Hammill Institute on Disabilities

Assessment for Effective Intervention
2021, Vol. 46(2) 121–131

DOI: 10.1177/1534508419862025
aei.sagepub.com
SAGE

Technical Adequacy of Procedural and Conceptual Algebra Screening Measures in High School Algebra

Vincent R. Genareo, PhD[1], Anne Foegen, PhD[2], Barbara J. Dougherty, PhD[3], William W. DeLeeuw, PhD[4], Jeannette Olson, MS[2], and Ruveyda Karaman Dundar, PhD[5]

Abstract

Although algebra often functions as a gatekeeper to higher-level mathematics courses and higher education admissions, few quality measures exist for assessing conceptual understanding. This study explored the technical adequacy of three procedural and two conceptual algebra screening measures. We administered three rounds of assessments throughout an algebra course to 2,021 secondary students of 31 teachers in three states. Assessments included procedural and conceptual screening measures and additional criterion measures: teacher ratings of students' algebra proficiency, course grades, results of two project-specific algebra proficiency exams, and state test scores. Descriptive and correlation analyses were used to investigate measure scores, alternate-form and test–retest reliability, concurrent validity, and predictive validity. Procedural measure results indicated high levels of reliability (r = .72–.99), and moderate concurrent and predictive validity (r = .36–.64; .36–.58). The conceptual measures produced moderate to low levels of validity (r = .10–.44). The procedural measure results suggest they may be suitable for use as screening measures, pending further revision and diagnostic testing, while the conceptual measures did not produce acceptable results for current implementation. The findings contributed to measure redesigns to bolster their use as mathematics proficiency assessments with algebra students.

Keywords

curriculum-based measurement, screening/benchmarking, math

Proficiency in high school algebra often functions as a gatekeeper to students' future opportunities and achievement, including successful completion of advanced mathematics courses in high school (Spielhagen, 2006), access to postsecondary education (Adelman, 2006), and the subsequent and profound impact of a college degree on earned income (Bureau of Labor Statistics, 2015). The importance of algebra is reflected in state graduation requirements, most of which include passing an Algebra 1 course (Center for Public Education, 2013). Consequently, the future success of students with and at risk for disabilities will be limited if they do not have access to successful experiences in algebra.

Scholars advocate for balancing conceptual understanding with flexible skill proficiency, because positioning algebraic procedures within conceptual frameworks can allow students to think more deeply and connect content to past and future learning (Choike, 2000; Kilpatrick, Swafford, & Findell, 2001). Given that mathematics outcomes for students with disabilities have lagged significantly behind those of their non-disabled peers (Council for Exceptional Children Policy Insider, 2013), teachers need powerful instructional methods and materials, as well as effective assessment techniques to determine students' progress in developing proficiency in algebra. One approach to assessment, curriculum-based measurement (CBM), holds potential as a valuable tool.

CBM in Secondary Mathematics

CBM (Deno, 1985, 2003) is an empirically based assessment method that teachers can use to inform instructional decisions and effect improved achievement for students (Stecker, Fuchs, & Fuchs, 2005). CBM employs formative assessment to measure student growth by administering

[1]Salisbury University, MD, USA
[2]Iowa State University, Ames, IA, USA
[3]University of Hawai'i at Mānoa, Honolulu, HI, USA
[4]Arizona State University, Mesa, AZ, USA
[5]Bartin University, Turkey

Corresponding Author:

Vincent R. Genareo, Salisbury University, 1101 Camden Avenue, CH281G, Salisbury, MD 21801, USA.
Email: vrgenareo@salisbury.edu

brief, technically adequate measures that sample important end-of-year outcomes in mathematics. By collecting repeated samples of student performance throughout a course, graphs of students' scores can be used to evaluate the level and pace of learning and signal needed changes in intervention (Fuchs, Fuchs, Hamlett, & Stecker, 1991). Because the measures remain constant in content and difficulty (and do not change to reflect the topics in each unit of instruction), the graphed data serve as indicators of general algebra achievement.

CBM measures have been used as screening tools for identifying struggling learners (Deno, 2003), in part because of their short duration and known technical adequacy levels. The primary function of screening measures is to identify students in need of further evaluation because their performance is at the extremes of the population distribution. Many school districts have used these measures as they implemented tiered systems of instructional support such as Response to Intervention (Fuchs, Fuchs, & Compton, 2012). CBM mathematics screening measures have demonstrated that they are a potentially effective predictor of students' standardized assessment performance; therefore, there is a call for increased use of CBM in schools (Shapiro, Keller, Lutz, Santoro, & Hintze, 2006).

Conventional CBM Mathematics Measures

At the elementary grade levels, an extensive research base supports the technical quality of CBM measures for mathematics, including consequential validity (Messick, 1989); teacher use of CBM data to inform instructional decisions is associated with greater improvements in student achievement (Stecker et al., 2005). A review of research on CBM mathematics measures and their technical adequacy found very few studies addressing measures for middle school and high school mathematics (Foegen, Jiban, & Deno, 2007).

Foegen and colleagues (Foegen, 2000, 2008b; Foegen & Deno, 2001) have demonstrated promising evidence of reliability and validity for middle school mathematics (grades 6–8). The measures investigated included both existing K-8 measures that assessed computation and concepts/applications, as well as measures addressing number sense and numeracy concepts (estimation, strategic counting, quantity comparisons). At the middle school level, reliability estimates generally ranged between .70 and .90, with concurrent validity coefficients generally between .40 and .60.

More recently, Foegen and colleagues have developed and evaluated procedural measures for beginning algebra (Espin, Chung, Foegen, & Campbell, 2018; Foegen, 2008a). Three measures have been developed. Algebra Basic Skills (ABS), Algebra Foundations (AF), and Algebra Content Analysis (ACA) were designed to examine core algebraic skills ranging from rational number operations to solving systems of equations. See Online Supplemental Figure 1 for sample items from each type of measure. Estimates of technical adequacy have included median reliability coefficients in the .75–.85 range and median criterion validity coefficients in the .50–.60 range. Although these results offer support for screening of procedural outcomes, no measures investigated to date have included items for conceptual understanding.

Exploring Conceptual Outcomes From a CBM Perspective

Conceptual knowledge and procedural fluency are interrelated, and strong conceptual foundations lead to students' success in mathematics (Rittle-Johnson, Siegler, & Alibali, 1999; Star, 2005). Assessing students' conceptual understanding is increasingly important given the emphasis on conceptual instruction in the Common Core State Standards-Mathematics (CCSS-M, National Governors Association Center for Best Practices, Council of Chief State School Officers, 2010; Porter, McMaken, Hwang, & Yang, 2011). As part of a larger project (Foegen & Dougherty, 2010), we sought to create measures that fit common CBM constraints (e.g., brief administration time, multiple forms, efficient to score) while shifting from procedural skill items to items representing conceptual understanding.

The conceptual screening measures examined in this study were developed to address two critical areas of understanding in algebra: variables and proportional reasoning (PR). These areas were chosen because of their foundational importance to algebra and because they were not well represented in the existing procedural measures. Within each domain, items were created using three processes identified by Krutetskii (1976), a Russian psychologist, and adapted by the Curriculum Research & Development Group (Rachlin, 1998) and others (Dougherty, Bryant, Bryant, Darrough, & Pfannenstielf, 2015).

Krutetskii's (1976) multiyear program of research identified characteristics of information processing used during mathematical problem solving that differentiated more capable from less capable students. Three of these processes were especially useful in guiding the development of items on the conceptual measures: generalization, flexibility, and reversibility. These three processes form a framework for constructing questions that delve deeper into mathematical understandings. Generalization questions focus attention on patterns that can be generalized and broadened. Flexibility questions motivate students to consider how a problem is related to another problem and use that relationship in the solution method. Reversibility questions ask students to reverse their thinking, in some cases, thinking backwards.

We drew upon these question types as we developed items for the conceptual measures (c.f., Dougherty et al., 2015). We classified items as representing generalization, flexibility, and reversibility. See Online Supplemental

Figure 1 for example items representing these characteristics for the two conceptual algebra measures: Concept of Variable (CoV) and PR. The conceptual items were designed to require attention to problem structure (e.g., CoV flexibility item example) or item features (such as the range of possible substitutions in the Bart problem for CoV), rather than rapid solution procedures. An expert panel of mathematics education and special education researchers reviewed the measures for content and alignment to the CCSS-M (Dougherty et al., 2017).

The purpose of the study reported here was to determine how well the procedural and conceptual algebra screening measures met expectations for technical adequacy. Given the absence of screening tools that reflect conceptual outcomes, particularly those aligned with a CBM approach, our study represents a first examination of the viability of such measures. If the procedural and conceptual measures examined in the study prove to have acceptable technical characteristics, they may provide teachers formative, data-based indicators of learning to identify struggling learners, inform instructional decisions, and provide interventions to maximize learning for all students. In this study, we addressed the following research questions:

1. What levels of reliability (alternate-form and test–retest) were produced by the procedural algebra measures?
2. What levels of concurrent criterion validity were produced by the procedural and conceptual algebra measures?
3. What levels of predictive criterion validity were produced by the procedural and conceptual algebra measures?

Method

The data we report represent the second year of a 4-year, federally-funded grant project to develop and refine procedural and conceptual progress monitoring measures in algebra, collaboratively led by faculty at two research universities (Foegen & Dougherty, 2010).

Setting and Participants

The study took place in three districts, each from a different state (States A, B, and C). District 1 in State A was located in a Midwestern city and enrolled 14,100 students, with race/ethnicity proportions of 57% White, 29% Hispanic, 5% Black, and 9% other ethnicities. Three high schools in this district participated in the study. Approximately 55% of the district's students qualified for free or reduced-price meals, 14% received special education services, and 19% were English Language Learners (ELLs). In State B, District 2 was located on the outskirts of a major Midwestern urban area and enrolled about 14,000 students, with student race/ethnicity representations of 69% White, 13% Hispanic, 11% Black, and 7% other ethnicities. Three high schools in this district participated in the study. In District 2, approximately 66% of the students qualified for free or reduced-price meals, and 11% received special education services; ELL data were not available for District 2. In State C, District 3 was on the border of a major urban area in a Southern county. District 3 enrolled over 32,000 students and had race/ethnicity representations of 61% White, 6% Hispanic, 31% Black, and 2% other ethnicities. Two high schools in District 3 participated in the study. In District 3, approximately 66% of the students qualified for free or reduced-price meals, 14% were receiving special education services, and 4% were ELLs.

Within each district, we recruited teacher participants who were teaching traditional Algebra 1 courses or Algebra 1A and 1B courses (content of Algebra 1 taught over 2 years). A total of 31 teachers participated in this study. See Online Supplemental Table 1 for demographic information for the teachers. Students within these teachers' classes were invited to participate in the study. Because the teachers were administering the assessments to the entire class, the study measures were deemed exempt for consent purposes, and all available data were used in the analyses. A total of 2,021 students were represented in the data. Because access to student record data (demographics, state test scores, Individualized Education Plan [IEP], and English as a second language [ESL] status) was limited to students for whom we were able to obtain parent consent and student assent, a significantly smaller set of students were included in the analyses involving these variables. Table 1 reports the demographic characteristics of the 1,200 students for whom consent was obtained. Remaining students ($n = 821$) were not included in the calculations. Before limiting the analyses to students with demographic information, the full sample included $N = 2{,}021$ (State A $n = 573$, State B $n = 586$, and State C $n = 862$). The sample was roughly evenly divided between males and females, primarily ninth grade students (State A/District 1 was an exception, with a large proportion of students in 10th grade), and approximately 70% White students. ELL students represented approximately 4% of the consented sample, 10% were students receiving special education services, and more than 40% were receiving free or reduced-price meals.

Data Sources

The data sources for the study included five screening measures and five criterion measures. Each of these is described further in the section below.

Screening measures. Three procedural measures and two conceptual measures were administered to students. The

Table 1. Student Participant Demographic Data for Consented Students.

	Full Sample		State A		State B		State C	
Characteristic	*N*	%	*n*	%	*n*	%	*n*	%
Female	625	52	135	58	196	54	297	49
Male	575	48	96	42	166	46	313	51
Grade								
9	973	81	113	49	315	88	545	89
10	165	14	81	35	41	11	43	7
11	50	4	32	14	3	<1	15	3
12	13	1	5	2	1	<1	7	1
Race								
Black	151	13	12	5	83	23	56	9
Hispanic	88	7	n/a	n/a	11	3	77	13
Other	76	8	23	10	11	3	66	11
White	860	72	196	85	253	71	411	67
IEP	115	10	18	8	49	14	48	8
ELL	59	5	39	17	2	<1	18	3
FRL	514	43	n/a	n/a	105	29	409	67
Total consent	1,201	100	231	100	360	100	610	100

Note. Demographics of those who consented for release of their information. Total *N* = 2,021. Percentages refer to only those who supplied demographic information. IEP = Individualized Education Plan; ELL = English Language Learners; FRL = Free/Reduced Lunch eligible.

procedural measures (ABS, AF, and ACA) were initially developed in an earlier funded project (Foegen, 2003). The present project (Foegen & Dougherty, 2010) expanded this initial work by examining larger, more representative student samples; refining the procedural measures; and developing screening measures designed to tap conceptual understanding of algebra. The mathematics education co-principal investigator led the development of the conceptual measures. See Online Supplemental Table 2 for a summary of the features of each measure and Online Supplemental Figure 1 for sample items for all of the measures.

Among the procedural measures, ABS focused on skills for which some level of automaticity was required. Problems included solving simple equations, using the distributive property, computing with integers, combining like terms, and using PR. In previous research (Espin et al., 2018), the ABS measure produced median reliability estimates of .81 (alternate-form) and .83 (test–retest) and median criterion validity estimates of .53 (concurrent) and .56 (predictive). AF items represented five core skills/concepts essential to conceptual understanding in algebra: writing and evaluating expressions, calculating with real numbers (integers, exponents, and order of operations), graphing inequalities and interpreting linear equations, solving simple equations and simplifying expressions, and generalizing relations and functions. Some of the AF items were similar to items used in ABS. Both measures included items that support computational aspects of algebra and reduce cognitive load for problem solving; the range of content was more expansive in the AF measure relative to ABS. In previous research (Espin et al., 2018), this measure produced median reliability estimates of .84 (alternate-form and test–retest) and median criterion validity estimates of .57 (concurrent) and .58 (predictive). ACA problems representing a sampling of core concepts in the first two thirds of a traditional Algebra 1 course and its items addressed solving equations, evaluating expressions, finding the slope of a line, solving linear systems, and interpreting graphs of inequalities. Students could show work to obtain partial credit on the ACA measure. In previous research (Espin et al., 2018), this measure produced median reliability estimates of .79 (alternate-form) and .77 (test–retest) and median criterion validity estimates of .58 (concurrent) and .54 (predictive).

Two conceptual measures were also administered in this study. The first conceptual measure, CoV, had items addressing relationships between and among related equations and expressions, and generalizations about the behavior of algebraic expressions and equations. The items on the second conceptual measure, PR, addressed relationships between and among quantities, generalizations about the behavior of quantities in relationships, and the multiplicative nature of ratios and proportions.

Criterion measures. The criterion measures for the study included teacher ratings, performance on two algebra achievement measures, state test scores, and course grades. Because state test scores and grades were gathered from students' school records, these data were only available for the consented students within the sample.

Teacher ratings. Teachers received rating forms at the beginning and end of their course. Using a scale from 1 (*low*) to 7 (*high*), teachers rated the overall algebra proficiency of each student in their class.

Performance on algebra achievement measures. Two measures of algebra achievement were administered to students in all three districts. Developed at the University of Hawai'i, the first algebra criterion measure, the Formative Assessment in a Networked Classroom (FANC), measured conceptual understanding of algebraic processes. The FANC was a 45-min assessment on which students could score up to 33 points. Content validity for the FANC was established using a systematic, multistep process involving selection and modification of released items from the National Assessment of Educational Progress and the Scholastic Aptitude Test (P. Brandon, personal communication, December 14, 2011). Following a pilot test to refine items, the final version of the measure was administered to nearly 1,700 students and produced an internal consistency reliability coefficient of .90. The second algebra achievement criterion measure, the Iowa End-of-Course Algebra 1 Assessment (IEOC), was a standardized assessment of algebra content developed at the University of Iowa. The IEOC was a 40-min assessment on which students could score up to 30 points. In 2008 and 2009, the IEOC was administered to nearly 9,000 students at 73 schools (68 public, 5 private, 45 total districts; Iowa Testing Programs, 2010). The internal consistency coefficient for the IEOC was .80, with a standard error of measurement of 2.40 (D. Henkhaus, Iowa Testing Programs, personal communication, September 10, 2010).

State tests. Students' scores on the mathematics portion of their respective standards-based state tests were obtained. Each state administered a different test. State A's test included 40 items and represented mathematics proficiency across a range of domains (number sense and operations; algebraic patterns and connections; data analysis, probability, and statistics; geometry; and measurement). Internal consistent coefficients (Kuder-Richardson 20) exceeded .91 for grades 9–11. Validity coefficients with the mathematics section of the ACT ranged between .74 and .76. State B's test included 53 items and was developed by Educational Testing Service to reflect the state's course-level expectations for Algebra 1. The internal consistency coefficient (Cronbach's alpha) was .91, kappa values ranged from .70 to .95, and classification accuracy analyses for predicting student performance levels found low rates of false positives and false negatives (3% or lower). State C's test, which had 35 items, was developed by Riverside to reflect the state's course-level expectations. The Algebra 1 exam represented four strands. The internal consistency coefficient was .86, and Cohen's kappa of .47 was in the moderate range for interrater agreement. The strands were moderately correlated with each other, with coefficients ranging from .54 to .69. Readers should note that references for state tests have been omitted, as they identify the participating states; they are available from authors by request.

Course grade. Students' first and second semester algebra course grades were obtained from school records. Each state used a letter grade system of A through F (including plus and minus grades, when applicable). Letter grades were converted to a numerical scale (e.g., A = 4.00, A– = 3.67, B+ = 3.33, B = 3.00).

Procedures

Students completed three rounds of data collection. The first two occurred within a week of each other near the beginning of the course; the third occurred near the end of the course. Each class section completed two forms of one of the three procedural measures; this assignment was counter-balanced across classes to avoid order effects. All class sections completed both forms of the conceptual measures. In the first round, students completed two forms of one of the three procedural measures, the first form of CoV, and the first form of PR.

In the second round (within 2 weeks of Round 1), students completed the same two forms of the procedural measure as the first round. In the third round (end of course), the students took the same two forms of the procedural measure, the second form of the CoV and PR, the IEOC, and the FANC. In Districts A and B, which used daily 45-min period or modified block bell schedules, approximately 7 months elapsed between the administration of the brief predictor measures and the administration of the criterion measures. In District C, which used a semester block schedule (90 min of algebra daily for a semester), the time lapse was approximately 4 months). The brief measures were hand scored by the researchers and the algebra achievement measures and state tests were machine scored.

Analyses

In each round, students' scores were determined by using the raw score (conceptual measures) or calculating the mean of two forms (procedural measures). Pearson product–moment correlations were computed for the bulk of the analyses, including alternate-form reliability (i.e., ABS 1 with ABS 2) within rounds, and single measure, test–retest reliability coefficients (i.e., ABS 1 Round 1 with ABS 1 Round 2) across administrations. Correlation analyses were used to explore concurrent and predictive validity relations between the conceptual and procedural measures and other mathematics proficiency indicators (criterion measures). Analyses involving state tests were computed within state

Table 2. Reliability of Procedural Measures.

	ABS		AF		ACA	
Reliability Type	*r*	*N*	*r*	*n*	*r*	*N*
Alternate-Form						
Round 1	.84	506	.82	571	.72	530
Round 2	.88	496	.86	564	.81	514
Test–Retest						
Form 1	.88	458	.84	514	.77	489
Form 2	.86	454	.86	524	.78	498
Mean of 2 Forms	.91	525	.90	595	.84	555
	α[a]	*n*	α[a]	*n*	α	*n*
Internal consistency						
Form 2	.99	423	.86	473	.81	76
Form 3	.89	435	.88	470	.79	74

Note. All correlations significant at the $p < .01$ level. ABS = Algebra Basic Skills; AF = Algebra Foundations; ACA = Algebra Content Analysis.
[a]Cronbach's alpha computed using a score of 0 for all skipped items.

only, because each state administered a different test. Taylor (1990) and subsequent researchers informed correlation coefficient interpretations; validity coefficients were substantial at .45 or above and reliability coefficients were weak at .30, moderate at .50, and high at .70 or greater.

Results

Means and standard deviations for the measures are available in Online Supplemental Table 3, reported in number of points earned in the time allowed. Online Supplemental Table 4 also includes the number of items in each of Krutetskii's categories for each of the conceptual measures. Mean scores increased (Beginning, Round 1 to End, Round 3) on all procedural measures (ABS = 3.0 points, AF = 2.6 points, ACA = 5.3 points). Conceptual measure mean increases were relatively small from the beginning to the end-of-course administrations (CoV = 0.1 points, PR = 0.5 points). We used paired sample *t*-tests to determine the statistical significances of the differences; in all cases, *p* values were less than .05 and 95% confidence intervals did not include 0 (though differences in CoV were very close to these thresholds).

Research Question 1: Reliability of Procedural Measures

We examined reliability only for the procedural measures. We were unable to examine alternate-form and test–retest reliability for the conceptual measures because of the data collection schedule (the first form was administered in Round 1 at the beginning of the course and the second was administered at the end of the course). The results for reliability analyses are reported in Table 2. Alternate-form reliability coefficients for single measures within a round ranged from .72 to .88, with ACA coefficients slightly lower than those for ABS and AF. Test–retest coefficients for a single form across Rounds 1 and 2 were similar, ranging from .77 to .88. Again, coefficients for ACA were lower than those for ABS and AF. Finally, correlating the mean of two forms in Round 1 with the mean of two forms in Round 2 provided a more stable estimate of student performance, with all reliability coefficients reaching .84 or greater. We used Cronbach's alpha to examine internal consistency; given the timed nature of the measures, fewer than 10 cases could be included in the analyses for ABS and AF. To address this concern, we replaced missing item data with scores of 0 for the internal consistency analyses.

Research Question 2: Concurrent Criterion Validity of Procedural and Conceptual Measures

We examined concurrent criterion validity by correlating students' procedural and conceptual algebra screening scores gathered at the beginning and end of the course with criterion measures gathered during the corresponding time period. At the beginning of the course, student scores were correlated with beginning of the course teacher ratings. At the end of the course, criterion measures included teacher ratings, grades, and test scores. The coefficients for the overall sample and by state are available in Table 3; nearly all are statistically significant.

Among the procedural measures, coefficients ranged from .36 to .64 for the overall sample, with a median value (.55) in the moderate range. Individual state coefficients reflected greater variability, ranging from .25 to .71. Concurrent validity was lower for the conceptual measures for the overall sample, with coefficients ranging from .13 to .44, with a median value (.25) in the low range. Relations between the criterion measures and CoV were stronger than comparable relations with PR. See Table 3 for full concurrent criterion validity results.

Research Question 3: Predictive Criterion Validity of Procedural and Conceptual Measures

We evaluated predictive validity by correlating procedural and conceptual measure scores from the beginning of the year with scores obtained on end of the year criterion measures. These results, reported in Table 4, reflect patterns similar to those obtained for concurrent validity. For the procedural measures, coefficients ranged from .36 to .58, with a median of .47. For the conceptual measures, coefficients ranged from .10 to .33, with a median of .22. All of the procedural measures demonstrated similar predictive ability. The CoV measure produced stronger predictive validity coefficients than did PR. Both procedural and

Table 3. Concurrent Validity of Measures, Overall and by State.

Concurrent Validity Measures	ABS	AF	ACA	CoV	PR
Beginning					
Teacher rating	.56	.49	.42	.28	.21
	(.49, .60, .60)	(.36, .58, .54)	(.41, .63, .36)	(.16, .41, .29)	(.09^{ns}, .28, .24)
End					
Teacher rating	.55	.56	.56	.34	.13
	(.53, .64, .52)	(.47, .63, .65)	(.43, .66, .68)	(.23, .40, .37)	(.04^{ns}, .11*, .20)
Grade	.36	.37	.53	.30	.13
	(.28*, .39, .32)	(.20^{ns}, .44, .37)	(.33, .65, .65)	(.25, .31, .29)	(.01^{ns}, .15*, .16)
FANC	.54	.60	.59	.44	.22
	(.52,.46, .60)	(.55, .56, .71)	(.65, .51, .64)	(.32, .37, .48)	(.21, .18, .28)
IEOC	.48	.60	.64	.42	.13
	(.39, .43, .45)	(.47, .65, .59)	(.55, .38, .58)	(.26, .31, .40)	(.20, –.00^{ns}, .19)
State test	(.33, .46, .48)	(.43, .54, .48)	(.50, .59, .53)	(.40, .41, .40)	(.09^{ns}, .06^{ns}, .18)

Note. Coefficients of states are reported in parentheses (State A, State B, State C). Unless otherwise noted, all correlations significant at the $p < .01$ level. ns = not significant; ABS = Algebra Basic Skills; AF = Algebra Foundations; ACA = Algebra Content Analysis; CoV = Concept of Variable; PR = proportional reasoning.
*$p < .05$.

Table 4. Predictive Validity of Measures, Overall and by State.

Predictive Validity Measures	ABS	AF	ACA	CoV	PR
Teacher	.58	.53	.39	.27	.21
Rating	(.49, .60, .59)	(.43, .53, .63)	(.28, .49, .45)	(.15, .29, .34)	(.11*, .20, .27)
Semester	.37	.48	.37	.22	.13
1 grade	(.34, .48, .32)	(.30*, .44, .51)	(.25*, .54, .35)	(.07^{ns}, .26, .23)	(–.00^{ns}, .24, .12)
Semester	.36	.38	.36	.22	.10
2 grade	(.42, .31, .32)	(.16^{ns}, .43, .41)	(.18^{ns}, .47, .41)	(.13^{ns}, .31, .23)	(.03^{ns}, .26*, .10)
FANC	.50	.53	.47	.33	.24
	(.55, .41, .60)	(.48, .47, .62)	(.51, .31, .50)	(.31, .27, .36)	(.24, .19, .28)
IEOC	.45	.53	.56	.31	.18
	(.52,.35, .38)	(.48, .49, .54)	(.41, .32, .41)	(.29, .15*, .30)	(.12, .09^{ns}, .22)
State test	(.33, .46, .48)	(.43, .54, .48)	(.50, .59, .53)	(.33, .30, .32)	(.19, .16, .25)

Note. Coefficients of states are reported in parentheses (State A, State B, State C). Unless otherwise noted, all correlations significant at the $p < .01$ level. ns = not significant; ABS = Algebra Basic Skills; AF = Algebra Foundations; ACA = Algebra Content Analysis; CoV = Concept of Variable; PR = proportional reasoning.
*$p < .05$.

onceptual measures had the lowest coefficients when predicting end of course grades.

Discussion

The goal of the larger grant project within which this study took place was to expand the range of tools available to teachers for algebra screening and progress monitoring to include conceptual understanding. In this study, we examined data on two newly developed conceptual measures and three refined procedural measures. Reliability results, examined only for the procedural measures, were similar to previous findings using earlier versions of the measures. Concurrent criterion validity estimates in this study were stronger for the procedural measures than for the conceptual measures. A similar pattern was obtained for predictive validity. The validity evidence for the conceptual measures obtained in this study fell short of conventional expectations for technical adequacy of screening tools.

The results for the procedural measures can be compared to previous research. Espin et al. (2018) reviewed existing research on curriculum-based mathematics measures at the secondary level, including a summary of the 4-year program of research during which the procedural measures were developed. This study used revised versions of the procedural measures and different criterion measures and obtained findings similar to the earlier results. With respect to reliability, alternate-form coefficients were similar across all three measures, but test–retest coefficients were comparable or higher in this study. When examining the criterion

validity data for the full sample with respect to algebra outcome measures (IEOC, FANC), the results of the present study were similar or higher for concurrent validity (with the exception of ABS and IEOC), but similar or slightly lower for predictive validity across all three measures. Disaggregated data by state revealed considerable differences in coefficients across the measures.

Although technical adequacy results for the procedural measures were encouraging, we were disappointed by our findings for the conceptual measures. Given the importance of developing effective conceptual assessments, we considered multiple factors that may have influenced these results. First, our data collection schedule did not permit evaluation of the reliability of the conceptual measures. If the measures produce scores that are inconsistent or include a substantive level of error, examination of validity evidence for these measures will be affected. Second, we speculate that the differential results may reflect a mismatch between the demands of the conceptual items and the focus of instruction and assessment occurring in the participating schools. As part of the larger grant project, we conducted observations of classroom instruction across all three states during the same year in which these data were collected. Our observational data were gathered using an instrument adapted from the Reformed Teaching Observation Protocol (Sawada et al., 2002) to include coding for every question teachers posed to students during instruction. Our data revealed predominantly factual, lower-order questions, and a heavy emphasis on procedural, algorithmic instruction; we observed virtually no examples of teachers asking higher-level questions, engaging students in robust discussion, or emphasizing conceptual development. Third, we questioned whether items designed to assess conceptual understanding could fit within the constraints of a CBM-based assessment model.

We noted that the open-ended format of many of the items on the conceptual measures proved difficult for students. Large numbers of items were left blank or marked with "IDK" (I don't know) as the response. Focus group interviews with students during the development process were illuminating, as students reported unfamiliarity with items requiring explanations of their reasoning or responses other than an algorithmic solution.

The stronger results for the CoV measure relative to the PR measure were also of interest. We speculate that there may have been a closer alignment between the instructional focus and the content of the items given the context of an Algebra 1, 1A, or 1B course. In our discussions with teachers as part of the larger project (we met with teachers in each state three times each year of the project), they noted that proportions were not an explicit focus of the content in algebra courses and therefore they spent little time on those topics. They expressed frustration with students' entering mathematical proficiency levels, but felt obligated to focus their instruction on the designated curriculum objectives for the Algebra 1 course even when students demonstrated significant gaps in prerequisite skills related to PR.

The mean scores on the procedural measures were relatively low, given the simplicity of the items, and students' improvement over the course of the academic year was quite limited. These results call into question the general instructional quality and the effectiveness of existing instructional emphases on skills and procedures within algebra classrooms, as the current instruction does not appear to produce expected gains even on procedural and skill-based outcomes. Star et al. (2015) noted the need for evidence-based instruction and assessment of secondary students' procedural and conceptual knowledge, but our data suggest that current practices are not achieving these goals.

We note three limitations in this study. First, our data collection design precluded examination of the reliability of the conceptual measures. Future studies will address this limitation. Second, the challenges we experienced in obtaining student assent and parent consent within a high school population constrained our ability to conduct specific analyses of subgroups (e.g., free/reduced price meals, special education). The majority of the analyses reported here were focused on project-administered measures (procedural and conceptual assessments, FANC, IEOC), which had been determined to be exempt through the human subjects review process. Finally, we did observe some instances of lackluster participation among students. Although we are unable to estimate the specific effects on our results, we recognize the formidable challenge of engaging high school students in research tasks (even when they were administered by students' classroom teachers) when students were fully aware that their scores on the assessments did not affect their classroom grades.

Implications for Practice and Research

The results of the study have implications for both practitioners and researchers. Given the evidence of reliability and validity obtained in this study, the three procedural measures showed promise for screening purposes, pending further revisions and research on their diagnostic accuracy. This is important, as existing options for screening and progress monitoring in secondary mathematics are quite limited and often focus on general, or multitopic, mathematics (e.g., elementary and early middle grades), rather than the more content-specific courses (algebra, geometry) that are typical in later middle school and high school. Practitioners and their students benefit when the measures used to screen for potential risk status and subsequently monitor progress are more closely aligned with the instructional content. Practitioners are also informed by the multistate sample and the data for technical adequacy reported at

the state level. These results suggest state tests may not be similar in their relations with the algebra measures, although not all states released reports of test blueprints during these years of administration, so we cannot conclude what specific algebra areas or items the differences occurred in and why. Practitioners would be wise to examine the specific relations between the algebra measures and respective local or state outcome measures to confirm the strength of the predictive ability of the measures within the local or state context. Although the CCSS-M have given states consistency in their focus on conceptual mathematics learning (Phillips & Wong, 2010), the standards do not guide how they are to be implemented in practice, and there have been concerns about whether the CCSS-M will raise the quality of instruction, particularly for students with mathematics difficulties (Powell, Fuchs, & Fuchs, 2013).

Our findings with respect to the conceptual measures also hold implications for practitioners. As expectations for general education mathematics increasingly move beyond procedural and algorithmic understandings, it is important that students with disabilities, particularly those who will pursue postsecondary education, also develop conceptual understanding in algebra. Doing so will require that teachers serving students with disabilities both incorporate conceptual understanding goals into their instruction and have assessment tools by which to evaluate students' developing understandings of algebra. Although the present findings suggested these conceptual algebra measures did not demonstrate acceptable levels of validity and reliability for use within CBM, special educators should ensure they select assessments appropriate for their students that do assess conceptual learning. They should also design conceptually focused pedagogy that engages students in reasoning and communicating about their understandings of algebra.

The study has implications for researchers seeking to advance the options available for screening and progress monitoring in secondary mathematics. As the general education curriculum increasingly reflects learning outcomes that go beyond procedural competence, it is important to pursue efforts to develop assessment strategies that will provide educators with data about these goals. We are uncertain about whether our insufficient technical adequacy results for the conceptual measures reflect inherent flaws in the design of the measures or whether the results are indicative of a general instructional climate that did not develop and support students' understanding and ability to reason with and explain algebraic concepts. Future research should examine the technical adequacy of the measures (both procedural and conceptual) in classroom settings in which the instruction is explicitly focused on developing student understanding beyond a procedural emphasis.

We also recommend that future research explore additional refinement of the conceptual measures to increase the accessibility of the items for students who found the open-ended item formats overwhelming. Revisions should make the items more accessible across a range of student abilities while maintaining a focus on student reasoning. In our future work, we plan to use common student misconceptions identified from the constructed response items in this study to craft multiple-choice options that reflected multiple levels of conceptual understanding. We will also examine the scoring rubrics and consider awarding differential point values to different distractors in a multiple-choice item, with more points linked to distractors reflecting increasingly sophisticated understanding. In addition, an important area of future research will be in the diagnostic accuracy of the measures with respect to the external criteria, including states tests. We are hopeful that these efforts will increase the range of scores produced when the measures are used, as well as increasing the technical adequacy of the conceptual measures. Finally, future research should examine relations between performance on procedural and conceptual assessments relative to emerging high-stakes assessments (Partnership for Assessment of Readiness for College and Careers [PARCC] and Smarter Balanced, Herman & Linn, 2013) that are designed to be more reflective of the deeper understandings of mathematics imbedded in the CCSS-M.

Although our initial efforts to develop conceptual measures of algebra proficiency that would be suitable for screening did not produce desired results for technical adequacy, we remain optimistic that refinements to the measures may improve the evidence of their reliability and validity. The fact that skills and procedures are easier to measure in a technically adequate manner does not diminish the importance of and need for assessment tools that tap conceptual understanding. Procedural fluency in algebra is essential, but it must be balanced with a strong conceptual understanding (Alberti, 2012; Loveless, 2013). As the K-12 mathematics landscape shifts with the CCSS-M and more students enroll in algebra courses than ever before, technically adequate algebra assessment measures have potential to support teachers by identifying students with disabilities and others at risk for poor performance in algebra and evaluating the effects of efforts to provide intervention to improve their achievement.

Authors' Note

An earlier version of the paper was presented at the American Educational Research Association conference in April 2016.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R324A110262 to Iowa State University as part of the Algebra Screening and Progress Monitoring Project. The opinions expressed here are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

ORCID iD

William W. DeLeeuw https://orcid.org/0000-0003-2319-1404

Supplemental Material

Supplemental material for this article is available online at https://journals.sagepub.com/doi/suppl/10.1177/1534508419862025.

References

Adelman, C. (2006). *The toolbox revisited: Paths to degree completion from high school through college*. Washington, DC: U.S. Department of Education. Retrieved from http://www.ed.gov/rschstat/research/pubs/toolboxrevisit/toolbox.pdf

Alberti, S. (2012). Making the shifts. *Educational Leadership*, *70*, 24–27.

Bureau of Labor Statistics. (2015). *Employment projections*. Retrieved from http://www.bls.gov/emp/ep_chart_001.htm

Center for Public Education. (2013). *Detail on mathematics graduation requirements from public high schools, by state*. National School Boards Association. Retrieved from https://docplayer.net/7119182-Detail-on-mathematics-graduation-requirements-from-public-high-schools-by-state-as-of-june-5-2013.html

Choike, J. R. (2000). Teaching strategies for "algebra for all." *Mathematics Teacher*, *93*, 556–560.

Council for Exceptional Children Policy Insider. (2013, November 14), NAEP results show wide achievement gaps between children with, without disabilities (Blog). Retrieved from http://www.policyinsider.org/2013/11/naep-results-show-wide-achievement-gaps-between-students-with-without-disabilities.html

Deno, S. L. (1985). Curriculum-based measurement: The emerging alternative. *Exceptional Children*, *52*, 219–232. doi:10.1177/001440298505200303

Deno, S. L. (2003). Developments in curriculum-based measurement. *The Journal of Special Education*, *37*, 184–192. doi:10.1177/00224669030370030801

Dougherty, B., Bryant, D. P., Bryant, B. R., Darrough, R. L., & Pfannenstielf, K. H. (2015). Developing concepts and generalizations to build algebraic thinking: The reversibility, flexibility, and generalization approach. *Intervention in School and Clinic*, *50*, 273–281. doi:10.1177/1053451214560892

Dougherty, B., DeLeeuw, W., Foegen, A., Olson, J., Genareo, V., & Karaman, R. (2017). *Expert review of conceptual measures* (Technical Report 6). Ames: Algebra Screening and Progress Monitoring Project, School of Education, Iowa State University. Retrieved from https://www.education.iastate.edu/aspm/technical/

Espin, C., Chung, S., Foegen, A., & Campbell, H. (2018). Curriculum-based measurement for secondary students. In P. Pullen & M. Kennedy (Eds.), *Handbook on response to intervention and multi-tiered instruction* (p. XX). London, England: Routledge.

Foegen, A. (2000). Technical adequacy of general outcome measures for middle school mathematics. *Assessment for Effective Intervention*, *25*, 175–203. doi:10.1177/073724770002500301

Foegen, A. (2003). Project AAIMS: Algebra assessment and intervention—Meeting standards (Field-initiated Research Grant from the U.S. Department of Education, Office of Special Education Programs, PR/Award Number: H324C030060). Retrieved from https://www.education.iastate.edu/aaims/case_studies/District%20A%20Case%20Study.pdf

Foegen, A. (2008a). Algebra progress monitoring and interventions for students with learning disabilities. *Learning Disability Quarterly*, *31*, 65–78.

Foegen, A. (2008b). Progress monitoring in middle school mathematics: Options and issues. *Remedial and Special Education*, *29*, 195–207.

Foegen, A., & Deno, S. L. (2001). Identifying growth indicators for low-achieving students in middle school mathematics. *The Journal of Special Education*, *35*, 4–16. doi:10.1177/002246690103500102

Foegen, A., & Dougherty, B. (2010). *Algebra screening and progress monitoring* (Measurement (Goal 5) Award from the Institute for Education Sciences, U.S. Department of Education. Award Number: R324A110262). Retrieved from https://www.education.iastate.edu/wp-content/uploads/2018/05//Conceptual-Measures-User-Agreement.pdf

Foegen, A., Jiban, C., & Deno, S. (2007). Progress monitoring measures in mathematics: A review of the literature. *The Journal of Special Education*, *41*, 121–139. doi:10.1177/00224669070410020101

Fuchs, D., Fuchs, L. S., & Compton, D. L. (2012). Smart RTI: A next-generation approach to multilevel prevention. *Exceptional Children*, *78*, 263–279. doi:10.1177/001440291207800301

Fuchs, L. S., Fuchs, D., Hamlett, C. L., & Stecker, P. M. (1991). Effects of curriculum-based measurement and consultation on teacher planning and student achievement in mathematics operations. *American Educational Research Journal*, *28*, 617–641. doi:10.3102/00028312028003617

Herman, J., & Linn, R. (2013). *On the road to assessing deeper learning: The status of smarter balanced and PARCC assessment consortia* (CRESST Report 823). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing. Available from https://cresst.org/publications

Iowa Testing Programs. (2010). *Iowa End-of-Course assessment programs: Spring 2010 results*. Retrieved from http://itp.education.uiowa.edu/ieoc/documents/IowaEOCAssessmentsSpring10Results.pdf

Kilpatrick, J., Swafford, J., & Findell, B. (Eds.). (2001). *Adding it up: Helping children learn mathematics*. Washington, DC: National Academy Press. doi:10.17226/9822

Krutetskii, V. A. (1976). *The psychology of mathematical abilities in schoolchildren* (J. Teller, Ed., J. Kilpatrick & I. Wirszup, Trans.). Chicago, IL: University of Chicago Press.

Loveless, T. (2013). *The algebra imperative: Assessing algebra in a national and international context*. Washington, DC: Brown Center on Education Policy at Brookings.

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher*, *18*(2), 5–11. doi:10.3102/0013189x018002005

National Governors Association Center for Best Practices, Council of Chief State School Officers. (2010). *Common core state standards mathematics*. Washington, DC: Author.

Phillips, V., & Wong, C. (2010). Tying together the common core of standards, instruction, and assessments. *Phi Delta Kappan*, *91*(5), 37–42.

Porter, A., McMaken, J., Hwang, J., & Yang, R. (2011). Common core standards the new US intended curriculum. *Educational Researcher*, *40*, 103–116. doi.org/10.3102/0013189x11405038

Powell, S. R., Fuchs, L. S., & Fuchs, D. (2013). Reaching the mountaintop: Addressing the common core standards in mathematics for students with mathematics difficulties. *Learning Disabilities Research & Practice*, *28*, 38–48.

Rachlin, S. (1998). Learning to see the wind. *Mathematics Teaching in the Middle School*, *3*, 470–473.

Rittle-Johnson, B., Siegler, R., & Alibali, M. W. (1999). Conceptual and procedural knowledge of mathematics: Does one lead to the other? *Journal of Educational Psychology*, *91*, 175–189. doi:10.1037//0022-0663.91.1.175

Sawada, D., Piburn, M. D., Judson, E., Turley, J., Falconer, K., Benford, R., & Bloom, I. (2002). Measuring reform practices in science and mathematics classrooms: The reformed teaching observation protocol. *School Science and Mathematics*, *102*, 245–253. doi:10.1111/ssm.2002.102.issue-6

Shapiro, E. S., Keller, M. A., Lutz, J. G., Santoro, L. E., & Hintze, J. M. (2006). Curriculum-based measures and performance on state assessment and standardized tests: Reading and math performance in Pennsylvania. *Journal of Psychoeducational Assessment*, *24*, 19–35. doi:10.1177/0734282905285237

Spielhagen, F. R. (2006). Closing the achievement gap in math: The long-term effects of eighth-grade algebra. *Journal of Advanced Academics*, *18*, 34–59. doi:10.4219/jaa-2006-344

Star, J. R. (2005). Reconceptualizing procedural knowledge. *Journal for Research in Mathematics Education*, *36*, 404–411.

Star, J. R., Foegen, A., Larson, M. R., McCallum, W. G., Porath, J., & Zbiek, R. M. (2015). *Teaching strategies for improving algebra knowledge in middle and high school students. Educator's practice guide* (NCEE 2015-4010). What Works Clearinghouse. Retrieved from https://ies.ed.gov/ncee/wwc/docs/practiceguide/wwc_algebra_040715.pdf

Stecker, P. M., Fuchs, L. S., & Fuchs, D. (2005). Using curriculum-based measurement to improve student achievement: Review of research. *Psychology in the Schools*, *42*, 795–819. doi:10.1002/pits.20113

Taylor, R. (1990). Interpretation of the correlation coefficient: A basic review. *Journal of Diagnostic Medical Sonography*, *6*, 35–39.