

Validity Evidence Based on the Consequences of Interim Assessments

Angela E. Broaddus, Laura M.B. Kramer, and Neal Kingston

Center for Educational Testing and Evaluation, Lawrence, Kansas

Author Note

Angela E. Broaddus, Center for Educational Testing and Evaluation, University of Kansas; Laura M.B. Kramer, Center for Educational Testing and Evaluation, University of Kansas; Neal Kingston, Center for Educational Testing and Evaluation, University of Kansas.

This research was conducted in support of the Kansas Assessment Program.

Correspondence concerning this article should be addressed to Angela Broaddus, Center for Educational Testing and Evaluation, University of Kansas, Lawrence, KS 66047.

E-mail: broadus@ku.edu

Validity Evidence Based on the Consequences of Interim Assessments

The Kansas interim assessments were designed to be managed predominantly by classroom teachers. The summative assessment program in Kansas is most often managed by district and building level administrators, but there was an interest in providing teachers with a tool for gauging student preparedness for the summative assessments, which was the main purpose of the interim assessment. This paper presents the issues related to the implementation of the interim assessment in addition to a discussion of the implications for the future of this program.

Kansas State Department of Education (KSDE) representatives and CETE staff examined a wide range of evidence throughout the implementation of the interim assessment. These data included descriptive information about the students who took interim assessments and their scores, the length of time students spent taking interim assessments, i.e. test times, and educator surveys, which included teachers' and administrators' feedback about their experiences with and suggestions for interim assessments.

Implementation Dates

The interim assessment was administered in 137 of the 317 school districts in Kansas, and 91,716 tests were completed. Interim assessments were available during three administration windows: Fall 1 (September 15, 2010, to October 22, 2010); Fall 2 (October 25, 2010, to December 31, 2010); and Winter (January 1, 2011, to February 28, 2011). The greatest number of tests, 37,501, was administered during the Fall 2 window. The interim assessment was used slightly more often at higher grade levels than at lower grade levels during all three windows. The percent of students per grade who took the test is displayed in Table 1.

Table 1
Student Participation by Grade and Test Administration Window

Grade	Test Administration	Percent of Population
3	Fall 1	11%
3	Fall 2	15%
3	Winter	10%
4	Fall 1	11%
4	Fall 2	16%
4	Winter	11%
5	Fall 1	12%
5	Fall 2	18%
5	Winter	12%
6	Fall 1	13%
6	Fall 2	19%
6	Winter	13%
7	Fall 1	15%
7	Fall 2	21%
7	Winter	18%
8	Fall 1	14%
8	Fall 2	19%
8	Winter	15%

Student Descriptive Information

The demographic characteristics of the students who took interim assessments are somewhat similar to those of the entire student population in Kansas. Comparisons of the percent of students in each demographic category for interim and prior year summative assessments indicate that slightly more students from traditionally lower-performing subgroups took interim assessments. One possible explanation for this finding could be because the interim assessment

was provided to all Kansas schools at no cost. Thus, schools operating on strained budgets, such as those in areas with depressed economies, were able to use these assessments without placing any additional strain on their finances. Many districts operating on more substantial budgets indicated that they purchased commercially produced tests to gauge student preparedness for the Kansas state assessments, which was the primary purpose of the interim assessment. In many cases, these districts did not implement the interim assessment.

Table 2 displays the percent of students in one grade by demographic category for each interim assessment administration window and for prior year summative assessments. The percentages in this table show that on average, more students in lower socio-economic categories took interim assessments when compared to the entire population that took the prior year’s summative assessment. Likewise, on average, more Native American and Hispanic students took interim assessments when compared to the entire population that took the prior year’s summative assessment.

Table 2
Percent of Students by Demographic Categories for Interim and Summative Assessments for Eighth Grade

Test	Gender		Lunch Status			Race and Ethnicity					
	Female	Male	Full	Reduced	Free	Native American	Asian	Black	Hispanic	White	Multi
Interim Fall 1	52	48	53	12	35	2	2	6	16	69	5
Interim Fall 2	52	48	56	12	32	2	2	5	17	69	5
Interim Winter	51	49	52	12	35	2	2	5	22	64	4
Prior Year Summative	50	50	60	10	30	1	3	7	13	72	4

Student scores. Students who took the interim assessment received scores on the same scale as the scale of the same grade’s summative assessment scores. Therefore, students, teachers, and other stakeholders could use these scores to estimate student proficiency regarding the tested curriculum after each test administration. Students are placed into proficiency categories according to their summative assessment scores each year. The five categories, from highest to lowest achievement, are “exemplary,” “exceeds standards,” “meets standards,” “approaches standards,” and “academic warning.” For each grade and subject, different cut scores define these five categories.

Student test scores can be used to inform a variety of judgments and decisions. At the state level, CETE and KSDE observed that not only did interim assessment scores improve throughout the year, but also a relatively large proportion of students “met standards” during the first interim assessment administration. Table 3 lists the percent of students per grade level that earned scores equal to or higher than the “meets standards” cut score during each interim assessment administration. This evidence suggests that a large number of students acquired acceptable knowledge of the tested curriculum early in the school year. This predicament should cause educators to consider whether these students should continue to be taught and required to practice the same concepts and skills throughout the year in order to earn higher scores or whether their curriculum should be adjusted or enriched with more advanced material.

Table 3
Percent of Students that “Met Standards” for each Mathematics Interim Assessment Administration

Grade	Fall 1	Fall 2	Winter
3	33	45	44
4	46	57	54
5	51	53	54
6	44	49	53
7	43	49	48
8	43	50	50

Student Test Times

The length of time students spent taking interim assessments was much longer than CETE anticipated. The test, which contained between 30 and 42 items depending on the grade level, was advertised to require only one average-length class period to administer. However, teachers notified CETE very early that students were spending up to two hours or more testing.

In response to teacher concern and frustration over long test times, CETE modified the testing software early during the first interim assessment administration window. Originally, the software was designed under the assumption that students would complete an entire interim assessment in one sitting. Thus, no feature existed that permitted the student to log out and return later. After teachers explained the difficulties, CETE quickly modified the software to allow them to reactivate students who left a test session before completion. Although this solution met an immediate need, a more refined solution was implemented for the second year of the assessments: beginning in Fall 2011, students will have an option at the end of each testlet to continue testing or to save and return later. Educators in Kansas and CETE staff continued to investigate possible causes of lengthy test times. One possible contributor was the reading load

associated with the items on the mathematics interim assessments. Teachers said that for younger students, particularly at the beginning of the school year, the reading demand caused some of the long test times. Third-grade students are still learning to read throughout the school year, and the test items were developed to be parallel to those used for summative tests administered at the end of the school year. Thus, the vocabulary and text complexity of some test items may have caused unanticipated difficulty for some students.

Other possible factors contributing to the long test times were less obvious. One possibility may be because the interim assessment measured all tested standards each time, and students may not have received instruction on some topics when they took these assessments. Teachers likely advised their students to react differently to material that looked unfamiliar. To investigate the nature of these differences, CETE staff asked teachers in a survey to select the description that best captured their students' testing conditions. Table 4 shows the text for the three response options, along with the percent of respondents that chose each option. These data suggest that many students were required to work very hard to answer test items for which they may not have received instruction, which may explain some of the long test times.

Table 4

Response Options and Response Percentages to the Survey Question "Describe the conditions under which your students took interim assessments."

Response Option	Percent of Teachers
I encouraged my students to do their best work with each question and not worry too much about what they didn't recognize.	44
My students raised their hands when they had questions about the material presented in the test items, and I went to their desks to help them with these items.	3
My students diligently worked through all problems regardless of whether they were familiar with the material.	53

Long test times coupled with analysis of the interim assessment's reliability eventually led to a modification of the original test design. CETE staff determined that shortening each interim assessment would not cause significant reductions in score reliability. Therefore, each interim assessment was reconfigured to contain just two items per tested standard, which resulted in test lengths ranging from 24 to 30 items, down from 32 to 42 items, depending on the grade level. The items were arranged in two testlets rather than three, thereby meeting a teacher preference for these tests to be administered in less total time and fewer sittings. The reliabilities of the interim assessments for each grade and for the two different test designs are shown in Table 5. These data show that the shorter tests that will be administered in 2011-2012 should produce scale scores with close to the same reliability as the longer tests that were administered in 2010-2011.

Table 5
Interim Assessment Test Lengths and Reliabilities for Two Administration Years

Grade	Number of Items (2010-2011)	Reliability (2010-2011)	Number of Items (2011-2012)	Reliability (2011-2012)
3	32	0.87	24	0.84
4	33	0.84	28	0.82
5	35	0.87	30	0.84
6	38	0.89	28	0.86
7	38	0.89	30	0.87
8	42	0.91	28	0.87

Interim Assessment Compared to Formative Assessment Tools

In order to introduce the interim assessment and situate it among the other assessment tools provided by CETE to Kansas educators, two professional development modules were assembled and delivered via webinar to educators across the state. One goal of the training was to raise administrator and teacher awareness of different types of assessments and how they should be implemented to maximize the effectiveness of teaching and learning. A second goal was to introduce administrators and teachers to the design of the interim assessment and to describe what the anticipated student testing experience would be. A third goal was to explain how to use the teacher interface webpage to organize classes, administer tests, and view test results.

It became evident early on that educators at all levels had difficulty distinguishing formative assessments from interim assessments. This confusion stemmed from at least two hypothesized sources. First, evidence suggested that teachers had been administering the greatest numbers of formative assessments shortly before administering summative assessments. Data supporting this claim were the number of formative test sessions recorded by the KCA management tool for each day during the school year for two years. The pattern of test sessions and dates implied that teachers were using the formative assessment tools as practice tests rather

than embedding them into their instructional activities. Second, teachers expressed great disappointment that the interim assessment reports did not provide student performance for each tested standard. They were frustrated particularly by the amount of instructional time needed to administer interim assessments, and that the reports did not provide what the teachers deemed instructionally relevant information. Teachers preferred to receive student item responses, like those provided in the formative assessment reports, in order to inform their instructional decisions.

For several years before the introduction of the interim assessment, Kansas educators have enjoyed access to thousands of practice test items from the Center for Educational Testing and Evaluation (CETE) website. Teacher survey responses demonstrate that tools from the website, commonly referred to as formative assessments and items, are routinely included as part of instructional activities. For example, 82% of teachers surveyed had used classroom-created formative assessment materials, and 84% had administered formative assessment materials from the CETE website. The website enables teachers to not only use what is already available, but to create and upload additional items. As for testing students, teachers can print tests from the website for their students, or they can administer formative assessments online using the Kansas Computerized Assessment (KCA) software, the same software used to administer the interim and summative assessments. If teachers use the KCA software, then they receive detailed reports that include every student's response to every test item. These reports contain one row per student, and every row includes an entry for each test item. Correct responses are noted by the number "1," and incorrect responses by the answer choice – the letter A, B, C, or D – that the student chose instead. An example of a formative assessment report is shown in Figure 1; test and student names have been removed. The teacher can also click on a link and review each item's

content and answer choices, thereby adding meaning to the student responses. Many teachers have grown to depend on these reports and use them to make instructional decisions for particular students.

1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	20/20 100.0%	
1	1	1	1	B	C	1	1	D	1	1	1	1	1	C	1	1	1	D	A	14/20 70.0%	
1	1	1	1	1	1	1	1	1	1	1	1	1	1	B	B	B	D	B	B	1	14/20 70.0%
B	1	1	A	1	1	A	1	B	C	D	C	B	C	A	1	A	D	D	1	7/20 35.0%	
1	1	1	1	1	1	1	1	1	1	1	1	1	1	C	1	1	1	D	A	17/20 85.0%	
1	1	1	1	1	1	1	1	A	B	A	1	1	B	B	C	1	A	1	D	B	11/20 55.0%
A	1	1	1	B	C	1	1	1	B	D	1	B	C	B	1	1	B	D	A	9/20 45.0%	
B	1	1	1	1	1	1	A	1	A	1	1	1	B	B	1	1	1	C	1	A	13/20 65.0%

Figure 1. Class Report of Formative Assessment Item-level Responses.

Professional development modules provided information on interim assessment implementation and score interpretation. While teachers quickly learned how to use the management tools and interpret individual student interim assessment scores, they struggled to understand the parts of the score reports that described class score distributions. They also continued to question the utility of these assessments for their classroom use. Many conversations with teachers combined with survey data support the possibility that many teachers who used the interim assessments did not receive adequate professional development.

Face-to-face professional development sessions were conducted during the statewide assessment conference and at two state-level professional teacher meetings. However, most attendees were administrators rather than classroom teachers. Professional development webinars were better attended by teachers; however, in many cases, grade-level or building-level representatives attended them and then conducted local sessions on their own to classroom teachers. Besides conference sessions, teacher meetings and webinars, the assessment’s user’s manual was available on the CETE website throughout the entire school year. The Interim Assessment Examiner’s Manual describes the purpose of design of the Kansas interim assessment and contains detailed instructions about how to enroll students, administer tests, retrieve reports, and interpret results. Table 6 shows the percentage of teachers who received professional development and who claimed to have read the user’s manual. These data support the interpretation that teachers required more information or professional development about the purpose and utility of the interim assessments and their scores.

Table 6
Teacher Response to Survey Questions Pertaining to Professional Development

Survey Question	“YES” Responses (percent)	“NO” Responses (percent)
Did you receive professional development about the interim assessment?	51	49
Did you download and read the Interim Assessment Examiner’s Manual?	71	29

Score reports for interim assessments were provided to teachers within one day of test administrations. Both individual student scores and class score distribution information were provided in the online reports. The class roster report contained a scale score for each student as

well as an instructed standard score for each student when the teacher checked the standards that were taught prior to the test administration. A screen shot of a class roster report is shown in Figure 2. In addition to the class roster report, teachers were provided score distribution information in the form of a box-and-whisker plot to represent each class and different test administration. Score distribution information was intended to illustrate how well each class was progressing toward mastery of the tested curriculum and to inform

Class Roster Report
Kansas Mathematics Interim Assessment Scale Scores

Student Name	Kansas Student ID	Grade	Instructed Indicators Scale Score Fall 1	Instructed Indicators Scale Score Fall 2	Interim Scale Score Fall 1	Interim Scale Score Fall 2	Interim Scale Score Winter
Blunt, Derek	111999005	5	88	82	70	82	N/A
Darling, George	111999007	5	N/A	98	N/A	98	N/A
Darling, Mary	111999011	5	N/A	80	N/A	83	N/A
Dear, Jim	111999009	5	N/A	78	N/A	78	N/A
Duck, April	111999003	5	N/A	89	N/A	88	N/A
Flaversham, Olivia	111999013	5	98	89	64	78	N/A
Gander, Gladstone	111999008	5	30	80	30	61	N/A
Mouse, Timothy	111999015	5	98	58	98	60	N/A
Mulan, Fa	111999006	5	N/A	98	N/A	76	N/A
Porter, Jane	111999010	5	37	75	58	75	N/A
Rabbit, Roger	111999014	5	88	65	82	51	N/A
Radcliffe, Anita	111999002	5	88	89	89	90	N/A
Slade, Amos	111999001	5	78	10	89	10	N/A
Tremaine, Nancy	111999012	5	78	53	63	20	N/A

[View all Student Reports >>](#)

Figure 2. Class Report of Interim Assessment Scores.

class-level instructional decisions. A screen shot of a class distribution plot is shown in Figure 3. However, when surveyed about the report’s usefulness, teachers favored student-level scores and reports over grouped information. Table 7 displays teacher responses to a question about the

report’s features. Response data support the interpretation that teachers remain focused on individual student performance and are less interested in group-level information that could inform class-level planning and decisions or be used during reflection of their professional practice.

Although originally developed predominantly for teachers and students, it became evident throughout the year that teachers required assistance and support for effective interim assessment implementation and valid score interpretation. Personal communication with various district leaders described an implementation model that promoted effective and efficient interim assessment administrations, which in turn provided teachers with instructionally useful information. In these implementations, teachers interested in

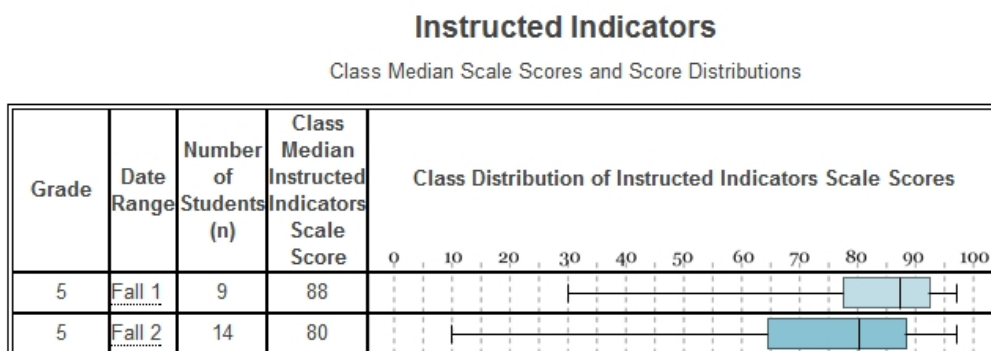
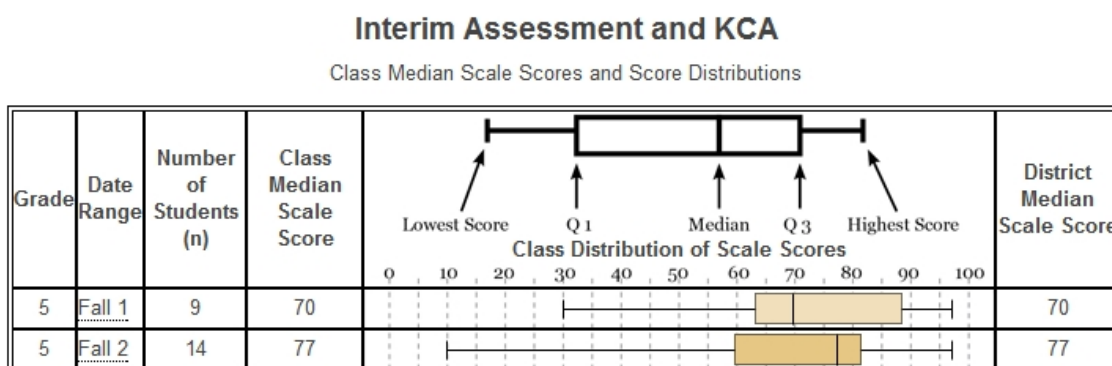


Figure 3. Class Report of Interim Assessment Scores.

Table 7

Teacher Responses to “Please rate the usefulness of these features of the interim assessment teacher report.”

Report Feature	Very Useless (percent)	Useless (percent)	Neutral (percent)	Useful (percent)	Very useful (percent)
Student scale scores	9	12	29	41	9
Instructed standard scores	10	11	23	43	14
Student level reports	10	12	24	42	12
Box and whisker plots	13	18	41	23	4

using interim assessments met with district leadership to receive focused professional development about the interim assessment’s purpose and how to use test scores to influence instructional decisions. Teachers were advised to administer the interim assessment to all students, and to use the scores to determine which students required more focused formative assessment. In this way teachers did not necessarily administer more assessments to all students, which was a common complaint, but rather, focused their efforts to identify the misconceptions and skill deficiencies of the students who struggled or were at risk of failure. These teachers effectively implemented interim assessments for screening and for gathering class-level information, and they followed up with formative assessments to target individual students and particular concepts and skills.

The implementation of interim assessments in Kansas illustrates the importance of educating test users about the purposes and best practices recommended for different types of assessments. Teachers and administrators appreciate the value of using assessment data to inform instructional and programmatic decisions aimed at improving teaching and student learning. With so many assessments available to educators, it’s imperative that test developers provide accurate descriptions of tests, the tests’ purposes, and recommendations on test implementation in educational settings. Furthermore, teachers and school administrators must improve their

understanding of different types of assessments, their varied purposes, and their best applications. These two communities must continue to gain knowledge of each other in order to advance the quality and effectiveness of assessments in schools.