Validity Evidence Based on Interim Assessment Content

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

Center for Educational Testing and Evaluation, Lawrence, Kansas

Author note

Laura M.B. Kramer, Center for Educational Testing and Evaluation, University of Kansas; Neal Kingston, Center for Educational Testing and Evaluation, University of Kansas; Angela E. Broaddus, 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 Laura M.B. Kramer, Center for Educational Testing and Evaluation, University of Kansas, Lawrence, KS 66047. E-mail: Laura.Kramer@ku.edu

Validity Evidence Based on Interim Assessment Content

With the continuing emphasis on using the results of large-scale, high-stakes academic achievement tests for accountability at federal, state, and local levels, many states and school districts have invested in the development of comprehensive assessment systems. Educational assessment programs should include components that effectively and validly inform instructional, programmatic, and policy decisions for the purpose of improving teaching and student learning. A comprehensive assessment system includes a cohesive array of achievement assessment instruments, procedures, and practices encompassing summative, interim or benchmark, and formative tools.

Because of their prominence in educational, policy, and social decisions, summative tests undergo a great deal of scrutiny, and a great deal of attention is paid to their technical attributes. Formative-assessment processes have, within the past decade, also gained prominence in the research literature as well as in teaching practice. Somewhat famously, these two assessment types have been described as representing, respectively, assessment OF learning, and assessment FOR learning (Stiggins, 2002). Later authors recharacterized the process of formative assessment representing assessment AS learning (Earl, 2003). To improve the value of test-based information and prevent misuse of data, it is imperative to state clearly the purposes of each component of a comprehensive assessment system.

Elements of a Comprehensive Assessment System

Summative assessments are typically conceived as tests given at the end of a school year or instructional period and often are aggregated in accountability systems. Results from these tests are used to measure the mastery (or non-mastery) of the expectations expressed in a state's

content standards and provide a deeper and broader assessment of what a student has learned. While summative assessment usually occurs too infrequently and too late to directly impact student learning, these results also are sometimes used to influence policy decisions or identify areas for additional professional development at the state level. In this way, the data produced can support in-depth review of curriculum and instruction to improve the learning of future students. Summative assessment data can effectively inform program strengths and illuminate areas where professional development might positively influence teaching and learning.

Summative test results are also often used in high-stakes decisions. For example, since they represent the culmination of a student's experience in a grade or course, these test scores frequently contribute to student promotion or graduation requirements, or to a student's final grade. Summative assessment results are also used in school and teacher evaluations; in teacher, principal, or superintendent retention or promotion decisions; and in some states, as the basis to award teacher bonuses. And of course, with state and federal accountability models, school and district test results determine sanctions (rarely rewards) in terms of state and federal funding as well as school or district takeovers or restructuring. The sometimes excessive reliance on summative test scores can even influence property values and the decisions of business and industry to locate in a certain jurisdiction.

Formative assessment provides information focused on the direct improvement of student learning either by providing students with information they need to regulate their own learning or teachers with information they need to modify instruction while they are teaching. As such, formative assessment must focus on a narrow set of related concepts or skills – either a single lesson or a small unit of study. Feedback from formative assessments must identify sources of misconceptions or skill deficiencies in such a way as to assist students in correcting their

understanding. Formative assessment thereby impacts learning on a frequent and immediate basis.

However, "formative assessment" is rather a misnomer, leading one to think of a single assessment event or series of events. Instead, formative assessment should be envisioned as a process (Popham, 2011). The Socratic ideal is a teacher actively assessing students every minute of instructional time with ongoing questioning to gauge students' comprehension and needs, and using this feedback to make appropriate changes to instruction moment-by-moment. More practically enacted, formative assessment is an ongoing questioning and evaluating process that may be punctuated by assessment events; teachers then use the evidence gained from the process to inform instructional changes day-by-day or week-by-week. Formative-assessment processes are not high-stakes, not reported to the federal government or the media, and most certainly not a single "test." Formative-assessment events do not even need to be graded. Formative-assessment processes are immediate, instructionally actionable, and relevant to both students and teachers to improve learning.

The newest vision of assessment to join the comprehensive assessment system falls between The Test (when summative events are spoken of, one can generally infer the capital letters), and good instructional practices as reflected in the formative-assessment process. This "middle of the road" type of assessment provides information that focuses on either identifying student needs for special program placement or the need to modify existing educational programs. While this type of assessment does not provide the content focus or descriptive feedback of formative assessment, it yields information that can be used to appropriately serve students' needs and to adjust instructional programs. This assessment type is helpful for identifying students who are not on track to make sufficient progress without additional

programmatic assistance or the need for within academic year modification of the existing program of study. This type of assessment impacts educational programs and student learning on a periodic basis – perhaps two to four times a year.

Depending on how this type of test is implemented, and to some extent local preference, these assessments may be referred to as benchmark or interim assessments; there is no "official" definition, and in some quarters these terms are used interchangeably along with progress monitoring, medium-scale assessment, and other terms to distinguish these assessments from summative and formative (Herman, Osmundson, & Dietel, 2010). These assessments are discrete events, given under fairly standardized test administration conditions, and often mimic the summative assessments in terms of item types and format. Depending on the origins of the mandate for their use, these assessments may have stakes attached for students (for example, a midterm exam) or teachers (evaluating instructional effectiveness). Common uses for these assessments are to aid in developing or modifying curriculum and planning instruction, and to communicate to both teachers and students how expectations for learning will be measured. These assessments may also serve additional purposes, such as informing programmatic or instructional decisions, or to predict eventual performance on the summative test.

For the current purpose, the difference posited between a "benchmark" test and an "interim" test is a matter of content. Here, interim assessment will refer to a test that covers the entire set of content standards expected to be mastered. By contrast, a benchmark assessment would cover a defined subset of the content; for example, a quarterly benchmark assessment would cover the 25% of the material in each test, presumably the content that had been covered during that quarter (or an accumulation of the content that had been covered up to that point).

This distinction is important to make in light of the purposes for which such an assessment might be used. For an assessment that is used to make predictions about a student's eventual performance on a summative test, a test or series of tests that cover discrete portions of the material might not produce the most accurate prediction of full-scale performance, particularly in content areas where early material provides basic content knowledge on which the later material builds in increasingly complex or abstract ways. Additionally, this distinction is important to consider given the scale of a test administration of this nature. A benchmark test, as described here, presupposes that all students taking the assessment have been instructed on the same content during the time frame proscribed by the periodic assessment. Within a classroom this makes sense, as all students have the same teacher; to some extent within a school or district, an explicit curriculum and pacing guide could heighten the validity of an evaluation of instructional or program efficacy, and allow for timely changes to be made to benefit student learning. However, on a larger scale such as a statewide administration, particularly in a "local control" state, often a constrained approach to curriculum will meet with resistance, if not complete rebellion, and cries of "teaching to the test" and "crippling creativity in the classroom."

The strengths of an interim assessment, as defined here, are precisely the opposite of the weaknesses of a benchmark model. Teachers are free to cover the content in the order and at the pace they deem most appropriate, allowing innovation and individualization. As the school year progresses, teachers and administrators can see the progress the students are making toward mastery of the entirety of the content standards, not just fragments; this model also lends itself quite well to predictive modeling of student attainment on the summative assessment. Likewise, the strengths of the benchmark model are weaknesses of the interim model. Because there is not just one course charted through the content standards, programmatic evaluations or judgments of

teacher efficacy are more difficult to make. Teachers and students both experience frustration and anxiety when they are being evaluated on content that has not yet been covered in class. Even when administrative and interpretive materials are provided that explain the purpose of the assessment, and that students will be exposed to material that has not yet been taught, the refrain is that such a test is “too difficult,” “unfair,” and “demoralizing.”

Another type of assessment, not often considered as a part of the comprehensive pantheon, will be mentioned only briefly here. Practice tests help to insure that a student's lack of experience with a testing program does not interfere with accurate educational measurement. Inaccurate tests scores can derive from confusion or anxiety and produce concomitant negative effects on accountability test scores. In a comprehensive testing system where students have experience with formative or interim/benchmark assessments that have similar interfaces and item types as are on the summative assessment, practice tests should be unnecessary. Nonetheless, while many educators use tests intended for practice, others use tests intended to be used for formative or interim purposes as practice tests. Thus, it is the responsibility of both test publishers and those who mandate assessments within a state to ensure that the purpose of each type of test in a comprehensive assessment system is clearly communicated to the users of the tests.

The Kansas Interim Assessment: Purpose and Design

When Kansas was investigating adding this “middle” component to its comprehensive assessment system, an advisory panel of district leaders met with the state department of education to determine what decisions, features, and outcomes would be desirable. The department and stakeholders wanted a test that would be predictive of the summative test, would

be shorter than the summative event, and that would not require a common curriculum / pacing guide across the state. Some districts were already using commercially available products, but there were concerns as to the alignment between the commercially available products and the Kansas content standards, as well as issues of comparability among the many commercial products and the Kansas summative test. Additionally, of course commercially available products can be costly, and in the current economic climate, having a single medium-scale component, aligned to the standards and developed in parallel to the summative test, and provided for free to the districts was a very attractive option to the districts. The assessment model eventually decided upon was a test that would cover all of the assessed content standards, could be administered up to three times a year, and would be testlet-adaptive.

In Kansas, the state's content standards are organized into Standards (most general level), Benchmarks, and Indicators (most specific level). Although teachers are expected to cover the entire content standards during classroom instruction, within a grade and subject the indicators build upon each other so that the expectation for mastery is reflected in higher-order indicators that may integrate many of the prerequisite or precursor indicators. Thus, rather than testing all indicators, Kansas identifies "Tested Indicators," success on which presupposes mastery of the prerequisite skills and successful integration of the precursor indicators. There are generally 12 to 15 Tested Indicators for a grade and subject, and each is tested with a minimum of 4 items on the summative test. The summative test is thus able to provide student-level reports at the indicator level in addition to an overall score.

The summative tests are computer-delivered to over 99.5% of the tested population for both the general assessment (in English or Spanish for subjects where a Spanish version is allowed) and the modified assessment (AA-MAS, or "2%" test). Paper forms are available for

certain students with disabilities, such as Braille or large-print beyond a size that the computer can deliver. The tests are linear (non-adaptive), and have many built-in tools and accommodations such as read-aloud (except for reading comprehension passages; the items can be read aloud), highlighter, and calculator (on items where calculator is allowed).

Because there is not a state-driven curriculum or pacing guide, each interim assessment was developed to cover all the Tested Indicators. In order to be most strongly predictive of the summative tests, the interim assessment blueprints covered all the Tested Indicators in the same proportion as the summative test. The summative tests are administered over multiple days (typically three, as the tests are divided into three sections), but the desire was to have the interim assessment fit into a single class period. And in order to be able to evaluate their own instructional effectiveness, teachers are able to select the indicators on which they had delivered instruction to their students. The expectation being, of course, that class performance on these "Instructed Indicators" should be higher than on the indicators which had not yet been taught.

Like the Kansas summative assessment, the Kansas interim assessment is delivered on the computer, using the same teacher and administrator management tools as well as the same student interface. Limited test versions and somewhat fewer accommodations are currently available to students taking the interim. For example, due to the adaptive nature of the test, paper forms and scripted read-alouds are not available as accommodations. The interim assessment is not available in Spanish or Braille, nor is there a separate interim assessment for students who would normally take the modified assessment. However, most of the test interface tools are still available, such as the highlighter and calculator (for items where a calculator is allowed).

The original interim assessment design incorporated three stages, echoing the three sections of the summative test. The first stage is a routing test, which had one item for each Tested Indicator. Depending on the ability estimate for a student, the student was routed to a second stage of either harder items, easier items, or items of roughly the same difficulty. This second stage contained one or two items from about half of the Tested Indicators. Again depending on the student ability estimate, students were routed to the third stage, which again contained one or two items from about half of the Tested Indicators. A schematic of the interim assessment is included as Figure 1. Each testlet within a stage was parallel in content to the others in that stage; however, of course, they differed in terms of difficulty.

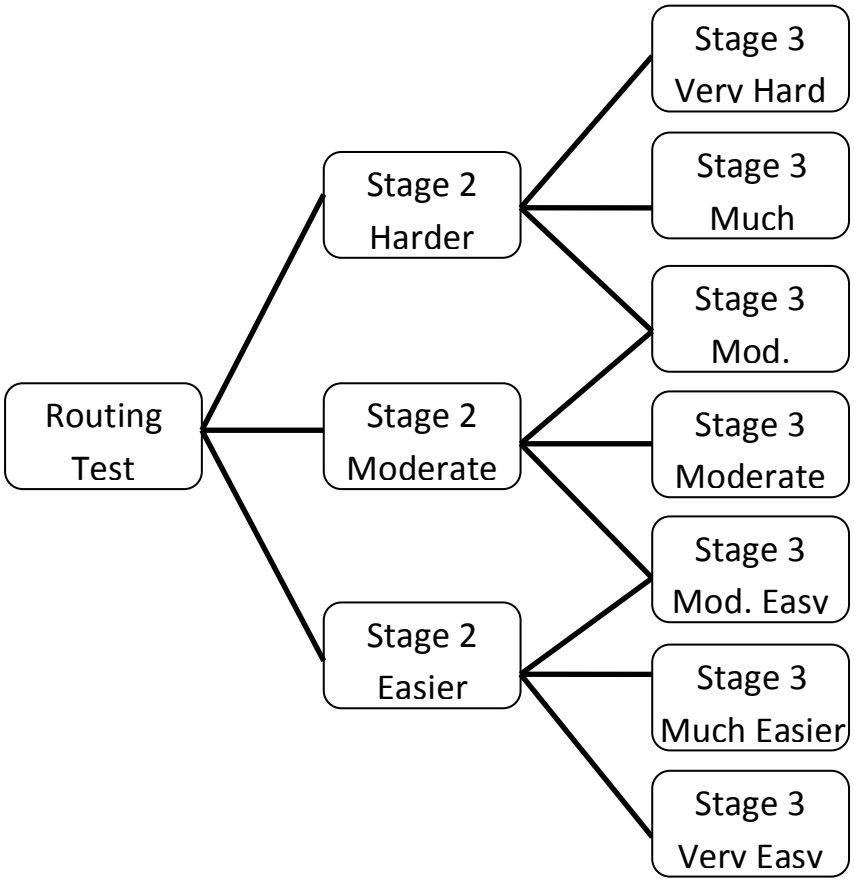


Figure 1. Three-Stage Adaptive Model.

Since the goal was to have a test that could effectively and efficiently predict performance on the summative test, but take one class period instead of three, an adaptive test seemed the most appropriate format. Among the many advantages of an adaptive design, a primary consideration was efficiency. An adaptive test, even a fixed-length, multi-stage adaptive test, is theoretically able to provide a more reliable estimate of student ability than a static form of the same length. Providing the adaptive algorithm is working correctly, the test engine will not present students with test items that are either too difficult or too easy, instead presenting students with items that should target the student's ability more closely to provide maximum information about the student's mastery of the content.

To ensure a strong predictive component, the interim test needed to reflect the same content, with the same pattern and degree of emphasis, as the summative test. Not all content is equally difficult – geometry and algebra items are generally more difficult than items relating to number sense and numerical operations – so an item-level adaptive test might omit large swaths of content for students who are very successful or very unsuccessful on early items. Furthermore, since there is not a proscribed sequence of instruction, not all students would have been taught the same content. The routing test needed to include all Tested Indicators so that student ability estimates would be more accurate. Since the interim assessment covers the entire range of Tested Indicators, not just the Instructed Indicators selected by the teacher, one would expect that the content that has not yet been taught would appear more difficult to the students. The most effective way to ensure that each interim test covered the required material and that student ability was accurately estimated was to have a multi-stage adaptive model rather than an item-adaptive model.

Overall, throughout the three stages, each path through the interim test includes two to four items from each Tested Indicator, and each interim assessment is about half the length of the general summative assessment. Tables 1 and 2 show two representative grades' indicator coverage. Small differences in the percent by indicator result from having odd numbers of items in the summative test (therefore not being well able to divide by half), and rounding with small numbers.

Table 1.

Grade 3 Interim and Summative Blueprints.

Indicator	Interim	% of Interim	Summative	% of Summative
3.1.1.K2	3	9.4%	6	8.6%
3.1.1.K3	3	9.4%	8	11.4%
3.1.1.K4	2	6.3%	4	5.7%
3.1.4.A1	2	6.3%	7	10.0%
3.1.4.K7	2	6.3%	4	5.7%
3.2.1.A2	3	9.4%	6	8.6%
3.2.3.K.3	3	9.4%	6	8.6%
3.3.1.K4	2	6.3%	5	7.1%
3.3.2.K2	3	9.4%	4	5.7%
3.3.2.A1	3	9.4%	7	10.0%
3.4.1.K2	3	9.4%	5	7.1%
3.4.2.K3	3	9.4%	8	11.4%
Total	32	100%	70	100%

Table 2.

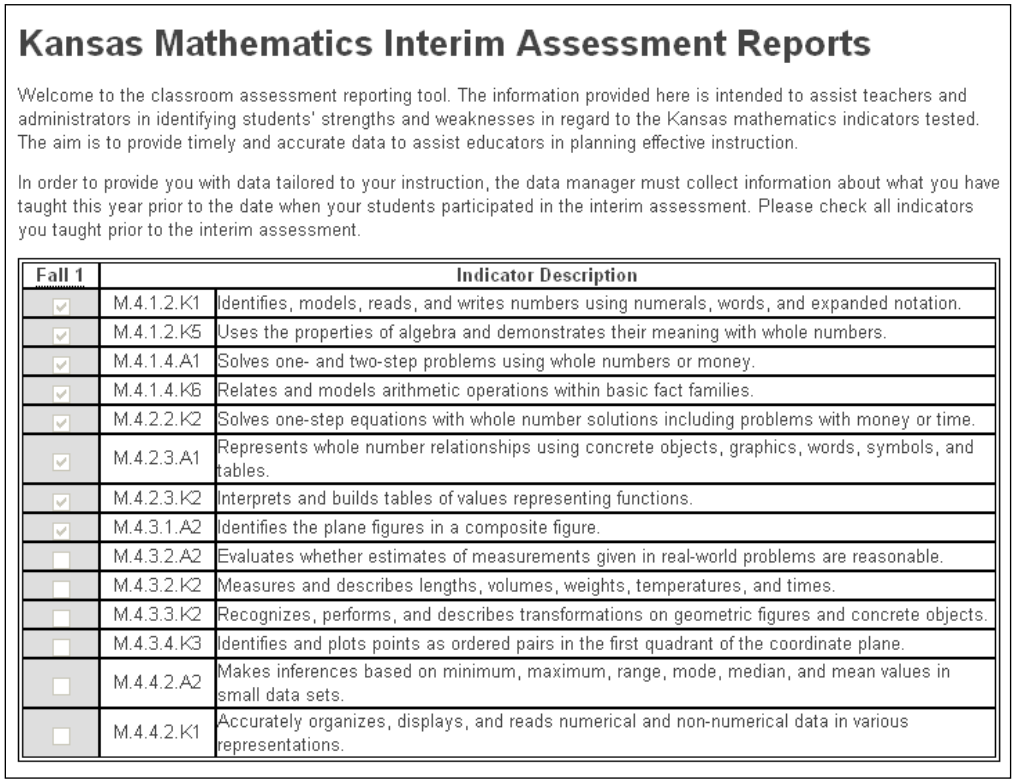
Grade 8 Interim and Summative Blueprints.

Indicator	Interim	% of Interim	Summative	% of Summative
8.1.1.K5	2	4.9%	6	7.0%
8.1.2.A1	3	7.3%	6	7.0%
8.1.2.K2	3	7.3%	4	4.7%
8.1.4.A1	3	7.3%	6	7.0%
8.1.4.K2	3	7.3%	6	7.0%
8.2.2.A1	2	4.9%	7	8.1%
8.2.2.K3	2	4.9%	5	5.8%
8.2.3.A3	3	7.3%	7	8.1%
8.2.4.A2	3	7.3%	4	4.7%
8.3.1.A1	3	7.3%	5	5.8%
8.3.1.K6	2	4.9%	5	5.8%
8.3.4.K1	3	7.3%	8	9.3%
8.4.1.A4	3	7.3%	7	8.1%
8.4.1.K3	3	7.3%	4	4.7%
8.4.2.K3	3	7.3%	6	7.0%
Total	41	100%	86	100%

Interim Score Reporting

Interim assessment scores were placed on the same scale as the same grade's summative assessment scores. Because each interim assessment contained only two to four items per indicator, and students received different items depending on their route through the assessment,

subscores by indicator were not reported. In order to provide teachers with information that could be used to relate their students' test performance to instruction, teachers were able to request subscores of the items that aligned to indicator that were taught prior to each test ("Instructed Indicators"). After logging in to the website, the teacher welcome screen included a list of all the Tested Indicators at that teacher's grade level (Figure 2). The teacher had the option using radio buttons to identify which indicators were taught prior to the test administration. If a teacher selected at least five Instructed Indicators, then subscores based on only the items aligned to these indicators were computed and available within one day.



Kansas Mathematics Interim Assessment Reports

Welcome to the classroom assessment reporting tool. The information provided here is intended to assist teachers and administrators in identifying students' strengths and weaknesses in regard to the Kansas mathematics indicators tested. The aim is to provide timely and accurate data to assist educators in planning effective instruction.

In order to provide you with data tailored to your instruction, the data manager must collect information about what you have taught this year prior to the date when your students participated in the interim assessment. Please check all indicators you taught prior to the interim assessment.

Fall 1	Indicator Description	
☑	M.4.1.2.K1	Identifies, models, reads, and writes numbers using numerals, words, and expanded notation.
☑	M.4.1.2.K5	Uses the properties of algebra and demonstrates their meaning with whole numbers.
☑	M.4.1.4.A1	Solves one- and two-step problems using whole numbers or money.
☑	M.4.1.4.K6	Relates and models arithmetic operations within basic fact families.
☑	M.4.2.2.K2	Solves one-step equations with whole number solutions including problems with money or time.
☑	M.4.2.3.A1	Represents whole number relationships using concrete objects, graphics, words, symbols, and tables.
☑	M.4.2.3.K2	Interprets and builds tables of values representing functions.
☑	M.4.3.1.A2	Identifies the plane figures in a composite figure.
☐	M.4.3.2.A2	Evaluates whether estimates of measurements given in real-world problems are reasonable.
☐	M.4.3.2.K2	Measures and describes lengths, volumes, weights, temperatures, and times.
☐	M.4.3.3.K2	Recognizes, performs, and describes transformations on geometric figures and concrete objects.
☐	M.4.3.4.K3	Identifies and plots points as ordered pairs in the first quadrant of the coordinate plane.
☐	M.4.4.2.A2	Makes inferences based on minimum, maximum, range, mode, median, and mean values in small data sets.
☐	M.4.4.2.K1	Accurately organizes, displays, and reads numerical and non-numerical data in various representations.

Figure 2. Selecting Instructed Indicators.

Teacher reports listed scale scores and instructed scores for each student in a class. This information was displayed in a table with a line for each student and a column for each test score. A screenshot of the roster report is shown in Figure 3.

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 3. Sample Class Roster Report for Total Score and Instructed Indicator Score.

Below this table, class distributions of scores were graphed in box-and-whisker plots. Each plot displayed the class median, first quartile, third quartile, minimum, and maximum scores. After each administration window closed, the district median was also displayed to give the teacher an idea of how the class compared to other classes at the same grade in the same district. A screenshot of class distribution information is shown in Figure 4. In order to assist

teachers in interpreting these charts, professional development included a pamphlet and presentation that explained each statistic in a box-and-whisker plot and how to interpret the plot in terms of class performance on a test. These materials were available online throughout the school year for any teacher or administrator to download and use freely.

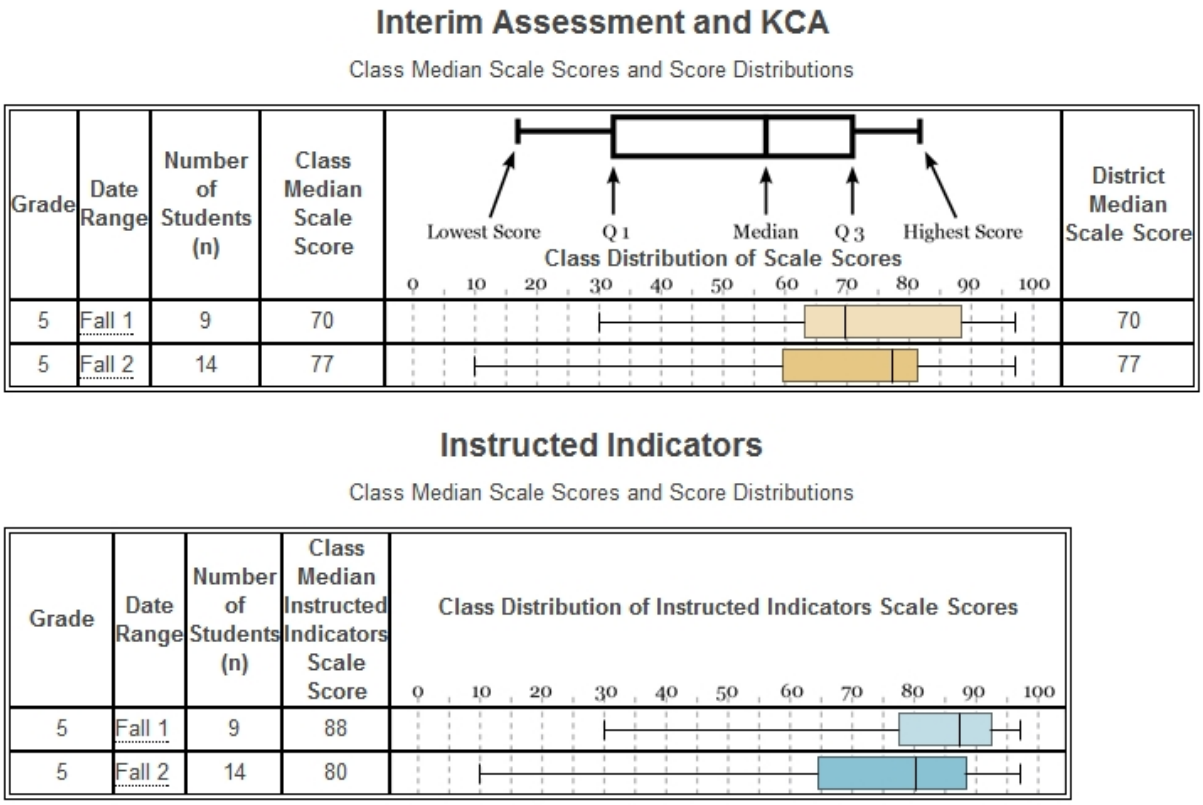
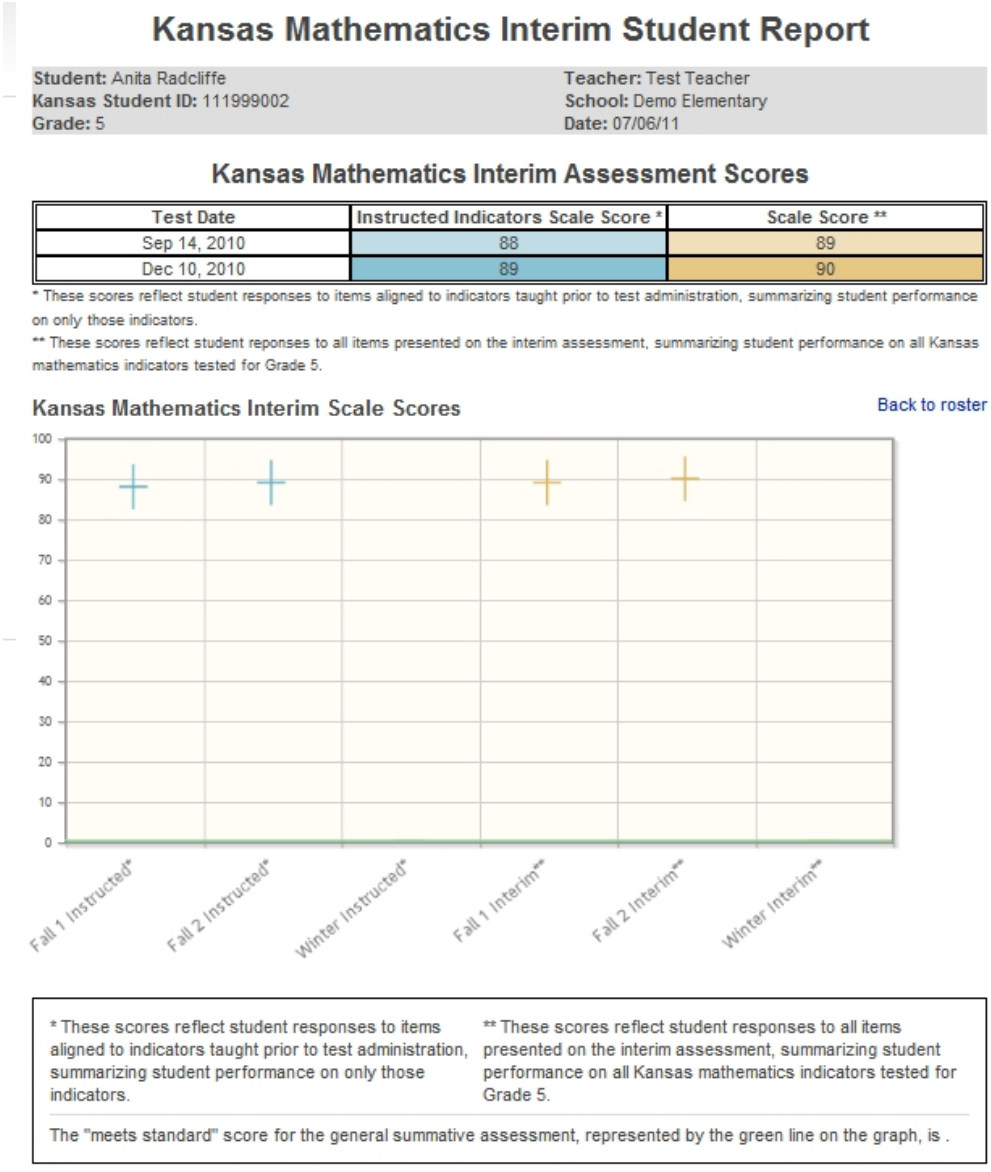


Figure 4. Sample Classroom Score Distribution Information.

Individual student reports were provided and listed all the information previously described in addition to a graphical representation of the student's individual scores. Great care was taken to provide detailed definitions of all the terms used in these reports in case they were printed and sent home to parents or guardians without additional explanation. Because of the very real concern that low scores on interim assessments could prompt negative reactions from

stakeholders who were unaware of the test's purpose, design, or scale, educators and administrators were advised that interim assessment reports be shared with parents only in face-to-face meetings, during which the educators could explain the meaning of the reports. A screenshot of an individual student report is shown in Figure 5.



Kansas Mathematics Interim Student Report

Student: Anita Radcliffe
Kansas Student ID: 111999002
Grade: 5

Teacher: Test Teacher
School: Demo Elementary
Date: 07/06/11

Kansas Mathematics Interim Assessment Scores

Test Date	Instructed Indicators Scale Score *	Scale Score **
Sep 14, 2010	88	89
Dec 10, 2010	89	90

* These scores reflect student responses to items aligned to indicators taught prior to test administration, summarizing student performance on only those indicators.
** These scores reflect student reponses to all items presented on the interim assessment, summarizing student performance on all Kansas mathematics indicators tested for Grade 5.

Kansas Mathematics Interim Scale Scores

Back to roster

* These scores reflect student responses to items aligned to indicators taught prior to test administration, summarizing student performance on only those indicators.

** These scores reflect student responses to all items presented on the interim assessment, summarizing student performance on all Kansas mathematics indicators tested for Grade 5.

The "meets standard" score for the general summative assessment, represented by the green line on the graph, is .

Figure 5. Individual Student Report.

As the interim assessment becomes better understood and implemented to inform program-level decisions, building reports will be available for entire buildings so that administrators can gather all the data from interim assessments in a single download.

Validity Evidence for the Kansas Interim Assessment Based on Test Content

Among the multiple conceptions of validity, the main idea that bubbles to the top is the question, "Do the test scores support the meaning and interpretations they are expected to support?" Validity is not a unitary thing; rather it is the collection of evidence starting with the initial test development and continuing into test administration and the use of the test results. The Joint Standards (AERA/ APA/NCME, 1999) discuss five sources of validity evidence: evidence based on test content, evidence based on response processes, evidence based on internal structure, evidence based on relations to other variables, and evidence based on consequences of testing. Other papers in this symposium deal with internal structure (Wang), relations to other variables (Kingston), and consequences of testing (Broaddus); the current paper addresses test content.

With the reauthorization of the Elementary and Secondary Education Act in 2001, the legislation known as No Child Left Behind ushered in a related set of standards for technical quality of assessments. These standards must be met in order for a state's assessments to be used for federal accountability purposes. As part of the Elementary and Secondary Education Act, each state undergoes a technical peer review of the state's standards and assessment system. The peer reviewers are assessment and policy experts from across the nation who study each state's submission and evidence and make recommendations to the U.S. Department of Education's Standards and Accountability division. The peer review guidance itself is divided into seven

sections; section 4 deals with technical quality, and critical element 4.1 is validity. The peer review guidance proposes that states document several aspects of validity. Although a state's interim assessment need only go through peer review if results from the test are used in federal accountability, the peer review guidance provides some additional considerations for documenting validity evidence. Most relevant to this paper are those related to content:

- *Has the State specified the purposes of the assessments, delineating the types of uses and decisions most appropriate to each?*
- *Has the State ascertained that the assessments… are measuring the knowledge and skills described in its academic content standards and not knowledge, skills, or other characteristics that are not specified in the academic content standards or grade-level expectations?*
- *Has the State ascertained that its assessment items are tapping the intended cognitive processes and that the items and tasks are at the appropriate grade level?* (USED, p. 42)

Other sources of validity evidence can be thought of as more universal as well as decentralized from the test itself. In a newer validity framework (Embretson, 2007), additional evidence can come from practical constraints such as standardization and appropriateness of conditions under which the test is administered, the principles followed during item design and development, and from the conditions laid out in the test specifications.

The state department of education and the stakeholders, when describing their needs, specifically requested a test that could be used to predict summative assessment scores that teachers could use to examine the effectiveness of their teaching strategies on the Instructed

Indicators, and that would provide instructionally useful feedback on the class's attainment of the Tested Indicators as a whole. The interim assessments, each being about half as long as the summative assessment, were never intended or designed to provide feedback at the individual student level. Indeed, given that there are only two or three items per indicator, and the adaptive nature of the test, student-level indicator scores would be unreliable as well as non-comparable.

As previously described, the test interface for students is the same as is used for the online administrations of the summative assessment. To the extent possible, the same tools and accommodations are available, with the exceptions noted above such as Braille, Spanish, and scripted read-aloud. The interim tests were designed following the same test specifications and blueprint; in fact, the items that were used to develop the first interim math test forms were initially developed for use in the summative assessments but not selected when the summative forms were built. From a practical standpoint, the adherence of the interim tests to the identical test development processes, specifications, and administration procedures provide strong validity evidence.

The Interim Assessment Examiner's Manual, produced by the test developers and the state department of education, opens with the purpose of the Kansas Interim Assessment:

> ***Purpose of the Kansas Interim Assessment***
>
> *The Kansas Interim Assessment is a program of the Kansas State Board of Education provided as a resource for educators and students in Kansas schools. The interim assessment program is designed to:*
>
> - *Measure the same specific indicators within the Kansas Curricular Standards that are measured by the summative Kansas Assessments.*

- *Provide students and teachers with estimates of student achievement on tested indicators at three time points prior to the summative assessment.*
- *Report individual student scores along with class score distribution information and district average scores.*
- *Provide subscale and total scale scores that can assist in making instructionally relevant decisions.* (KSDE, p. 4)

As this assessment was offered for the first time in the fall of 2010, many online and face-to-face trainings and professional development sessions were offered throughout the first half of the 2010-2011 school year to introduce this new assessment to teachers and administrators. These trainings included guidance on what teachers and administrators could or ought not do with the results. Predictive data of course was not available during the first year of implementation, but stakeholders were informed about how the data collected during the first year would contribute to the development of the predictive model. In spite of the training sessions and public relations work in the first year of the assessment, most of the negative feedback received was that teachers still wanted individual student indicator-level scores, as are made available with some of the commercial products that were in use in the districts.

When the Kansas math standards were adopted, thousands of items were developed and reviewed; most of them were field-tested; and many that were field-tested were used on the summative forms. There were, however, many items that were left over from the summative test development. Some had been field tested but just did not fit onto a form, some that had been through review but there were already enough from that indicator being field tested, some that had been to review but needed revisions, and so forth stopping at various points in the item development process. These items were recalled to service. Items that needed editing were

changed according to the reviewers' notes, items that had not been all the way through the review process were reviewed, and gaps in the coverage of the content standards were identified. When necessary, additional items have been written explicitly for the interim assessment program, and reviewed according to Kansas' item review procedures. When items were ready to be field tested again, they were embedded into the operational summative forms so that the most accurate data could be collected to recalibrate the items.

The Kansas item development and review process, as with many other states, includes several layers of review. First, items are written by trained item writers to align to the Kansas standards. They are reviewed internally for content alignment, appropriateness of context and vocabulary for the grade level and for cultural mores, edited for grammar and mechanics, reviewed for accessibility to all tested sub-populations including students with disabilities and English language learners, examined for clarity of both the question itself and any stimulus materials (graphs or artwork) that accompany the question, and reviewed for the many psychometric rules of good item writing. The items then go to external reviews . First they are reviewed by the Kansas State Department of Education for alignment and for adherence to the intent of the standards. Items are reviewed by panels of teachers drawn from across the state, representing large and small districts; drawn from rural, urban, and suburban districts; and representing the state in terms of racial diversity as well as gender. The teacher panels include teachers of special education students and English language learners, so that the particular needs of those students are also reflected in the development of test items. A panel of non-educator stakeholders is convened to review the items for potential bias or sensitivity issues. Finally, items are fact-checked for the accuracy of content beyond the mathematics concepts (for example, if a

math item has a science context, the science has to be right also – when a ball is dropped, the function describing its fall in feet per second should include the term $^{-}16x^2$).

When the items had passed through all this scrutiny, they were embedded into the operational forms for field testing in the manner least likely to interfere with the behavior of the summative test and to provide the most seamless transitions for the students. Items for a given indicator were field tested in positions within or adjacent to the operational items for that same indicator. Items that were to be used in the calculator-active portion of the interim assessment were field tested in the calculator-active portion of the summative assessment. All allowable tools were active for the field tested items as they would be for the formative. Due to logistical constraints, however, the field tested items were only embedded in the computer-administered assessments; the handful of students who took paper forms were not included in the item calibrations. As previously noted, however, over 99.5% of students in Kansas take the summative assessments online.

After field testing, the items were reviewed once more with their data. Items with differential item functioning were reexamined for bias. Items with "bad" statistics were reviewed and some were excluded from the item pool. The same criteria for item selection for the summative item pool were used in selecting items to be included in the interim item pools.

Once the interim item pools were established, the "testlets" that were represented in Figure 1 were developed according to the interim test blueprint. Recall that the interim test blueprint was derived from the summative test blueprint to be about half the length of the summative test and with resultant very minor changes in emphasis of the tested indicators. The testlets were reviewed for match to the blueprint, for clang associations among items, and for

adherence to the principles of Universal Design. The adaptive algorithm was checked to ensure that students would be routed correctly and a final check was done of the scoring algorithm to ensure that students received accurate test scores.

There are a few differences in these final stages of development between the interim assessments and the summative assessments. First, the summative tests are not adaptive and are based on classical test theory; the interim tests take advantage of item response theory to estimate student ability and route the student to the most appropriate testlet in the next stage. Additionally, the initial calibrations for the items used in the summative test were collected from an explicit (stand-alone) field test design; subsequent item calibrations, including the interim items, were collected from embedded field testing. Finally, when the summative test forms were originally developed, they were not reviewed for Universal Design.

The parallel (and for the most part, coincident) item and test development processes followed in creating the interim assessments when compared to the summative assessment development processes provide strong evidence of validity for the interim tests' purposes. Furthermore, the many layers of reviews by curriculum specialists, content experts, and active classroom practitioners ensure that the items used on the interim measure the Kansas tested indicators at the targeted grade level and at the intended level of cognitive complexity. Including diverse reviewers and multiple levels of content, bias, and sensitivity review, as Kansas and many other states do, serve to remove potential sources of challenge or construct-irrelevant noise from the items.

With the success of the mathematics interim assessment, interim assessments are being developed for reading as well. Since the Kansas reading comprehension items are passage based,

additional reviews are necessary. In addition to the item-level reviews as described above, the reading passages are reviewed separately for accessibility, interest, and potential concerns with bias or sensitivity. Additionally, multiple readability indices are calculated on each passage, and word count of each passage is also recorded. Reviewers also indicate if the passage contains unfamiliar or low-incidence words that should either be defined with a glossary or footnote, or if the word would be a good candidate for a question about vocabulary in context or determining word meaning by decomposing the word into the root word and its prefixes or suffixes. All passages are reviewed by classroom teachers as well as staff from the Kansas School for the Blind and the Kansas School for the Deaf to ensure the passages do not contain imagery or content that would be inaccessible to students with these disabilities.

In spite of the expectation that each interim assessment could be administered comfortably within a single class period, this turned out to not be the case, particularly in the lower grades. Additionally, analysis of actual student data revealed that not all testlets were used as students went from the routing test to stage 2 to stage 3. After consulting with the technical advisory committee and prophesying the effect on reliability (Spearman, 1910; Brown, 1910), the decision was made to redesign the interim assessments from three stages to two. The test blueprints were pared back, and the criteria for item placement into each testlet were reevaluated so that there was more differentiation between testlets in the second stage. It is now expected that most students will be able to complete the interim assessment in 45 to 50 minutes. The overall test reliability is prophesied to drop given that the number of items is decreased; the lowest tests' reliabilities drop from an observed coefficient alpha (Cronbach, 1951) value of 0.84 to a prophesied value of 0.82. Table 3 contains the observed reliabilities for each of the three interim test windows from last year and the prophesied reliability for next year.

Table 3.

Observed and Prophesied Reliability for Mathematics Interim Assessment.

	Grade	Window	Observed Reliability	Prophesied Reliability
Math	3	Fall1	0.87	0.83
		Fall2	0.87	0.84
		Winter	0.87	0.84
	4	Fall1	0.84	0.82
		Fall2	0.84	0.82
		Winter	0.84	0.82
	5	Fall1	0.86	0.84
		Fall2	0.87	0.85
		Winter	0.86	0.84
	6	Fall1	0.87	0.83
		Fall2	0.90	0.87
		Winter	0.90	0.87
	7	Fall1	0.89	0.86
		Fall2	0.89	0.87
		Winter	0.89	0.87
	8	Fall1	0.91	0.88
		Fall2	0.91	0.88
		Winter	0.91	0.88

It is hoped that the tighter item selection criteria and better discrimination among testlets will recoup some of that drop. However, the other common complaint, that teachers want student indicator-level feedback, could not be addressed at the same time. In fact, by shortening the test further, there are only two items per indicator, which would result in extremely unreliable indicator-level information for individual students.

Interim assessments will be available in the coming school year for grades 3-5 reading, grades 3-8 mathematics, and high school mathematics; passages and items are being developed for grades 6-8 reading interim assessments for the following year. Plans are already underway

for additional professional development for teachers and administrators to help them incorporate the interim assessments effectively into their instructional planning, to understand the proper uses and interpretations of the interim test data (as well as the limitations), and to better understand where the interim assessments fit into a comprehensive assessment system.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). Standards for Educational and Psychological Testing. Washington, DC: AERA.

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology, 3,* 296–322.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16,* 297-334.

Earl, L.M. (2003). *Assessment as Learning: Using Classroom Assessment to Maximize Student Learning.* Thousand Oaks, CA: Corwin Press.

Embretson, S.E. (2007). Construct Validity: A Universal Validity System or Just Another Test Evaluation Procedure? *Educational Researcher*, Vol. 36, No. 8 (November, 2007): 449-455.

Herman, J.L., Osmundson, E., & Dietel, R. (2010). *Benchmark assessments for improved learning* (AACC Report). Los Angeles, CA: University of California.

Kansas State Department of Education. (2010). *Kansas Interim Assessment – Examiner's Manual.* Topeka, KS: KSDE.

Popham, W.J. (2011). Combating Phony Formative Assessment—With a Hyphen. *EdWeek*, Vol. 30, No. 21 (February 23, 2011): 35.

Spearman, Charles, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology, 3,* 271–295.

Stiggins, R.J. (2002). Assessment Crisis: The Absence of Assessment FOR Learning. *Phi Delta Kappan*, Vol. 83, No. 10 (June 2002): 758-765.

U.S. Department of Education. (2009). *Standards and Assessments Peer Review Guidance: Information and Examples for Meeting Requirements of the No Child Left Behind Act of 2001.* (January 12, 2009). Washington, DC: USED.