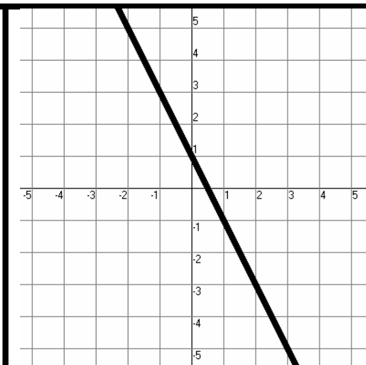
____



____



____



____

$y = -2x + 1$ ____	$y = x^2 + 2x$ ____	$y = 3^x$ ____	$y = x(x + 2)$ ____	$y = \frac{1}{2}x$ ____

Matt built a maze for his gerbil. Each time the gerbil comes to an intersection, it can go three possible ways. Matt made this table to show the total possible number of routes for the gerbil through the maze. ____

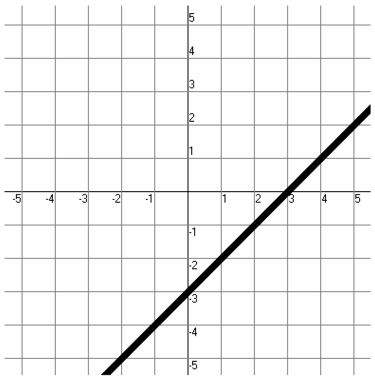
LaShaya's mom makes her save half of what she earns in the summer for college. She made this table to show how much money she will earn for her college fund this summer. ____

A diving board is one foot above the surface of the pool. An average diver drops twice his height when he steps off the board. Marcus made this table to show a diver's depth in the water. ____

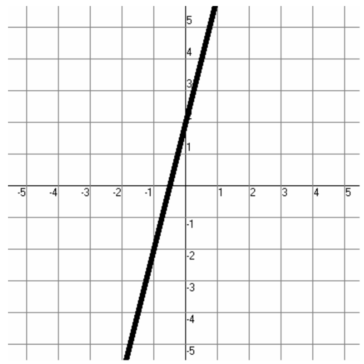
Ming Hui has two cats, Oscar and Otis. She knows that Oscar eats twice as much as Otis. She made this table to show how much Otis eats. ____

Tammy is making a backdrop for the school play. She needs to add on to a square piece of wood. The piece she will add is the same height as the square, but only 2 feet wide. Tammy made this table to show the area of the backdrop. ____

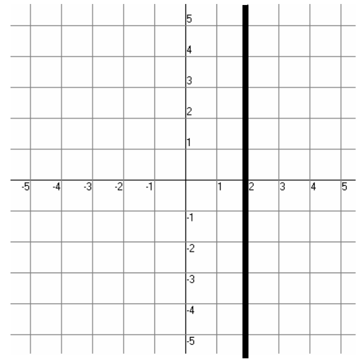
A



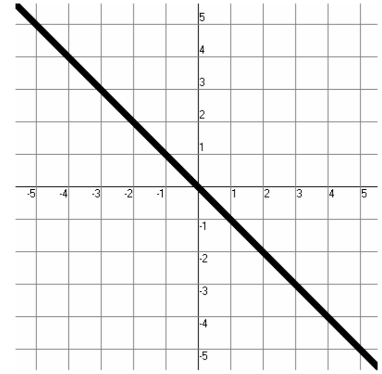
B



C



D



$3y = 3x - 9$ ____	$y = 4x + 2$ ____	$y = x - 3$ ____	$y = -x$ ____	$x = 2$ ____

x	*y*
4	-4
2	-2
0	0
-2	2
-4	4

____

x	*y*
2	4
2	2
2	0
2	-2
2	-4

____

x	*y*
2	-2
1	-1
0	0
-1	1
-2	2

____

x	*y*
4	18
2	10
0	2
-2	-6
-4	-14

____

x	*y*
4	1
2	-1
0	-3
-2	-5
-4	-7

____

Tim is collecting state quarters for his state. He started his collection with two quarters. He wants to trade in some dollar bills for quarters. Tim made this graph to show how many quarters he'll have after the trade.	____
Leah is three years younger than her sister. She made this graph to show the relationship between their ages.	____
Every time he gets home after curfew, he loses a chance to use the car. Joel made this graph to show the relationship between breaking curfew and his chances to use the car.	____
Sam is planning a basketball tournament. He made this graph to show the relationship between the number of teams in the championship game and the total number of teams in the tournament.	____
Teresa has taken four quizzes and gotten the same score on each one. She also has two extra credit points. Teresa made this graph to show how her total quiz points would be related to the score she gets on each quiz.	____

A	B	C	D
$y = 16(.5)^x$	$y = -2x - 1$	$y = x + 1$	$y = x^2 - 1$

x	y
2	-5
1	-3
0	-1
-1	1
-2	3

____

x	y
4	15
2	3
0	-1
-2	3
-4	15

____

x	y
2	4
1	8
0	16
-1	32
-2	64

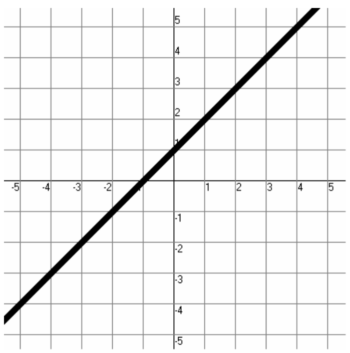
____

x	y
2	3
1	2
0	1
-1	0
-2	-1

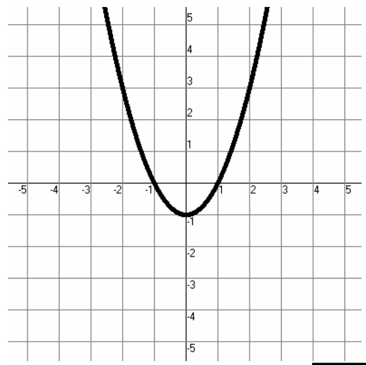
____

x	y
4	-9
2	-5
0	-1
-2	3
-4	7

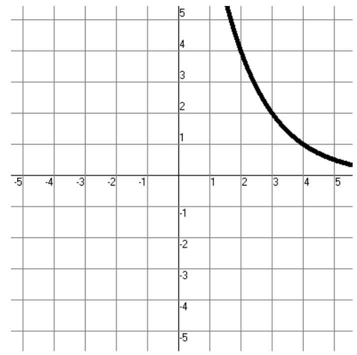
____



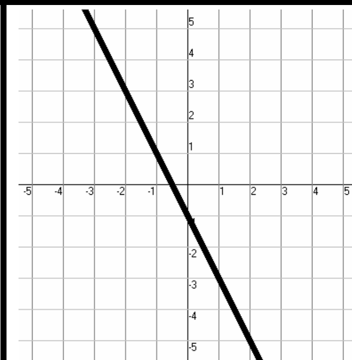
____



____



____



____

Pat is organizing the brackets for the doubles tennis tournament. Sixteen teams have entered. Pat wrote this equation to show how many teams will be left after each of the rounds. ____

LeRoy needs to buy tile for a square room. The tiles come in 1-foot squares. There is a post in the middle of the room that is the same size as one tile. LeRoy wrote this equation to find how many tiles he will need. ____

Elaine's mom gives her a list of chores to do each week. Before the week is over, she always finds one more thing that Elaine needs to do. Elaine wrote this equation to show the number of chores she does each week. ____

When Maria eats hot lunch, it costs two dollars. She already owes her sister a dollar. Maria wrote this equation to find out how much less money she'll have each time she eats hot lunch. ____

Ryan has a stool that is one foot tall. He wrote this equation to find the height of any person who stands on the scale. ____

A

x	y
4	14
3	6
2	2
1	0
0	-1

B

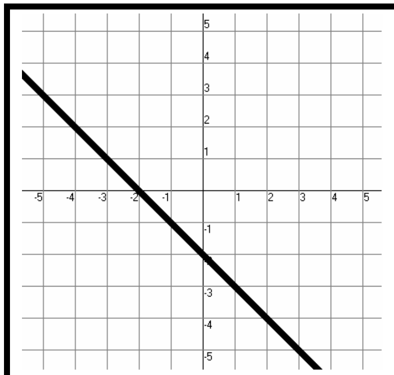
x	y
4	8
2	4
0	0
-2	-4
-4	-8

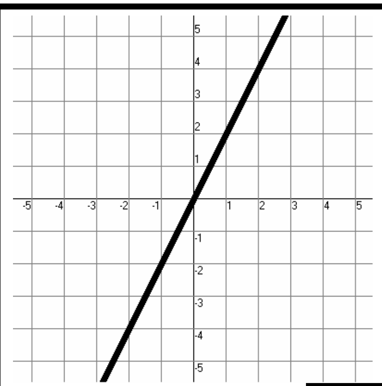
C

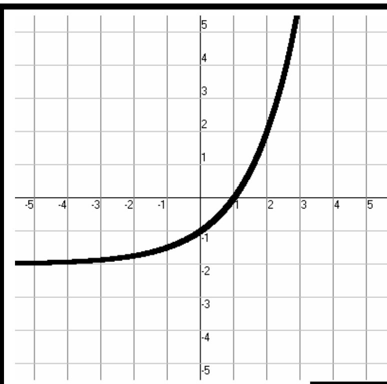
x	y
2	10
1	5
0	2
-1	1
-2	2

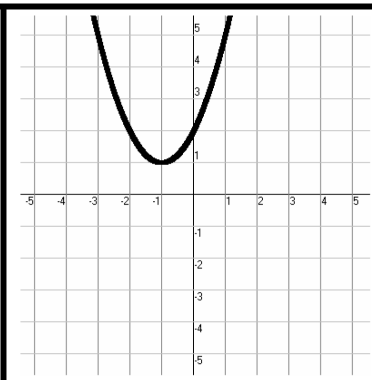
D

x	y
2	-4
1	-3
0	-2
-1	-1
-2	0


____


____


____


____

$y = x^2 + 2x + 2$ ____

$y = -x - 2$ ____

$y = 2x$ ____

$y = 2^x - 2$ ____

$y + 2 = 2^x$ ____

Bryan's dad will match his donation to the animal shelter. Brian made this table to show the relationship between how much he gives and his total donation to the shelter. ____

At the teachers' cookie swap, each teacher brings one cookie for all the teachers. The principal brings two cookies for each teacher. The cooks donate two cookies left from lunch. This table shows the number of teachers and cookies. ____

The class is planting trees for Earth Day. Each hole needs to be dug two feet deeper than the height of the root ball. This table shows the relationship between the root ball's height and the level of the ground. ____

Chris learned that a pair of mice will produce one litter of two baby mice and that when each baby matures, it will do the same. Chris made this table to show the relationship between the generations and the total mice if the original two mice die. ____

Jean changed jobs and doubled her hourly pay rate. This table shows the relationship between Jean's old and new hourly pay rates. ____

Evaluate $b^2 - a^2$ when $a = 4$ and $b = 5$

a) 21
b) 1
c) 11
d) 9

Rewrite this expression without parentheses:
$(-5)(4 - y)$

a) $9 - y$
b) $-20 + 5y$
c) $-1 - 5y$
d) $-20 - 5y$

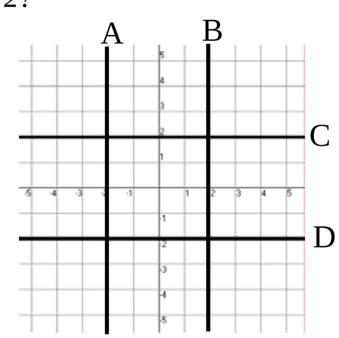
Solve:
$2t - 5 = 7$

a) $\frac{1}{2}$
b) 6
c) 1
d) 2

Solve:
$\frac{y}{3} = 4$

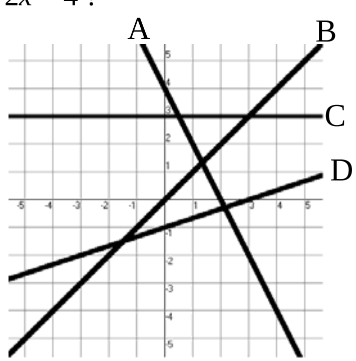
a) -10
b) 7
c) $\frac{4}{3}$
d) 12

Which line on the graph is
$y = 2$?



a) Line A
b) Line B
c) Line C
d) Line D

Which line on the graph is
$y + 2x = 4$?



a) Line A
b) Line B
c) Line C
d) Line D

Write the equation in slope-intercept form:
$m = (\frac{1}{2})$ $b = 3$

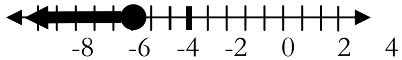
a) $y = 2x + 3$
b) $y = 3x + \frac{1}{2}$
c) $x = \frac{1}{2}y - 3$
d) $y = \frac{1}{2}x + 3$

Rewrite this equation in standard form using integer coefficients.

$$-4y + \frac{1}{2}x = 2$$

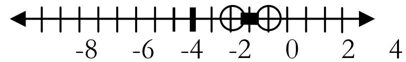
a) $-8y + 2x = 4$
b) $x - 8y = 4$
c) $y = 4x + 8$
d) $4y - 2x = 4$

This graph shows the solution for which equation?



a) $x > -3$
b) $2x \leq -6$
c) $-3x > 9$
d) $3x \geq 9$

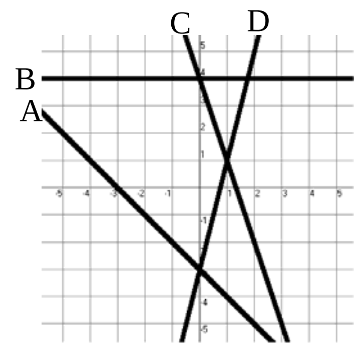
This graph shows the solution for which equation?



a) $3x > 6$ or $2x < 2$
b) $2 < 4x - 6 < 10$
c) $2 < |x| < 4$
d) $2x < 6$

Circle the TWO lines that show the solution to this linear system:

$4x - y = 3$
$3x + y = 4$



a) Line A
b) Line B
c) Line C
d) Line D

Evaluate the expression:

$$4^{-2}$$

a) -16
b) $\frac{1}{8}$
c) $\frac{1}{16}$
d) -8

Simplify

$$\sqrt{32}$$

a) $4\sqrt{2}$
b) $8\sqrt{4}$
c) $\sqrt{16} \bullet \sqrt{2}$
d) $8\sqrt{2}$

Add:
$(-x^2 + x + 2) + (3x^2 + 4x - 5)$

a) $4x^2 + 5x + 7$
b) $2x^2 + 5x - 3$
c) $2x^2 + 4x - 7$
d) $2x^2 + 3x + 3$

Simplify the expression:

$$\frac{x^2 + 4x + 4}{x^2 + 9x + 14}$$

a) $\frac{1}{5x+10}$
b) $\frac{(x+2)(x+1)}{(x+7)(x+2)}$
c) $\frac{x^2+2}{(x+2)(x+7)}$
d) $\frac{x+2}{x+7}$

Simplify:

$$4\sqrt{3} - 2\sqrt{3}$$

a) 24
b) $6\sqrt{3}$
c) $2\sqrt{3}$
d) 2

Evaluate $9 + (3 - 1) - 3^2$

a) 8
b) 2
c) 6
d) 0

Find the sum:
$9 + (-12) + 5$

a) 2
b) 26
c) 8
d) 16

Solve:
$9r - 2 = 24 - 4r$

a) $\frac{26}{9}$
b) $\frac{9}{26}$
c) $\frac{1}{2}$
d) 2

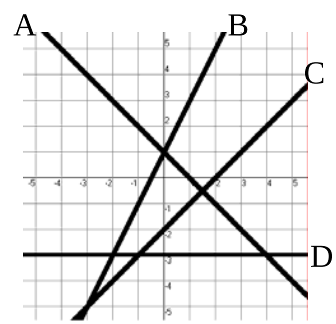
Solve:
$4x - 3 = 13$

a) 4
b) $\frac{13}{4}$
c) 10
d) 16

Find the slope of a line through (-3, 1), (2, 1)

a) $\frac{5}{2}$
b) 0
c) $-\frac{2}{5}$
d) -1

Which line on the graph is $2x + y = 1$?



a) Line A
b) Line B
c) Line C
d) Line D

Write the equation of a line through (-2, -8), (2, 4)

a) $y = 3x + 4$
b) $y = -2x + 8$
c) $y = 3x - 2$
d) $y = 2x + 4$

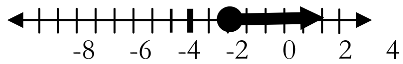
Write the equation in slope-intercept form if $m = 3$ and $b = 2$

a) $y = 3x + 2$
b) $3y = 3x + b$
c) $y = 2x - 3$
d) $y = 3x + 4$

Solve $|2x-3|=5$

a) 4, -1
b) 8, -8
c) 8, -4
d) 1, -1

This graph shows the solution for which equation?



a) $2x < 4$
b) $3x – 5 \geq 4$
c) $-6 \leq -8 + x$
d) $–x > 2$

Solve the linear system:

$2x + 5y = 7$
$7x + y = 8$

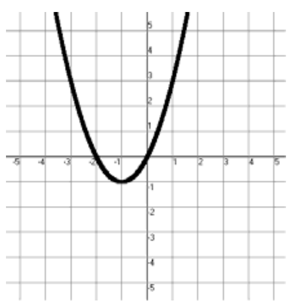
a) (-1, 1)
b) (1, 1)
c) (-2, 7)
d) (7, -8)

Simplify the expression:

$$\frac{a^2}{ab^3} \bullet \frac{b^4}{a^3}$$

a) $\frac{a^8}{a^3b^3}$
c) $\frac{ab^8}{a^4b^3}$
b) $\frac{b}{a^2}$
d) $\frac{b}{a}$

Which function matches this graph?



a) $y = 5x - 7$
b) $y = x^2 + 2x$
c) $y = -2x^2 -3$
d) $y = x^2 + 8$

Factor this trinomial:

$3x^2 - 8x + 4$

a) $(3x – 2) (x – 2)$
b) $(4x + 2) (-x + 4)$
c) $(3x – 2) (x + 2)$
d) $(2x + 2) (x + 2)$

Simplify the expression:

$$\frac{2x+1}{3x} + \frac{x+5}{3x}$$

a) $\frac{3x+6}{6x}$
c) $\frac{2x^2+11x+5}{9x^2}$
b) $\frac{2x+4}{3x}$
d) $\frac{x+2}{x}$

Solve the equation:

$$\sqrt{x-1}=5$$

a) $x^2 + 6$
b) 6
c) 26
d) $x - 4$

Teacher ___________________

Project AAIMS:
Algebra Assessment and Instruction:
Meeting Standards

District A Junior/Senior High

Directions: Below is a list of the students you teach. Please rate each student's proficiency in algebra in comparison to others in the student's grade. A rating of "1" indicates a low level of proficiency compared to others in the same grade, "3" indicates average proficiency and "5" indicates a high level of proficiency.

Student	Algebra Proficiency				
	Low		Average		High
	1	2	3	4	5
	1	2	3	4	5
	1	2	3	4	5
	1	2	3	4	5
	1	2	3	4	5
	1	2	3	4	5
	1	2	3	4	5
	1	2	3	4	5
	1	2	3	4	5
	1	2	3	4	5
	1	2	3	4	5
	1	2	3	4	5
	1	2	3	4	5
	1	2	3	4	5
	1	2	3	4	5
	1	2	3	4	5
	1	2	3	4	5

Algebra Probe Data Collection Procedures
Probes D and E
District A
October 19, 2004

Materials:
1. Student copies of the probes
2. Stopwatch/timer
3. Pencils for students

General Introduction:

As you all know, your class and other algebra classes at District A High School are working with Iowa State on a research project to learn more about improving algebra teaching and learning. Today I need your help in trying out some of the brief tasks that teachers may be able to use to track student progress. As you may remember, ALL students will complete the tasks, but we will only use your scores in the research project if you and your parent or guardian have both given us permission to do so. Please clear your desk—the only thing you'll need for this activity is a pencil. (Distribute pencils to any students who need them.)

There are a few things you should know about the tasks, or probes, we will ask you to complete today. First, we will be limiting the amount of time you have to work on the tasks. We EXPECT that you will NOT be able to finish the probes. These tasks are different from classroom tests or quizzes and are not meant to be completely finished. Second, there may be problems on the probes that are difficult or unfamiliar. Please work across each row and try each problem. If you do not know how to answer the question, skip it and go on to the next question. DO NOT spend a great deal of time on any one problem. If you get to the end of the probe and still have time to work, go back to the problems you skipped and try to solve them. Remember, your score on the probe will not hurt your grade in algebra class, but it is important for the research project that you do your best work. Do you have any questions at this point?

Directions for Reform-Oriented Probes: **D1, D2**
1. Distribute copies of the first Version D probe to all students in the group FACE DOWN. Ask students to keep the probes face down until they are told to begin.

2. Put the Sample Page on the overhead and say to the students:

This page shows a portion of [the first/another] type of task we are testing out. At the top of the page, you will see a row of graphs. On other pages, the top row may be tables or equations. Below this top row *(point),* You will see a set of four boxes (only two are shown here). As you move to the second row, your task is to match each item to one of the representations in the first row. Let's look at this sample page together.
[DO NOT REPEAT THIS PARAGRAPH FOR THE SECOND ADMINISTRATION.]

Please put your name, date, and period on the back page. This is a reform-oriented algebra probe. You will have 7 minutes to work. When I say 'begin,' turn the probe over and begin answering the problems.

3. Set timer for **3** minutes. Say ***Begin*** and start your stopwatch.
4. When timer reads **2** minutes, say ***Slash***.
5. When timer reads **1** minute, say ***Slash***.

6. When timer goes off, say ***Stop. Put your pencils down.***
7. Ask students to pass papers to the back of the room and prepare to repeat for second A version probe. Say, ***Now we will do another probe that is similar to the one you just finished.***

Directions for Content Analysis-Multiple Choice Probes: **E1, E2**

1. Distribute copies of the first Version E probe to all students in the group FACE DOWN. Ask students to keep the probes face down until they are told to begin.

2. Say to students,

This is [the first/another] type of task we are testing out. The problems on this probe represent the different types of problems that you are learning in your textbook. In general, you will probably find that the problems at the beginning are easier and those on the second page are more challenging. Look at each problem carefully before you answer it. The problems on this probe may seem more difficult than those on the probes you've already completed.

[DO NOT REPEAT THIS PARAGRAPH FOR THE SECOND ADMINISTRATION.]

When I say 'begin,' turn the sheet over and begin answering the problems. Start on the first problem on the left on the top row. Work across and then go the next row. If you can't answer the problem, make an 'X' on it and go on to the next one. Remember to make a slash mark when I say "slash." You will have 7 minutes to work.

3. Set timer for **7** minutes. Say ***Begin*** and start your stopwatch.
4. When timer reads **2** minutes, say ***Slash***.
5. When timer reads **1** minute, say ***Slash***.
6. When timer goes off, say ***Stop. Put your pencils down.***
7. Ask students to pass papers to the back of the room and prepare to repeat for second B version probe. Say, ***Now we will do another probe that is similar to the one you just finished.***

Directions for Version C Probes: **C1**

1. Distribute copies of the Version C probe to all students in the group FACE DOWN. Ask students to keep the probes face down until they are told to begin.

2. Say to the students:

3. ***When I say 'begin,' turn the sheet over and begin answering the problems. Start on the first problem on the left on the top row. Work across and then go the next row. If you can't answer the problem, make an 'X' on it and go on to the next one. Remember to make a slash mark when I say "slash." You will have 10 minutes to work.***

4. Set timer for **10** minutes. Say ***Begin*** and start your stopwatch.

5. IF ANY STUDENTS COMPLETE ALL PROBLEMS DURING THE 10 MINUTE PERIOD, PLEASE NOTE THIS ON THIS PAGE!
6. When timer goes off, say ***Stop. Put your pencils down.***
7. Ask students to pass papers to the back of the room
8. Say, ***That is the end of the tasks for today. Next Tuesday we will be back in your class to do some more of the probes. Thank you for your help with our research project!***
9. If there is time left in the period, ask the teacher if s/he wants the students to do any particular activity OR play hangman with students until the bell rings.