Kansas Assessments in Science

2008

TECHNICAL MANUAL

for the

Kansas Assessments of Modified Measures (KAMM)

And

Kansas Alternate Assessments (KAA)

Prepared by:

Patrick M. Irwin, Neal M. Kingston,
Douglas R. Glasnapp, and John P. Poggio

Center for Educational Testing and Evaluation
The University of Kansas

September 2008

Table of Contents

Kansas KAMM Assessment in Science

PURPOSE OF THE TECHNICAL REPORT

The *Standards for Educational and Psychological Testing* (AERA/APA/NCME, 1999) requires that test developers and publishers produce a technical manual that provides information documenting the technical quality of an assessment, including evidence for the reliability and validity of test scores. This report contains the technical information for the 2008 Science Kansas Assessments of Modified Measures (KAMM) for grades 4, 7, and high school. The information included in this report is intended for use by those who evaluate tests, interpret scores, or use test results in making educational decisions. It is assumed that the reader has some technical knowledge of test construction and measurement procedures.

The main body of this report addresses technical aspects focusing on scores from the 2008 Science Kansas Assessments of Modified Measures (KAMM) for grades 4, 7, and high school. Information on the Kansas Alternate Assessments (KAA) is found in Appendix A.

Information is provided addressing the technical quality of assessments developed to measure science learning outcomes specific to one distinct population of Kansas students, namely students with moderate disabilities. The Kansas Assessments of Modified Measures (KAMM) are intended for administration to students with such disabilities. Thus, the body of this report addresses technical aspects focusing on scores from the KAMM tests.

Section 1

INTRODUCTION AND ORIENTATION

KAMM TEST TECHNICAL CHARACTERISTICS

The Kansas Assessment of Modified Measures (KAMM) is a state assessment with modified achievement standards based on grade level content standards. A student with a disability whose Individual Education Plan (IEP) team used the KAMM eligibility criteria (available from the Kansas State Department of Education) and determined the KAMM/modified assessment is the appropriate assessment for the student who may take the KAMM. The basis for the KAMM is less complexity of test items. KAMM compares to the general assessments in science in that the same assessed indicators are used; however, the number of indicators assessed is reduced. Fewer multiple-choice items appear on the KAMM than on the general assessment. In science, there are 42 items at grade 4 and 60 items at grade 7. The high school science KAMM assessment is divided into a physical science section which is comprised of 30 items and a life science section which is also comprised of 30 items. There are three answer choices for all KAMM selected response items, compared to four answer choices on the general assessment. Students were allowed to take the KAMM over as many days as necessary.

Accommodations are allowed on the KAMM as the IEP team makes about accommodations for the KAMM the same as they do for the general assessment. Additionally, the KAMM is primarily a computer-delivered assessment. For students who cannot complete the KAMM assessment online, a paper and pencil accommodation can be made.

The population of students taking the KAMM did not result in a sufficient number of students within subgroups to conduct empirical Differential Item Functioning (DIF) analyses. However, items were reviewed by KSDE panels for bias, offensiveness, and insensitivity.

As background, new KAMM assessments in science were planned and developed, then administered for the first time in Spring 2008. WestEd served as the contractor for the development of test items based on test specification provided by KSDE. The Center for Educational Testing and Evaluation (CETE) at The University of Kansas served as the contractor for all other aspects of the program. Students with moderate disabilities in grades 4, 7, and 11 participated in the assessments.

The Spring 2008 administration of the KAMM serves as the baseline for the new cycle of state assessments. The assessments administered were all newly developed to measure the new targeted indicators (learning outcomes) in the most recent editions of the Kansas Curricular Standards for science. These documents should be referenced when examining and evaluating any of the information resulting from the state assessment programs. The Curricular Standards serve as the basis for what is assessed by the tests, and any interpretation and subsequent action based on student or group performance on these tests must focus on the assessed standards, benchmarks, and indicators. Copies of the Kansas Curricular Standards in the content areas are available from the KSDE website at www.ksde.org.

As the baseline year of the new round of assessments, the Spring 2008 administration incorporated important changes from prior KAMM assessments administered in the 2000 – 2007 testing cycle. Curriculum standards and targets for the assessments were changed and test specifications revised. Any comparisons to past student, building, district or state performance should be made cautiously.

KAMM assessments were developed by the state and made available in lieu of the general assessments for administration to students with moderate disabilities. Single forms measuring targeted learning outcomes in science were developed for each grade level, 4, 7, and high school. The item format used in these assessments was multiple-choice with three response options.

The KAMM assessments are planned and created to reflect and otherwise operationalize certain grade level learning outcomes that should serve as curricular and instructional targets in Kansas K-12 schools. As in previous years, the assessments contribute information for ongoing school accreditation status determinations, and results from the science assessments have a primary role in monitoring annual yearly progress (AYP) as part of the federally mandated No Child Left Behind assessment requirements. As related to the accountability demands, cut scores on each test were determined to classify students into one of five performance categories (Exemplary, Exceeds Standard, Meets Standard, Approaches Standard, and Academic Warning). The proportion of students classified in these categories becomes a primary source of information in determining AYP for schools, districts, and the state. Section 3 of this report provides additional details about procedures for setting the specific test score criteria used to classify students into one of the five performance categories established by the state.

As a final important aspect of the KAMM assessment program, administration of the tests are offered under one of two modes on a voluntary basis, a paper and pencil (P&P) test administration mode or an online administration using the Kansas Computerized Assessment (KCA) system developed by the Center for Educational Testing and Evaluation at The University of Kansas. Documentation describing the KCA system may be found at www.kca.cete.us.

Section 2

TEST DEVELOPMENT AND CONTENT REPRESENTATION

The content of the KAMM is derived from the Kansas Curricular Standards. These Curricular Standards define, for Kansas schools, what students should know and be able to do in the respective content domains at each grade level. The 2008 KAMM measured targeted indicators in the Curricular Standards for science in grades 4, 7, and high school.

Test Specifications

Test specifications provide the blueprint to be followed when writing items and constructing test forms. KSDE developed and provided the test specifications that guided all item and test development efforts. Test specifications were provided in matrix form that identified, by cognitive complexity level and targeted indicators (skill) to be assessed, the number and distribution of items to be on each test form at a grade level. These grade level and content area specifications guided the construction of operational forms development, but the order and manner in which items were placed throughout the forms was left to the collaborative efforts of CETE test development staff and KSDE content specialists. The most recent versions of the test specifications can be obtained through the KSDE website.

Item Type

The multiple choice item type is the only item type used on the KAMM in science. For all multiple choice items appearing on any KAMM form, students select the one best answer from among three choices provided.

Item Development

KSDE contracted with WestEd, to supply science items that were aligned with the content area Curricular Standards. The actual items that would make up the assessments at each grade level would come from these item pools after several rounds of reviews and empirical tryouts (pilot testing), the latter conducted by CETE.

The final rounds of item pool reviews involved content review and fairness review committees comprised of Kansas educators. Along with KSDE specialists, the content committees reviewed each item, focusing on its alignment to the table of specifications and to the Kansas Curricular Standards, as well as the appropriateness of item content, ensuring that each item accurately reflected what was intended to be taught in Kansas schools. The fairness review committees focused on language and content that might be inappropriate, offensive, or insensitive to students, parents, or communities, making sure that no individual or group would be unfairly favored or disadvantaged due to the content of the items. With both review committees, each item was accepted, edited, or rejected from its respective item pools.

KAMM Science Summary Statistics

Table 2.1 reports summary findings for the KAMM Science Assessments by grade level, identifying the number of items per KAMM form, the number of students administered at a particular grade level KAMM, reliability coefficients, and descriptive statistics in terms of raw total scores and percent correct total scores. Grade 11 is divided into two sections, life science (LS) and physical science (PS). All of the reliability coefficients are greater than 0.53. The percent mean correct ranged from 41.69% (SD = 12.40) at grade 11 (PS) to 70.77% (SD = 14.87) at grade 4. Table 2.2 reports reliability coefficients for the KAMM Science Assessment by demographic (specifically by race and students on free and reduced lunch programs). Students categorized as English language learners (ELL) were also included in this analysis; however, the sample size was not large enough (N > 100) at each grade level to produce stable estimates. This was also the case for students categorized as Hispanic at Grade 11 in both physical science and life science.

Table 2.1
2008 KAMM Science Summary Statistics

Grade	Test_ID	# of Items	N	Reliability (α)	Mean Raw Score	SD of Mean Score	Mean Percent Correct	SD of Mean Percent Correct
4	128950	42	740	0.82	29.72	6.25	70.77	14.87
7	128951	60	850	0.76	28.94	7.25	48.24	12.08
11(LS)	128952	30	1,006	0.53	14.09	4.30	46.97	14.34
11(PS)	128953	30	915	0.55	12.50	3.72	41.69	12.40

Table 2.2
Reliability Coefficients for 2008 Science KAMM Assessments Based on Demographic

				Race						Free/Reduced Lunch	
				Hispanic		African-American		Caucasian			
Grade	N of Students	N of Items	Reliability (α)	N	α	N	α	N	α	N	α
4	740	42	0.82	110	0.81	115	0.80	457	0.83	511	0.82
7	850	60	0.76	147	0.72	132	0.68	508	0.76	551	0.74
11 (LS)	1,006	30	0.53	96	NA	165	0.45	692	0.65	534	0.63
11 (PS)	915	30	0.55	92	NA	141	0.29	631	0.57	477	0.49

Item Delivery and Tryouts

All science items that were approved were delivered via electronic upload to the CETE server. Items received were subjected to reviews by CETE staff prior to being assembled onto pilot forms that would be administered in field tests to representative samples of Kansas students.

All Kansas schools were encouraged and invited to participate in the science assessment pilot testing. Due to the large number of items to be piloted across the science content area, a fourth test session was added to 2007 state testing of mathematics and reading at all grade levels. For grades 4, 7, and high school, the students took science pilot items and for the remaining grades took history and government, mathematics, or reading items. All items in the science item pools supplied to CETE were piloted. Science pilot tests were administered via the KCA delivery mode. Since the KAMM assessments were administered via KCA, paper and pencil pilot forms were not available. At grade 4, there were five KCA pilot forms; at grade 7, five forms; and at high school, six forms. When the students logged into the KCA system, they were randomly assigned to test form.

Pilot Item Analysis

Following the administration of the pilot test item sets, statistical item analyses were conducted to determine the effectiveness and quality of the items. For multiple choice items, the item means (p value) and item-test correlation coefficients (point biserial) were calculated. Further, statistics for each response alternative were also calculated and examined. The proportion of examinees responding to each response option was obtained, as well as the point-biserials for each response choice. In addition, the proportion of a low ability (lowest 27% based on total score) group and a high ability (upper 27%) group responding to each choice option was obtained. The difference in p-values for these two ability groups on the correct answer choice yielded another index of item discrimination (Kelly index) that provided information about the item's ability to differentiate between high and low scoring examinees.

Across grade levels assessed, just over 200 items were piloted and subsequently evaluated by CETE test development staff using classical item analysis procedures described above. To assist in the pilot item review process, a set of rules were adopted to assist in identifying poorly functioning (items that are too easy, too difficult, contain errors, or have low or negative discrimination information, for example). The rules or criteria for identifying poorly functioning items were the following.

Items were flagged for review if:

- $r_{pb} < 0.20$ for the keyed (correct) response
- $p > 0.95$ or $p < 0.25$ for the keyed response
- $r_{pb} > 0$ for any distractor (incorrect answer choice)
- $p > 0.25$ for a distractor for the high ability group OR $p > 0.15$ and $r_{pb} > 0.055$ for the low ability group
- the Kelly discrimination index for an item is less than 0.20

Each item that was flagged based on the criteria listed above was individually reviewed by KSDE. During these reviews, items were either accepted or rejected for the final pool of items.

Since only one form at each grade level was to be constructed, there was a limited number of items commissioned by West Ed and the number of spare items per indicator was considerably low. KSDE reviewed all discrepant items.

Section 3

STANDARD SETTING

2008 Kansas Science Performance Standards

Performance standards were set for the 2008 Kansas Science Assessments using a multistep process designed in keeping with the dictum that standard setting is a policy decision supported by data. Following are the major steps in that process. Science KAMM performance standards were set in conjunction with the Science General Assessments as well as the Science Alternate Assessments.

1. Development of performance level names
2. Development of performance level descriptors
3. Bookmark procedure
4. Standard setting policy advisory group
5. State Board of Education adoption of performance standards

The first and second steps are intended to provide guidance for subsequent steps. The third step provides an operational definition of each performance level (identify possible cut scores) consistent with the performance level descriptors. The fourth step provides an opportunity to identify the desirability of other forms of consistency, such as across grade level, test type (General-KAMM-Alternate), or academic discipline. The last step is to present the State Board of Education with the information it needs to set Kansas cut score policy.

1. Development of Performance Level Names

At its August 8, 2006 meeting, the Kansas State Board of Education adopted five performance level names to describe the quality of student achievement demonstrated in each tested discipline on the Kansas State Assessments. Those performance levels, from lowest to highest, were entitled as follows.

1. Academic Warning
2. Approaches Standard
3. Meets Standard
4. Exceeds Standard
5. Exemplary

While these performance level names were new, they were intended to clarify the meaning of the existing five categories which had previously been called Unsatisfactory, Basic, Proficient, Advanced, and Exemplary. The new performance level names were first applied to the results of the 2005-2006 test administration.

2. Development of Performance Level Descriptors

Performance level names create a shared understanding of the level of achievement indicated by each performance level but, in and of themselves, remain highly subjective. What one teacher thinks of as exemplary achievement will differ from another teacher unless steps are taken to clarify expectations. Reducing this inherent subjectivity requires the development of performance level descriptors – a verbal description of what it means to be in a particular performance level. While all tests (mathematics, reading, history and government, and science) share the same performance level names, each has its own performance level descriptors. In order to maximize clarity, Kansas has chosen to write the specific curriculum indicators addressed by each grade level assessment into each performance level descriptor. Since each grade level addresses (and therefore assesses) different indicators, there are separate performance level descriptors for science at grades 4, 7, and high school.

Also, while the performance level descriptors are quite similar for the Kansas General Assessment and the Kansas Assessment of Multiple Measures, they are not identical, and thus at each grade level, there are separate performance level descriptors for each, for a total of nine science performance level descriptors.

3. Bookmark Procedure

The Bookmark Procedure (Mitzel et al., 2001) was used as the next step in the standard setting process. For each test, items were ordered from easiest to hardest. For each performance level, participants were asked to make a judgment about the items that a student at the threshold of one category should have mastered versus those not necessary to be mastered. Panelists were advised that the distinction is not intended to be within the immediate pair of items (the one they were looking at now versus the previous item) but between several previous items and several subsequent items, the items before and after the bookmark. Panelists then placed the bookmark where they estimated a threshold student would have a 0.67 probability of responding correctly to a selected response item at the cut-point.

The Bookmark procedure was implemented by training the participants and then by performing three iterations. First, each panelist placed each bookmark independently. Then panelists were provided with their group's data, and they discussed where they placed their bookmarks as well as the rationale for their decisions. At this time, no attempt was made to come to consensus but instead to simply understand the issues considered. Then panelists went through a second round of placing bookmarks informed by those discussions. Results of the second round were provided to the groups as was consequence data, the estimated percent of students who would fall into each performance category if the average of the group's judgment was implemented.

Preparation of Item Ordered Booklets

The following describes the creation of ordered item booklets which were prepared for Bookmark standard setting activities that took place in Summer 2008. Standard setting activities

for science were conducted at grades 4, 7, and high school (life science and physical science). For each of these four grade level tests, a KAMM was also administered, thus ordered item booklets were also created for KAMM. The four tests for which ordered item booklets were created are listed in Table 3.1.

Table 3.1
Overview of Tests for Which Performance Standards Were Set

Subject	Grade	# Items (KAMM)
Science	4	42
Science	7	60
Life Science	High School	30
Physical Science	High School	30

Ordered item booklets were prepared according to guidelines prescribed by Mitzel, Lewis, Patz, and Green (2001). Ordering of items was accomplished by (1) fitting an Item Response Theory (IRT) model to the test data, (2) determining RP (response probability) -67 values for each item, and (3) ordering items from easiest to most difficult on the basis of RP-67 values. In IRT, an RP-67 value represents the point along the latent trait continuum where an examinee would have a 67% chance of correctly answering the item.

All item response data were fit to the three-parameter logistic (3-PL) IRT model (e.g., Lord, 1980) using BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 2002). The 3-PL model relates the probability of success for examinee *j* on item *i* (i.e., the item response, u_{ij} = 1 instead of 0) as a function of examinee ability and three item parameters, as follows:

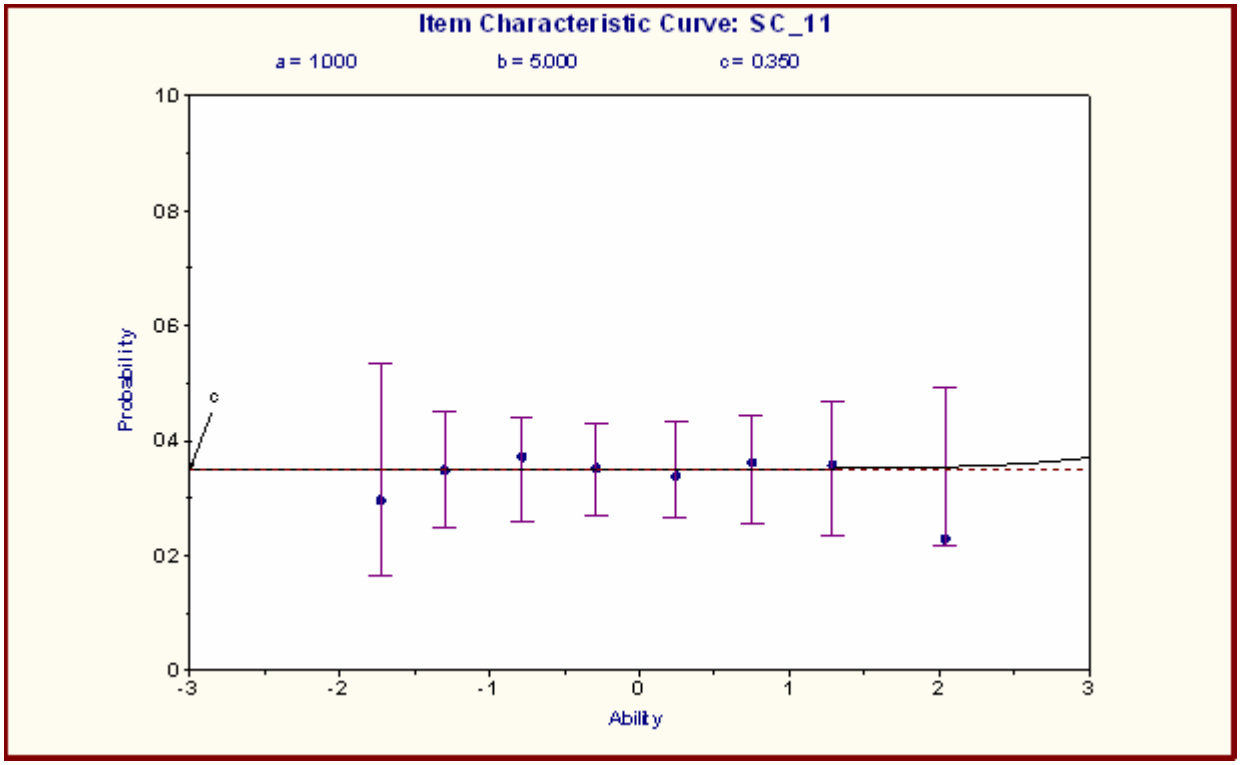
$$P(u_{ij}=1|\theta_j) = c_i + (1-c_i)\frac{e^{Da_i(\theta_j - b_i)}}{1+e^{Da_i(\theta_j - b_i)}},$$

where θ_j is the latent trait or ability parameter for examinee *j*, a_i is the slope or item discrimination parameter, b_i is the location or item difficulty parameter, c_i is the lower asymptote or guessing parameter, and *D* is a scaling constant equal to 1.7.

For some assessments, a small number of items (no more than four items for any one assessment) with poor statistical qualities (point-biserial correlations between item response and total test score approximately equal to zero or less) caused problems with the convergence of solutions. Inspection of item-true score regressions indicated that these items were difficult enough for examinees that no ability level performed better than chance-levels of success. As a result, maximum likelihood solutions for difficulty parameters could not be obtained. In this situation, the following procedure was employed: (1) all items were calibrated while excluding the items with poor statistical qualities, (2) item parameter estimates were fixed at their estimated values, and (3) the items with poor statistical quality were placed back in the dataset, with the a-parameter (discrimination) set equal to 1, the b-parameter (difficulty) set equal to 5, and the c-parameter (lower asymptote) estimated using BILOG-MG. This procedure produced item response functions like the example in Figure 3.1. This figure illustrates that the item characteristic curve is effectively flat in the region of ability where most examinees are located

(from -3 to 3 on a standardized metric). The only effect on the probability of success is the c-parameter. Items such as these, although they do not increase reliability nor the precision of ability measurement, were still included in the ordered item booklets so that no items would be eliminated. It should be noted that these items tended to be the most difficult on any assessment, and thus such items were always placed toward the very back of the ordered item booklets.

Figure 3.1. Example item characteristic curve for an item with poor statistical quality. This was item #11 on the High School KAMM Life Science Assessment.



Once all items were jointly calibrated, RP-67 values were calculated for each item, and then items were placed in ascending order on the basis of these values. RP-67 values were calculated using the following formula:

$$\theta_P = \frac{\ln\left[\frac{1-c_i}{P-c_i}-1\right]}{-Da_i} + b_i,$$

where θ_P is the ability level for an examinee for which $P(u_{ij}=1) = P$, P is the desired level of probability (in this case, $P = 0.67$), and all other terms have been defined previously. Table 3.2 contains an example of an item ordering on the basis of PR-67 values. A table like this was

created for each of the science assessments, and ordered item booklets were created on the basis of the item ordering.

Table 3.2
Item Ordering for the High School KAMM Science Assessment

Ordered Booklet	Item Number	RP-67
1	24	-0.4058
2	28	-0.3117
3	14	-0.2301
4	18	0.0990
5	21	0.3657
6	25	0.5486
7	10	0.7535
8	9	0.8236
9	29	0.8278
10	30	0.8696
11	12	0.8957
12	22	0.9648
13	27	1.2769
14	23	1.3227
15	15	1.3776
16	16	1.4125
17	17	1.6192
18	6	1.6925
19	26	1.8164
20	8	1.8618
21	19	1.9556
22	2	1.9809
23	20	1.9910
24	13	2.0166
25	5	2.0509
26	4	2.3833
27	3	2.7818
28	1	3.0098
29	7	3.1021
30	11	4.9819

(Note that item #11, the item demonstrated in Figure 1, is the last item on the list.)

Panel Participants

Invitation letters were sent to a random 50 percent of building principals in the state asking each to nominate a person. Some key language from the letter follows.

> *We are bringing together a group of Kansas educators to provide input toward setting performance standards (i.e., cut scores) for the new Kansas assessments in science. This is important work, and we plan to involve as many interested and available persons as possible. Below are the considerations for your building nominee.*
>
> 1. *Only nominate a person with whom you have spoken and that person has agreed to attend <u>both days</u> of scheduled meetings if selected.*
> 2. *The meetings will be held in Kansas City (meeting location was later changed to Overland Park) from 1PM to 5PM, Friday, June 20th, then continue on Saturday, June 21st starting at 8:30AM and ending by 2PM.*
> 3. *A person may only serve in one content area, the area in which he or she is nominated by you. Participants will receive travel and meal allowances along with a $175 stipend for their participation; lodging will be paid by CETE (double occupancy in the local hotel). Nominees will be notified by June 2 if they are selected, at which time details will be provided.*
>
> *From among the nominees, participants will be chosen based on expertise, instructional/supervisory experience, and qualifications at the specific content (science), grade (elementary, middle/jr. high, or high school), and test type (General or KAMM assessment) for the students being assessed. Other participation selection factors when considering prospective nominees will include:*
>
> - *highly regarded and respected local educators with at least three years experience;*
> - *instructional/supervisory experiences with students who have disabilities, students with limited English proficiency, and other subgroups;*
> - *balanced regional representation;*
> - *<u>access to an email address</u> to receive communication (even after schools close); and,*
> - *building or district administrators with qualifications are also eligible to be nominated.*
>
> *While we would like to involve as many educators as may be interested in this process; that is not possible. Based on KSDE information and random*

*selection principles, we request that you nominate <u>one</u> educator to represent your building who meets the qualifications and condition identified above **<u>at grade GG in science with the TEST</u>**. We invite you to nominate one person who, in your judgment, meets the experience, expertise, and training criteria to serve with other Kansas educators. Your nominee(s) must be highly qualified, held in esteem by peers, and have at least three years experience teaching at the grade and content area as a general or special educator (note: special educators may be nominated for the general assessment slots if desired, as many of their students take the general assessments). We rely on your professional judgment to nominate an individual who can help guide Kansas education and expectations for the future.*

If you do not have a person to nominate at the grade, content, and test area, please do not feel compelled to make a nomination.

Where ***<u>at grade GG in science with the TEST</u>*** is a placeholder for the grade, subject, and test type for which a nominee was sought at that school.

When sufficient numbers of nominees were not available from the first mailing, a second mailing was sent to the remaining 50% of the principals and the Department of Education directly contacted teachers who had previously served on state committees.

Each of 16 Kansas educators participated on one of three panels as part of the Bookmark process. Table 3.3 presents the number of participants on each panel. The same participants served on both the life and physical science panels.

Table 3.3
Number of Participants on Each Benchmark Panel

Grade and Subject	KAMM
Grade 4 Science	5
Grade 7 Science	5
HS Life Science	6
HS Physical Science	

Of these participants, 12 were female and 4 were male; 16 were Caucasian; 1 had fewer than two years teaching experience, 4 had 3-5 years experience, 4 had 6-10 years, 4 had 11-20 years, and 3 had more than 20 years; and 4 came from inner city schools, 1 from other urban, 4 from suburban, and 7 from rural.

Bookmark Results

After the third round of judgments, the average of the panelist cut scores was determined and transformed to the 0 to 100 reported score scale metric. Judgments were rounded to the nearest integer value. Table 3.4 presents the average Bookmark Procedure recommended score ranges for each performance category for the science assessments.

For the high school science assessment, the test is divided into two parts, one for life science and one for physical science. Students may take one or both parts, but a performance level assignment is not made until after both parts of the test are taken.

Table 3.4
Science Bookmark Procedure Recommended Performance Level Score Ranges

Performance Level	KAMM		
	4	7	HS
Academic Warning	0-39	0-35	0-28
Approaches Standard	40-58	36-48	29-44
Meets Standard	59-76	49-57	45-59
Exceeds Standard	77-90	58-65	60-69
Exemplary	91-100	66-100	70-100

Evaluation of Bookmark Procedure

At the end of the Bookmark Procedure meeting, participants were asked to evaluate the session. Key findings include: 100% of the participants in science found the training adequate or very adequate, and 89% of the science participants were comfortable with the final assignments of cut scores.

4. Standard Setting Policy Advisory Group

On July 26, 2008, a one day policy advisory group meeting was held in Topeka, Kansas. While the Bookmark procedure attempted to align cut scores with performance level descriptors, each test was reviewed by a separate panel, and consistency across grades or test types was not considered. Also, past experience suggests that, on occasion, standard setting panel results can be overly driven by a minority of participants who state their positions forcefully. The purpose of the meeting was to review the results of the Bookmark procedure for reasonableness and consistency.

Advisory Group Participants

Two members from each of the 14 Bookmark panels (14 table leaders and 14 who were nominated by their peers) were invited to participate in the policy advisory meeting to ensure that the process and results would be well represented. Of these 28 invitees, 27 were available and participated. An additional 29 people representing a variety of constituency groups also participated. Demographically, the 56 participants included 35 classroom teachers, three building administrators, 11 district administrators, five parents/grandparents of Kansas students, and two representatives of state educational organizations. These same 56 members included 51 Caucasians, three African Americans, one Hispanic, and one Native American. Thirty-two of these participants were female and 24 were male. Rural and urban, and eastern, central, and western areas of the state were all represented.

Policy Advisory Meeting Agenda

Following are the steps that took place during the science policy advisory meeting.

- *Introductions, logistics, and agenda*.
- *Context and purpose*. Participants were provided with an overview of the state assessment program and the steps that had occurred so far. Issues of consistency were discussed as was the task of ensuring an appropriate level of consistency and recommending specific cut scores the Kansas Department of Education.
- *Review of performance level descriptors*. Performance level descriptors were reviewed to provide further grounding and to ensure that the external consistency issues they were considering (cross-grade and cross-test type) were within the context of consistency with performance level descriptors.
- *Description of Bookmark standard setting process*. The Bookmark procedures were described so that advisory group members who did not participate in the Bookmark process would nonetheless understand it.
- *Results of Bookmark process*. The Bookmark recommended cut scores were presented to the participants.
- *Recommended consistent performance standards*. Using procedures explained below, panelists were presented examples of cut scores adjusted for inconsistencies within subject but across grades and two of the test types, general and KAMM. It was stressed that this was an example and that participants could indicate they agreed with the Bookmark assigned cut scores, with an example of more consistent scores, or with a different cut score. In order to make the task reasonable, data were presented both in terms of cut scores and also percent of students who would fall into each performance level. Table 3.5 and 3.6 and Figure 3.2 present some of the kinds of information presented to the participants. At the end of this presentation, the participants made their recommendations in terms of what percent of students they believed should be in each category.

Procedure for Creating Example of More Consistent Standards

To create the examples of cut scores that were more consistent across grades, the following procedures were followed.

1. Cut scores were transformed to the z-score corresponding to the proportion of students at or below that test score. For example, if 84% of the sample scored at or below the recommended cut score, the corresponding z-score was 1.0
2. The weighted average of the z-scores was taken, giving 50 percent weight to the z-score for the cut score under consideration and 25% for each of the other two grades. For example, if the z-scores for grades 4, 7, and HS were 0.6, 0.8, and 0.9, respectively, then the resulting z-score for grade 4 would be (0.6 x 0.5) + (0.8 x 0.25) + (0.9 x 0.25), or 0.725.
3. The proportion of a population corresponding to the weighted average z-score was calculated.
4. The raw score that has a cumulative frequency closest to the proportion from step 3 was selected as the more consistent example.

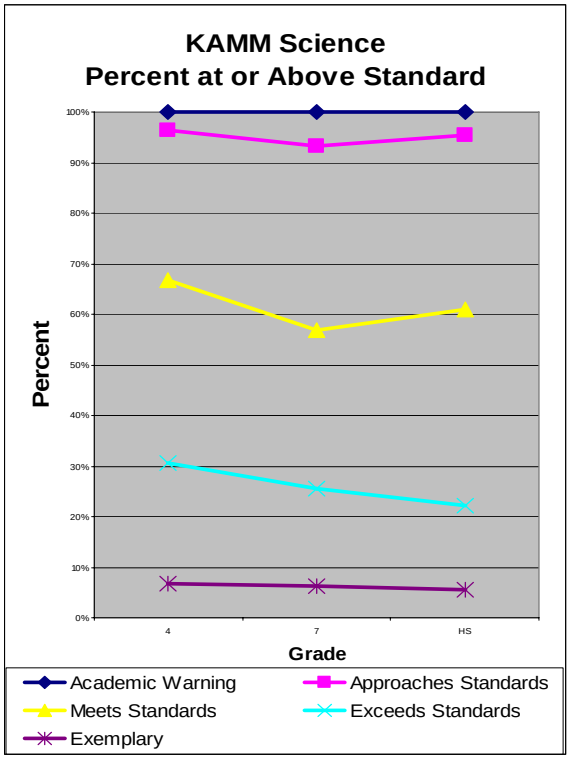
Table 3.5
General Science Effect of Cross-Grade Consistency on Scaled Cut Scores

Performance Level	Bookmark Recommended Maximum Possible Scaled Score in Performance Level			Cross-Grade Consistent Maximum Possible Scaled Score in Performance Level		
	4	**7**	**HS**	**4**	**7**	**HS**
Academic Warning	27	36	26	34	30	25
Approaches Standard	43	50	43	53	46	38
Meets Standard	66	67	75	76	68	63
Exceeds Standard	86	81	90	89	83	80
Exemplary	100	100	100	100	100	100

Table 3.6
General Science Effect of Cross-Grade Consistency on Percent in Category

Performance Level	Bookmark Recommended Percent in Category			Cross-Grade Consistent Percent in Category		
	4	7	HS	4	7	HS
Academic Warning	0%	4%	2%	1%	2%	2%
Approaches Standard	3%	15%	20%	8%	12%	13%
Meets Standard	19%	33%	64%	36%	39%	50%
Exceeds Standard	49%	30%	12%	40%	33%	28%
Exemplary	29%	17%	1%	16%	14%	7%

Figures 3.2 Examples of more consistent cut score graphs presented to participants



- *Other relevant information*. After participants indicated the percent of students they believed should be in each category, this information was tallied and presented. In addition, the percent if students in each category for the science general and KAMM assessments was presented. Participants were divided into groups of six to eight to discuss the data and were individually asked, in light of any information that came out during their discussions, to recommend the percent of students who should be in each category. It was stressed that the purpose of the discussion was to make sure everyone understood the various points of view, but that there was no need to come to consensus; each participant could make the set of recommendations that he or she saw fit.

Alternate Assessments

For several reasons a separate process was held for the alternate assessments. First, unlike for the general and KAMM assessments, the alternate assessments used a very different set of performance level descriptors. Moreover, the assessment process was completely different. Students participating in the alternate assessment worked with teachers to create a portfolio demonstrating competence for only three indicators per discipline. If a child met standards for those three indictors, they could not use them in the subsequent year – instead three different indicators would be chosen. Finally, portfolios for all disciplines were graded using a common rubric. Based on these considerations, a previous decision was made to use the same cut scores for the alternate assessments in reading and mathematics.

In a meeting of special educators held in Topeka (June 30, 2008, as part of the alternate portfolio scoring review) the consensus of the participants was that the same cut scores applied to Reading and Mathematics should also be used for History and Government and Science. This recommendation was presented to the Standard Setting Policy Advisory Group along with data as to how many students would fall into each category (presented in Table 3.7). No attempt to attain consensus was made, but a strong majority of participants agreed that the percent of students in each category consistent with the existing cut scores be maintained (between 75 and 82 percent).

Table 3.7
Alternative Assessment Results Maintaining Existing Cut Scores across Subjects and Grades

Performance Level	Score Range	Subject			
		Read.	Math	Sci.	H&G
Academic Warning	< 3.00	1.4%	2.8%	1.2%	1.1%
Approaches Standard	3.00 – 3.74	7.5%	8.1%	6.1%	6.6%
Meets Standard	3.75 – 4.24	15.7%	15.6%	12.8%	14.6%
Exceeds Standard	4.25 – 4.79	38.7%	38.1%	35.4%	36.4%
Exemplary	4.80 – 5.00	36.7%	35.4%	45.7%	41.3%

Resulting Recommendations from Standard Setting Policy Advisory Group

Table 3.8 presents the score ranges corresponding to the average recommended percent of students in each performance level from the 56 members of the Standard Setting Policy Advisory Group.

Table 3.8
Standard Setting Policy Advisory Group Recommended Performance Level Score Ranges for Science

Assessment Type	Grade	Academic Warning	Approaches Standard	Meets Standard	Exceeds Standard	Exemplary
KAMM	4	0-40	41-62	63-79	80-90	91-100
KAMM	7	0-30	31-43	44-55	56-67	68-100
KAMM	HS	0-27	28-40	41-53	54-65	66-100
Alternate	All	0.00-2.99	3.00-3.74	3.75-4.24	4.25-4.79	4.80-5.00

Evaluation of Standard Setting Policy Advisory Group Meeting

At the end, participants evaluated the meeting, and 95 percent of the participants found the training "adequate" or "more than adequate," while none found the training "not adequate." Some participants found the task of considering consistency issues to be very difficult. Table 3.9 presents the results when participants were asked how confident they were in their final decisions. All 56 people participated in all decisions, though not all responded to each question.

Table 3.9
Percent of Participants Indicating Confidence Level for Final Cut Score Decisions

Performance Category	No Response	Not Confident	Partially Confident	Confident	Very Confident
KAMM Science	4	5	21	38	32
Alternate Assessment	11	13	14	27	36

Although in all cases, more than 60% of the participants were confident or very confident in their decisions, confidence was lower for the alternate assessment and KAMM than for the general assessment. Written comments suggested that participants who had been members of the Bookmark panel had greater difficulty wrestling with these issues (perhaps because they had previously invested two days in that process). For future standard settings, it is recommended that Bookmark panel participants be advised in advance as to the difference in purposes between their task and the task of the policy advisory group. Also, perhaps the proportion of policy advisory group participants chosen from Bookmark participants should be limited to about 20 percent.

5. State Board of Education Adoption of Performance Standards

The performance level recommendations of the policy advisory group were reviewed by KSDE and submitted to the Kansas State Board of Education at its August 12, 2008 meeting. The performance level standards were accepted unanimously.

Section 4

RELIABILITY ANALYSES

Score Reliability

Detailed information on the reliability of test scores for each general assessment test form can be found in Section 2, Table 2.1. The information is condensed and presented below in Table 4.1. The score reliability estimates reported in the tables are Cronbach alpha coefficients. The coefficient values range from a low of 0.53 to a high of 0.82 across all the science KAMM forms. The overall general standard errors of measurement on the percent correct score scale range from 5.92 to 9.83 for scores on the KAMM Science Assessments.

Table 4.1
Science KAMM Reliabilities by Grade

Grade	Form	(α) Reliability	SEM % Correct
4	128950	0.82	6.31
7	128951	0.76	5.92
11 Life	128952	0.53	9.83
11 Physical	128953	0.55	8.32

The reliability of the composite high school science test was estimated as 0.73 with a corresponding standard error of measurement of 6.1. The estimation was based on the formula for the reliability of a composite score.

$$rel = 1 - \frac{\sum sem_i^2}{s^2}$$

Where sem_i^2 is the variance error of measurement of component i and s^2 is the variance of the composite score.

Classification Consistency for Science KAMM

Since the Kansas Assessment program is standards-based, it categorizes students into performance levels. The five performance levels are used to provide feedback to students, parents, and teachers and serve as the basis of accountability decisions. To help provide context for all of these uses, it is important to provide estimates of the consistency and accuracy of these categorizations. Consistency tells us how likely it is that a student categorized in a particular performance category would be categorized in that same category if that student took another form of the test. Accuracy tells us that if we knew the category to which a student truly belonged, what the probability is that the student would be so categorized when he or she took the test. Since consistency estimates contain two sources of error (one for each observed classification decision), but accuracy decisions contain only one (the hypothetical true classifications have no error), accuracy estimates are usually higher.

As stated in standard 2.15 in the current *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999): "When a test or combination of measures is used to make categorical decisions, estimates should be provided of the percentage of examinees who would be classified in the same way on two applications of the procedure, using the same form or alternate forms of the instruments" (p. 35).

Method

Classification indices were estimated by assuming a four-parameter beta compound binomial strong true score model (Hanson, 1991; Lord, 1965). The basic role of the psychometric model is to estimate the latent true score distribution and predict the observed score distribution. Then, classification consistency can be calculated based on the joint predictive probability of falling in the same performance category over two testing occasions, based on the estimated parameters of the true score model. Similarly, classification accuracy can be calculated based on the joint predictive probability of falling in the same performance category based on both observed and true test cut scores. The parameters of the true score model were estimated based on the actual data from a given base form at a particular grade and subject.

The BB-CLASS program (Brennan, 2004) was used to estimate consistency and accuracy using both the Hansen and Brennan (1990) and Livingston and Lewis (1995) approaches.

Procedures

Samples. A single KAMM science form was administered at each grade level (4, 7, and HS). For high school, performance levels were applied to a scaled composite score based on the average scaled score of the life sciences and physical sciences tests (each of which contained 30 items). Students can take one or both of these tests at one administration or take the two tests at different administrations. Only students who took both tests were included in the score distribution for the classification consistency and accuracy analyses.

For HS classification, consistency and accuracy were based on the distribution of total scores estimating the reliability of the composite score from the reliabilities of the life science and physical science tests. Table 4.2 presents descriptive statistics for the three analyses.

Table 4.2
Descriptive Statistics for Each Grade Analysis

Statistic	Grade 4	Grade 7	High School
n	738	845	669
Max possible	42	60	100
Mean	29.71	28.98	45.48
Variance	38.97	52.20	134.68
Reliability	0.82	0.76	0.69

Results

Table 4.3 summarizes the consistency and accuracy results for the three grade levels.

Table 4.3

Summary of Consistency and Accuracy Results

Approach		Grade 4	Grade 7	High School
Hanson & Brennan	Consistency	0.58	0.51	0.59
	Accuracy	0.69	0.63	0.71
Livingston & Lewis	Consistency	0.58	0.51	0.49
	Accuracy	0.69	0.63	0.62

Because the most critical decision for school accountability is whether a student is correctly classified as being at or above the Meets Standard performance level, Table 4.4 presents the results for that binary classification decision.

Table 4.4

Summary of Consistency and Accuracy Results, Academic Warning or Approaches Standard versus Meets Standard or Exceeds Standard or Exemplary

Approach		Grade 4	Grade 7	High School
Hanson & Brennan	Consistency	0.90	0.79	0.80
	Accuracy	0.93	0.85	0.86
Livingston & Lewis	Consistency	0.84	0.79	0.73
	Accuracy	0.89	0.85	0.80

Conditional Standard Errors of Measurement

The classical test theory standard errors of measurement (SEM) is calculated using both the standard deviation and the reliability of test scores. It is important to note that the classical SEM index only provides an estimated average test score error for all students regardless of individual proficiency levels. However, standard errors of measurement are different at different score levels. For this reason, it is useful to report not only a test level SEM estimate, but also an individual score level estimate. Individual score level estimates of error are commonly referred to as conditional standard errors of measurement (CSEM). The *Standards for Educational and Psychological Testing* (1999) recommends that test publishers provide CSEMs.

Procedure

Sample

The analysis of reliability was based on samples of students who were administered the Kansas KAMM assessments in science via the computer or paper-and-pencil. In 2008, test forms were constructed for grades 4, 7, and 11 (Life Science and Physical Science). There was one test form per grade except for grade 11 which had two test forms total, one each for life science and physical science.

Method

The binomial model for estimating both individual score-level CSEM and scaled-level CSEM was used because the tests consisted of dichotomously scored items. A modification method of estimation proposed by Keats (1957) for the error variance derived under the binomial error model of Lord (1955) was used. The raw score CSEMs ($\hat{\sigma}_{E \cdot X_p}$) were estimated by using Keats' (1957) modification equation:

$$\hat{\sigma}_{E \cdot X_p} = \sqrt{\frac{(n - X_p)(X_p)(1 - \hat{\rho}_{XX'})}{(n-1)(1 -_{21} \hat{\rho}_{XX'})}}$$

where *n* is the number of items on the test, X_p is the individual raw score, $\hat{\rho}_{XX'}$ is the most defensible estimate of reliability for the test, and ${}_{21}\hat{\rho}_{XX'}$ is Kuder-Richardson 21 for the test, which is expressed as

$$ {}_{21}\hat{\rho}_{XX'} = \left(\frac{n}{n-1} \right) \left(1 - \frac{\mu_x (n - \mu_x)}{n \sigma_x^2} \right) $$

where μ_x is the total test mean, and σ_x^2 is the total test variance. Keats recommended a parallel forms coefficient for $\hat{\rho}_{XX'}$, but in practice it might be necessary to use Cronbach alpha coefficients (Feldt & Brennan, 1989, pp. 123-124).

Because the Kansas Science KAMM Assessments results are not reported in terms of raw scores but rather in terms of scaled scores, the raw score CSEM had to be converted to a scaled score CSEM. A scaled score CSEM is simply the raw score CSEM multiplied by 100 and then divided by the number of items (*n*) on the test.

Results

Conditional Standard Errors of Measurement (CSEM)

Both a raw score CSEM and a scaled score CSEM were estimated for each grade. The results are presented in Tables 4.5 - 4.8. For each grade, the general trends are parabolic, which is concave downward. The peaking of CSEM occurs in the middle of the score range. Because the variance of binomial distribution is maximized when the probability of getting an item correct equals 0.5, the CSEM for the number of correct scores is usually greatest in this range, and scores are less reliable in this range. The distribution of scaled score CSEMs for each grade is summarized in Figure 4.1.

Table 4.5.
Conditional Standard Errors of Measurement (CSEM) for Grade 4 KAMM Science

Raw Score	Raw Score CSEM	Reported Scaled Score	Scaled Score CSEM
0	0.0	0	0.0
1	0.9	2	2.2
2	1.3	5	3.1
3	1.6	7	3.7
4	1.8	10	4.3
5	2.0	12	4.7
6	2.1	14	5.1
7	2.3	17	5.4
8	2.4	19	5.7
9	2.5	21	6.0
10	2.6	24	6.2
11	2.7	26	6.4
12	2.8	29	6.6
13	2.8	31	6.7
14	2.9	33	6.9
15	2.9	36	7.0
16	3.0	38	7.1
17	3.0	40	7.1
18	3.0	43	7.2
19	3.0	45	7.2
20	3.1	48	7.3
21	3.1	50	7.3
22	3.1	52	7.3
23	3.0	55	7.2
24	3.0	57	7.2
25	3.0	60	7.1
26	3.0	62	7.1
27	2.9	64	7.0
28	2.9	67	6.9
29	2.8	69	6.7
30	2.8	71	6.6
31	2.7	74	6.4
32	2.6	76	6.2
33	2.5	79	6.0
34	2.4	81	5.7
35	2.3	83	5.4
36	2.1	86	5.1
37	2.0	88	4.7
38	1.8	90	4.3
39	1.6	93	3.7
40	1.3	95	3.1
41	0.9	98	2.2
42	0.0	100	0.0

Table 4.6.

Conditional Standard Errors of Measurement (CSEM) for Grade 7 KAMM Science

Raw Score	Raw Score CSEM	Reported Scaled Score	Scaled Score CSEM	Raw Score	Raw Score CSEM	Reported Scaled Score	Scaled Score CSEM
0	0.0	0	0.0	31	3.7	52	6.1
1	0.9	2	1.6	32	3.7	53	6.1
2	1.3	3	2.2	33	3.7	55	6.1
3	1.6	5	2.7	34	3.7	57	6.1
4	1.8	7	3.1	35	3.6	58	6.1
5	2.0	8	3.4	36	3.6	60	6.0
6	2.2	10	3.7	37	3.6	62	6.0
7	2.4	12	3.9	38	3.6	63	5.9
8	2.5	13	4.2	39	3.5	65	5.9
9	2.6	15	4.4	40	3.5	67	5.8
10	2.7	17	4.6	41	3.4	68	5.7
11	2.9	18	4.8	42	3.4	70	5.6
12	3.0	20	4.9	43	3.3	72	5.5
13	3.0	22	5.1	44	3.3	73	5.4
14	3.1	23	5.2	45	3.2	75	5.3
15	3.2	25	5.3	46	3.1	77	5.2
16	3.3	27	5.4	47	3.0	78	5.1
17	3.3	28	5.5	48	3.0	80	4.9
18	3.4	30	5.6	49	2.9	82	4.8
19	3.4	32	5.7	50	2.7	83	4.6
20	3.5	33	5.8	51	2.6	85	4.4
21	3.5	35	5.9	52	2.5	87	4.2
22	3.6	37	5.9	53	2.4	88	3.9
23	3.6	38	6.0	54	2.2	90	3.7
24	3.6	40	6.0	55	2.0	92	3.4
25	3.6	42	6.1	56	1.8	93	3.1
26	3.7	43	6.1	57	1.6	95	2.7
27	3.7	45	6.1	58	1.3	97	2.2
28	3.7	47	6.1	59	0.9	98	1.6
29	3.7	48	6.1	60	0.0	100	0.0
30	3.7	50	6.1				

Table 4.7

Conditional Standard Errors of Measurement (CSEM) for Grade 11 KAMM Life Science

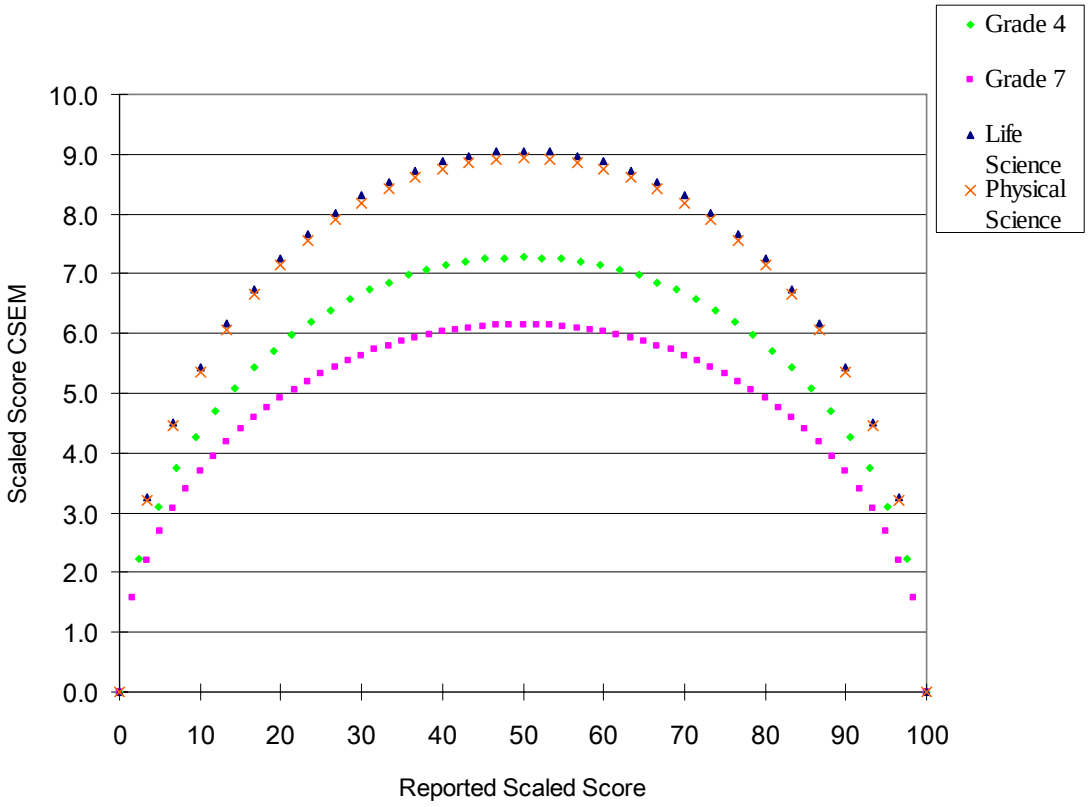
Raw Score	Raw Score CSEM	Reported Scaled Score	Scaled Score CSEM
0	0.0	0	0.0
1	1.0	3	3.3
2	1.4	7	4.5
3	1.6	10	5.4
4	1.8	13	6.2
5	2.0	17	6.8
6	2.2	20	7.2
7	2.3	23	7.7
8	2.4	27	8.0
9	2.5	30	8.3
10	2.6	33	8.5
11	2.6	37	8.7
12	2.7	40	8.9
13	2.7	43	9.0
14	2.7	47	9.0
15	2.7	50	9.1
16	2.7	53	9.0
17	2.7	57	9.0
18	2.7	60	8.9
19	2.6	63	8.7
20	2.6	67	8.5
21	2.5	70	8.3
22	2.4	73	8.0
23	2.3	77	7.7
24	2.2	80	7.2
25	2.0	83	6.8
26	1.8	87	6.2
27	1.6	90	5.4
28	1.4	93	4.5
29	1.0	97	3.3
30	0.0	100	0.0

Table 4.8.

Conditional Standard Errors of Measurement (CSEM) for Grade 11 KAMM Physical Science

Raw Score	Raw Score CSEM	Reported Scaled Score	Scaled Score CSEM
0	0.0	0	0.0
1	1.0	3	3.2
2	1.3	7	4.5
3	1.6	10	5.4
4	1.8	13	6.1
5	2.0	17	6.7
6	2.1	20	7.1
7	2.3	23	7.6
8	2.4	27	7.9
9	2.5	30	8.2
10	2.5	33	8.4
11	2.6	37	8.6
12	2.6	40	8.8
13	2.7	43	8.9
14	2.7	47	8.9
15	2.7	50	8.9
16	2.7	53	8.9
17	2.7	57	8.9
18	2.6	60	8.8
19	2.6	63	8.6
20	2.5	67	8.4
21	2.5	70	8.2
22	2.4	73	7.9
23	2.3	77	7.6
24	2.1	80	7.1
25	2.0	83	6.7
26	1.8	87	6.1
27	1.6	90	5.4
28	1.3	93	4.5
29	1.0	97	3.2
30	0.0	100	0.0

Figure 4.1. ***Conditional Standard Errors of Measurement (CSEM) for KAMM assessments in science by grade***



Section 5

VALIDITY

Validity is one of the most important attributes of assessment quality. It refers to the appropriateness or correctness of inferences, decisions, or descriptions made from test results about what students know and can do and is one of the fundamental considerations in developing and evaluating tests (AERA/APA/NCME, 1999). It is a complex construct that resides not in tests, but in the relationships between any test score and its context (including the instructional practices and the examinee), the knowledge and skills it is to represent, the intended interpretations and uses, and the consequences of its interpretation and use. Therefore, validity is not based on a single study or type of study but instead should be considered an ongoing process of gathering evidence supporting every intended interpretation and use of the scores resulting from a measurement instrument. As validity is not a property of a test or a test score or even of an interpretation, inference, or use of a test score, it cannot be captured conclusively. Rather, a judgment must be made regarding whether a body of evidence supports specific test claims and uses. This process begins with the test design and continues throughout the entire assessment process, including design, content specifications, item development, psychometric quality, and inferences made from the results.

While the primary evidence for the validity of the Kansas Assessments lies in the processes used to develop and design the system, it is also informative to collect evidence related to the degree to which a test correlates with one or more outcome criteria, or what is called criterion-related validity evidence. This type of validity evidence is needed to support inferences about an individual's current or future performance by demonstrating that test scores are systematically related to other indicators or criteria. The key is the degree of relationship between the assessment items or tasks and the outcome criteria. To help ensure a good relationship between the assessment and the criterion, the criterion should be relevant to the assessment, and it should also be reliable. Two analyses documenting the criterion-related validity evidence of the KAMM Assessment scores are detailed below.

INTERNAL EVIDENCE FOR THE VALIDITY OF KANSAS ASSESSMENT OF MODIFIED MEASURES (KAMM) ASSESSMENT SCORES

Benchmark Intercorrelations

The analysis of test structure based on examinee response data is an important part of test development and evaluation. A simplistic approach to conducting such an analysis is to determine the relationships among subsections of a test. Items on the KAMM Assessments are based on benchmarks (learning outcomes); there are 21 benchmarks measured on the grade 4 science assessment, 30 at grade 7, and 15 benchmarks measured on each grade 11 assessment (i.e., life science and physical science). There are two items that correspond to each benchmark across every grade. Results from such analyses provide empirical evidence for understanding the basic internal structure of the domain being measured. Evaluating the associations across these benchmarks, which are defined by groups of related items, is one method for providing evidence for construct validity. It is expected that these relationships would be low to moderate because while they are all intended to measure the same construct (i.e., science ability), they are simultaneously measuring different aspects of that construct. The size of the correlation coefficient between these scores will indicate the strength of the relationship between the criteria. Generally, the higher the correlation coefficient between the scores, the more valid the test is at assessing the criterion (Carmines & Zeller, 1979, p. 18).

Method

In the Kansas Science Assessments, there were four benchmarks on each test form per grade. Each benchmark tests a specific content area, which include scientific processes and connections, physical science, life science, and earth/space science in benchmarks 1, 2, 3, and 4, respectively. Table 5.1 below displays the number of items per benchmark at each grade level. Pearson Product-Moment Correlations were calculated among sub-domain scores at the benchmark level.

Table 5.1
Number of Items per Benchmark by Grade Level

Grade	Scientific Processes & Connections	Physical Science	Life Science	Earth/Space Science
4th	12	14	6	10
7th	16	18	16	10
HS	8	18	26	8

Results

Intercorrelations between benchmarks were calculated for the KAMM form at each grade in Science and are presented in Tables 5.2 to 5.5. Pearson product-moment correlations were calculated using the total scores for each benchmark. The observed intercorrelations for grade 4 ranged from 0.451-0.576, for grade 7 from 0.282-0.499, and for grade 11 from 0.212-0.371.

Table 5.2
Benchmark Intercorrelation Coefficients for 2008 Grade 4 Science KAMM

N = 740	**Scientific Processes & Connections**	**Physical Science**	**Life Science**
Physical Science	0.576		
Life Science	0.456	0.451	
Earth/Space Science	0.519	0.508	0.463

Note All correlations are significant at the 0.01 level (2-tailed)

Table 5.3
Benchmark Intercorrelation Coefficients for 2008 Grade 7 Science KAMM

N = 850	**Scientific Processes & Connections**	**Physical Science**	**Life Science**
Physical Science	0.402		
Life Science	0.499	0.385	
Earth/Space Science	0.380	0.282	0.430

Note All correlations are significant at the 0.01 level (2-tailed)

Table 5.4
Benchmark Intercorrelation Coefficients for 2008 Grade 11 Life Science KAMM

N = 1006	**Scientific Processes & Connections**	**Life Science**
Life Science	0.371	
Earth/Space Science	0.212	0.313

Note All correlations are significant at the 0.01 level (2-tailed)

Table 5.5
Benchmark Intercorrelation Coefficients for 2008 Grade 11 Physical Science KAMM

N = 915	**Scientific Processes & Connections**	**Physical Science**
Physical Science	0.331	
Earth/Space Science	0.271	0.245

Note All correlations are significant at the 0.01 level (2-tailed)

Intercorrelations across Content Area Tests

Criterion validity refers to "how adequately a test score can be used to infer an individual's most probable standing on some measure of interest – the measure of interest being the criterion" (Cohen & Swerdlik, 2002, p. 160). A criterion is "defined as the standard against which a test or a test score is evaluated" (Cohen & Swerdlik, 2002, p. 160). An assessment of criterion validity is conducted by correlating the group scores of each criterion (i.e. science, mathematics, and reading), For example in the case of state assessments, assessing if there are meaningful relationships among science, mathematics, and reading scores for grade 4. The size of the correlation coefficient between these group scores will indicate the strengths of the relationships among the measures.

An evaluation of the KAMM assessment science scores' criterion validity includes assessing the relationships of total KAMM scores in the areas of science with mathematics and reading for grades 4 and 7 and the relationships of life science, physical science, and combined sciences (both life science and physical science) with mathematics and reading for high school.

The evaluation of the strength of test relationships was based on samples of students who were administered any form of the Kansas general assessments in a given subject administered either via the computer or paper and pencil. Only one test form for each grade is available for the Science KAMM Assessments.

Pearson product-moment correlations were calculated using the total score for science, mathematics, and reading for grades 4 and 7, and correlating the total scores for life science, physical science, and combined science with the total scores for mathematics and reading for high school.

In order to estimate the strength of relationship of the underlying construct, correlations were corrected for attenuation using the following formula:

$$r_{x_t y_t} = \frac{r_{xy}}{\sqrt{r_{xx} r_{yy}}}$$

where $r_{x_t y_t}$ is the estimated correlation between the true scores of the measures x and y, r_{xy} is the observed correlation, and r_{xx} and r_{yy} are the reliabilities of x and y, respectively.

Results from the validity assessments are displayed in Tables 5.6, 57, and 5.8, which detail the intercorrelations for grades 4, 7, and high school, respectively, including the observed correlations and the correlations corrected for attenuation.

Intercorrelations for grade 4 ranged from 0.600-0.616 (observed) and 0.728-0.729 (corrected), for grade 7 from 0.407-0.550 (observed) and 0.523-0.681 (corrected), and for high school from 0.386-0.455 (observed) and 0.558-0.659 (corrected).

At grades 4 and 7, the observed and corrected correlations between reading and both of the other content areas, science and mathematics, are slightly higher than the correlations between mathematics and science. There does not seem to be any obvious patterns across content areas at high school.

Table 5.6
Intercorrelations and Sample Size (N) for 2008 Grade 4 Science KAMM

Grade 4	**Observed Correlation (N)**		**Correlation After Correction for Attenuation**	
	Mathematics	**Reading**	**Mathematics**	**Reading**
Science	0.600(573)	0.612(686)	0.728	0.729
Mathematics		0.616(776)		0.729

Note All correlations are significant at the 0.01 level (2-tailed)

Table 5.7
Intercorrelations and Sample Size (N) for 2008 Grade 7 Science KAMM

Grade 7	**Observed Correlation (N)**		**Correlation After Correction for Attenuation**	
	Mathematics	**Reading**	**Mathematics**	**Reading**
Science	0.407(677)	0.550(737)	0.523	0.681
Mathematics		0.498(785)		0.601

Note All correlations are significant at the 0.01 level (2-tailed)

Table 5.8
Intercorrelations and Sample Size (N) for 2008 High School Science KAMM

High School	**Observed Correlation (N)**		**Correlation After Correction for Attenuation**	
	Mathematics	**Reading**	**Mathematics**	**Reading**
Life Science	0.421(519)	0.433(565)	0.659	0.623
Physical Science	0.386(443)	0.397(531)	0.593	0.561
Combined Sciences	0.451(1,239)	0.455(729)	0.602	0.558
Mathematics		0.392(710)		0.468

Note All correlations are significant at the 0.01 level (2-tailed)

References

AERA, APA, & NCME (1999). *Standards for educational and psychological testing*. Washington, D.C.: Author.

Brennan, R. L. (2004). BB-CLASS: A computer program that uses the beta-binomial model for classification consistency and accuracy (Version 1.0) (CASMA Research Report No. 9). [Computer software and manual]. Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, The University of Iowa. (www.education.uiowa.edu/casma).

Carmines, E. G., & Zeller, R. A. (1979). Reliability and validity assessment. *Quantitative Applications in the Social Sciences, 17.* Sage Publications.

Cohen, R. J., & Swerdlik, M. E. (2002). *Psychological testing and assessment: An introduction to test and measurement* (5th ed.). Boston: McGraw-Hill.

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 123-124). Phoenix, AZ: Ornyx.

Hanson, B. A. (1991). *Method of moments estimates for the four-parameter beta compound binomial model and the calculation of classification consistency indexes*. ACT Research Report 91-5. Iowa City, IA: American College Testing.

Hanson, B. A., & Brennan, R. L. (1990). An investigation of classification consistency indexes estimated under alternative strong true score models. *Journal of Educational Measurement*, 27, 345-359.

Keats, J. A. (1957). Estimation of error variances of test scores, *Psychometrika*, *22*, 29-41.

Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement*, 32, 179–197.

Lord, F. M. (1965). A strong true-score theory, with applications. *Psychometrika, 30,* 239 – 270.

Lord, F.M. (1980). *Applications of Item Response Theory to Practical Testing Problems*. Hillsdale, NJ: Erlbaum.

Mitzel, H. C., Lewis, D. M., Patz, R. J., & Green, D. R. (2001). The Bookmark procedure: Psychological perspectives. In G. J. Cizek (Ed.), *Setting performance standards* (pp. 249-282). Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (2002). *BILOG-MG* [Computer software]. Chicago, IL: Scientific Software International, Inc.

Appendix A

KANSAS ALTERNATE ASSESSMENT (KAA) TECHNICAL CHARACTERISTICS

To provide a description of the Kansas Alternate Assessment procedures, the Implementation Guide found on the CETE website (www.cete.us) is reproduced as an introduction in Part 1. Following this introduction, available information addressing the reliability and validity of scores resulting from the KAA is provided as Part 2.

Part 1: Kansas Alternate Assessment Implementation Guide

The primary reference describing the Kansas Alternate Assessments is the *KS Alternate Assessment Teachers' Guide 08-09.* This document can be found by following the KS Assessment/KS Alt Assessment links at www.kansped.org, the KSDE Special Education website. The Teachers' Guide should be reviewed prior to undertaking an assessment. What is presented below is a summary of the requirements and procedures described in the guide. And more importantly, this document describes how to use tools available on the CETE website in administering Alternate Assessments.

As a quick overview, the following table identifies the major activities required to implement the Kansas Alternate Assessment.

I. Students meeting the criteria for receiving the Alternate Assessment must be identified.

II. Five indicators must be selected by the local IEP team as the focus of each content area assessment to be given.

III. For each content area assessment and for each of the five chosen indicators, three tasks must be defined that will allow the student to demonstrate level of mastery. Assembly of an Evidence Folio will begin with documentation of the assessment design.

IV. During the testing window (January 5 – April 14) the teacher will collect the evidence of student performance on the defined tasks and add it to the Evidence Folio.

V. Each piece of evidence in the folio will be independently rated by three local raters using Rater Worksheets.

VI. The Rater Worksheets will be used as entry forms to facilitate score entry on the CETE website and then added to the student's Evidence Folio.

VII. Evidence Files for each student will be maintained locally to document the student's progress.

Each of these activities is elaborated below.

I. Identification of Students Taking the Alternate Assessment

Use the *KS Alternate and KAMM Eligibility* document that can be found by following the KS Assessment/KS Alt Assessment links at www.kansped.org to determine who will be taking which Alternate Assessments. Students who would be eligible for Alternate Assessments, but enter the school on or after January 5, 2009, are exempt from state assessment for 2009.

Reading, Mathematics, Science and Writing assessments are mandated for the 2008-2009 school year. Reading and Mathematics are to be tested in grades 3 through 8 and once at the High School level. Science assessment will happen in grades 4 and 7 and once at the High School level; Writing, in grades 5 and 8 and once at the High School level. Government also is offered High School level, but is optional.

The types of assessment being given to students are identified through the KIDS system. Share your Alternate Assessment decisions with your district's staff that have KIDS system responsibility. Once a KIDS TEST collection containing the test type information for students has been successfully completed, any student designated as taking any alternate assessment will be available in the Alt. Assessment section of the CETE website. Note that to successfully register a high school student for a Science Alternate Assessment, a value of 3 (Alternate Assessment) must be entered in KIDS field D76 (High School State Life Science Assessment); for a Government Alternate Assessment, a 3 must be entered in KIDS field D79 (High School State History/Gov. Assessment: World Focus).

II. Selection of Assessment Indicators for a Student

Use the documents found under the heading *KS Extended Standards* (by following the KS Assessment/KS Alt Assessment links at www.kansped.org) to develop a full understanding of the extended indicators available for the definition of Alternate Assessments. In support of educators trying to define Alternate Assessments, CETE provides a link that will produce an EXCEL workbook with a sheet for each student to be tested in an academic area (or a selected subgroup of those students) showing the extended indicators available for the specific student. This is something that might be printed and carried to an IEP meeting to aid in discussion.

Since IEP conferences may be scheduled earlier than your district's KIDS staff can manage to get student names into KIDS, there is also an EXCEL workbook available on the CETE website that has the full set of indicators available in reading, mathematics and science at each grade level. The first work sheet in the workbook gives advice on how to find out which indicators should not be available for a specific student. It is suggested that you wait for the personalized lists when it is possible to do so, to eliminate the possibility of including inappropriate indicators.

Designing a Kansas Alternate Assessment requires that five extended indicators be selected for a student from each content area to be assessed. The indicators for a content area are grouped by Standards and the available indicators vary by grade level. In defining an Alternate Assessment care must be taken that the five chosen indicators adequately represent the various

Standards for the academic area. In Reading, there are two Standards: 1) Reading and 2) Literature. At least one indicator must be selected for a student's assessment in Reading from each of these two Standards. The other three indicators may be selected entirely from only one of the two Standards or may be distributed across the two Standards. In Mathematics, there are four Standard areas: 1) Numbers and Computation, 2) Algebra, 3) Geometry and 4) Data. Again, choose five; at least one from each area.

Science poses a special problem because there are seven Standards: 1) Science as Inquiry, 2) Physical Science, 3) Life Science, 4) Earth and Space Science, 5) Technology, 6) Science in Personal and Environmental Perspectives, and 7) History and Nature of Science. Not all Standards have indicators at every grade level, but there are always more than five Standards with indicators. Five of the available Standards must be chosen to be represented by an indicator. Writing has three Standards areas: 1) Writing, 2) Research, and 3) Communication for Social Interaction. Choose five indicators; at least one from each of the three areas.

Government (which is an option for High School in 2008-09) has but four Standards: 1) Government, 2) Economics, 3) Geography, and 4) History; choose five indicators with all Standards represented.

There are two further requirements concerning extended indicator selection. First, an indicator becomes unavailable for a student if the student achieved a performance level of Above Standard or Exemplary on that indicator in the previous year. Next, in Reading, Mathematics, Science and Government the extended indicators have been aligned with indicators being tested on the general assessment, and no two extended indicators that are aligned with a single general indicator may be chosen.

The Indicator Selection Page

To support Special Education staff in the selection of extended indicators CETE has a page on the CETE website for designating selected indicators. Only indicators available for a student are displayed while their indicators are being chosen. Both grade inappropriate indicators and personally unavailable indicators will not show on a student's list.

Here is the first segment of an indicator selection screen which illustrates the concepts laid out in the description that follows:

Standard ER1 - Reading: The student reads and comprehends text.

General Indicator R.5.1.3.1 - determines the meaning of words or phrases by using context clues (e.g., definitions, restatements, examples, descriptions) from sentences or paragraphs.

General Indicator R.5.1.3.1: The student

- ○ ER.1.2.1 - assigns meaning to natural gestures.
- ◉ ER.1.2.4 - assigns meaning to spoken words / manual signs.
- ○ ER.1.4.2 - understands multiple meanings of words.
- ○ None

On the selection page, the extended indicators are represented as groups of radio buttons; each group presented under the general indicator with which they are aligned. The radio buttons preclude the possibility of choosing more than one extended indicator aligned with a general indicator; clicking on a second aligned indicator deselects the first. In case an extended indicator is inadvertently chosen from a group aligned with a general indicator, the option "None" has been added to each group of aligned indicators to force deselection. It is not necessary to select "None" when choosing no extended indicator out of a group of extended indicators aligned with a particular general indicator; "None" is merely there to provide a mechanism for correcting accidental choices. It is the case that many general indicators are not represented on any given Alternate Assessment.

All of these groups of aligned indicators are arranged in list fashion under the Standard with which they are associated. This should make it possible to easily identify whether rules regarding distribution across Standards have been upheld.

Once five indicators have been chosen to define an assessment, they may be saved by scrolling to the bottom of the selection page and clicking the save button. The chosen indicators will be investigated to assure that the distribution across Standards is correct and that, indeed, five distinct extended indicators have been chosen. It is the case that some of the extended indicators are aligned with several general indicators. If the same extended indicator is inadvertently chosen from two different sets of radio buttons, this final check will identify the problem. If any problem is identified with the chosen indicators, a pink message box with an explanation of the problem appears on the screen. Only when there appears a green box with the message "The selected indicators have been saved." will it be true that indicators have been successfully chosen.

To choose indicators:

- Sign onto the CETE website, www.cete.us, using a Userid and Password that are of the Alt Assess District or Alt Assess Building type. To learn how to get a Userid if you have none, go to the site and click the "Need a new account?" link on the opening page just above the login boxes. The account registration process will be explained on screen.

- Once logged in, click on "Alt. Assessment" in the menu list on the left of the screen. A list of options will appear.
- Click on the "Students" option. This brings up a table of students for each building to which the user has access. One column is labeled "Indicators". This column shows the count of chosen indicators for each student, for each assessment being taken. The number will be either 5, indicating that a choice has been made; or 0, meaning that indicators have not been chosen. On the row with the student name and assessment for which you wish to choose indicators, click on the indicator count. This is a link to a checklist of the indicators already chosen for that student in that subject.
- On the checklist page, click the Modify Indicators button to choose or edit indicators. A list of all extended indicators available for the chosen student and academic area appears.
- Find the five indicators on the list that need to be selected for the student's assessment and click the radio buttons next to them.
- When five indicators have been selected, click the "Save Indicators" button at the bottom of the screen to complete the task. If an error has been made in selection (not exactly 5 indicators were chosen or not enough standards were represented), a pink error message explaining the problem appears on the screen. When no error exists, the click on "Save Indicators" will terminate the selection process.

III. Task Definition and the Evidence Folio

Three tasks must be defined to allow a student to demonstrate their skill level for each of the five indicators selected to constitute an Alternate Assessment. A separate Evidence Folio should be developed for each Alternate Assessment administered to a student. Each folio will ultimately contain 15 performance reports (the evidence) collected during the assessment window (January 5th to April 14th). Actual task definition is extensively discussed in the Teachers' Guide. Here is presented the description of how to build the task descriptions into an Evidence Folio:

1. The first page of the folio is a list of indicators to be assessed. This page may be printed from the "Alt. Assessment" section of the CETE website, www.cete.us. (This printed listing may serve as the Cover Sheet for the Evidence Folio.)

    - Once signed onto the site using an Alt Assess District or Alt Assess Building Userid and Password, click on "Alt. Assessment" in the menu list on the left side of the screen and a list of options will be presented.
    - Click "Students." This will bring up the "Alternate Assessment Student Management" Page. One feature of the page is a list of students registered for the Alternate Assessment. (If no student names are visible, no students have been designated in KIDS as needing Alternate Assessments. Talk to your KIDS administrator.)
    - Above the table of students is a green box with, among other things, a "Print Lists of Chosen Indicators" link. Click the link to bring up a printing-options list.
    - Links for printing all students or just chosen students in each subject will appear. Follow the screen directions to produce an EXCEL workbook with printable indicator lists.

2. There follows the Evidence Label and then the evidence for each of three pieces of evidence to be evaluated for each of the five indicators to be assessed.

    - The Evidence Label that precedes each piece of evidence describes the task and the assessment procedure for the piece of evidence. Evidence labels are made available on the CETE website in EXCEL workbooks. On the "Alternate Assessment Student Management" page click the "Print Evidence Labels" link, and on-screen instructions will guide you to an EXCEL workbook with Evidence Labels. Each label has the student, the indicator and the number of the piece of evidence identified. There are boxes on the label to enter the evidence description.
    - Behind each Evidence Label in the folio will be placed either student work that shows level of mastery or the test administrator's written description of student response that demonstrates the student's level of competence in addressing the task.

IV. Evidence collection.

The fifteen tasks that constitute an assessment will be presented to the student, and either student work that demonstrates level of mastery or a teacher description of student mastery will be collected.

The guidelines for the evidence collection are:

- When the student's task response is scored as correct/incorrect, a minimum of 5 trials/probes is required to complete a piece of evidence.
- Any written transcription of student responses must be verbatim.
- When using worksheets as data, 3 different worksheets may be used as 3 different pieces of evidence, but they may not be identical worksheets. Each must have 5 questions probing for student's level of performance.
- When responding to 3 different individuals, tasks, or environments, each encounter may be considered a different piece of evidence.
- The person collecting the evidence must not use the Alternate Assessment skill scoring rubric to make judgments about the student's performance level. The evidence is to be presented describing the assessment process and the response of the student, but judgment of performance level is in the purview of the rater.

As regards support that may be provided:

- The teacher should first ask the student to respond without support.
- If the teacher provides support, it should be documented on the Evidence Label.
- The support provided **should not exceed** the support provided during instruction.

- The teacher may use cues/prompts to direct the student's attention to the task to elicit a response, unless the target skill in the indicator calls for student attention, e.g., responds to stimuli. The teacher is **not** to lead the student to an answer or response with skill related cues/prompts.
- If the teacher is completing the task or leading the student to a correct response, that cannot be considered appropriate support for assessment.

  NOTE: Hand over hand assistance is considered appropriate for IEP goals, but is not considered appropriate support for the Alternate Assessment. If hand over hand assistance is required for a student to complete the task correctly, the data can only receive a rating of "1" on the skill rubric score scale.

V. Rating the Student's Skill Level Based on the Evidence Folio

Each piece of evidence in the folio is to be independently rated by three local raters. **A student's current special education teacher is a required scorer.** It is recommended that the other two scorers should be professionally licensed educators who do *not* work directly with the student. This will ensure a more objective review of the evidence. However, if only limited numbers of professional staff are available, then staff members who work directly with the student may be used.

Raters may be general education teachers, related service providers, other special education teachers, or administrators. Raters should be trained in the review, evaluation and scoring of student data folios. KSDE will be providing training. Information about the KSDE training dates, locations and materials can be found on www.kansped.org.

Checking for inter-rater reliability on evidence evaluations prior to rating students' evidence folios is an important component of the training process. However, checking for inter-rater agreement **should not occur** during the actual rating of student evidence folios. **These ratings must be done independently with one individual's ratings not shared with other raters.**

Workbooks for the collection of evidence ratings for each subject being assessed are available through the CETE website. These EXCEL workbooks should be downloaded in preparation for the rating process. A workbook's first worksheet has a list of "Rater Codes," a numbered list of the possible relationships that the rater might have to the student. A rater code is to be provided on each Rater Worksheet.

There follows a Rater Worksheet for each student; the student's name appears on the worksheet tab at the bottom of the EXCEL window. A worksheet has clearly labeled boxes for entering the scores for each of the three pieces of evidence for each of the five extended indicators that constitute an Alternate Assessment.

To download a workbook with Rater Worksheets:

- Sign onto the CETE website with an Alt Assess Building or Alt Assess District type Userid and Password.
- Click on "Alt. Assessment" in the menu list on the left side of the screen to get a list of options available.
- Click on the "Students" option to bring up a table of students registered for the Alternate Assessment.
- In the green box above the student table is the link “Print Rater Recording Worksheets.” Click on this link.
- A screen appears with options to generate all students’ Reading, Mathematics, Science, Government or Writing worksheets or to print the worksheets for selected students for an academic area. Click on the desired option. Worksheets will only be available for assessments that have already been defined by the selection of five indicators to be assessed.
- Print the materials that you need or save the workbook to your desktop to so it will be available for printing later. Take care to give any workbook being saved a unique name so it does not replace an existing workbook.

Below is an example of a worksheet.

RATER WORKSHEET FOR RECORDING KANSAS ALTERNATE SCIENCE ASSESSMENT
Student name: Rater Code: Rater name:

Indicator 1: ES.1.1.1 - explores objects and/or environments.

Record Evidence #1 Score:	Record Evidence #2 Score:	Record Evidence #3 Score:

Indicator 2: ES.2.1.1 - describes an object by one of its properties.

Record Evidence #1 Score:	Record Evidence #2 Score:	Record Evidence #3 Score:

Indicator 3: ES.3.2.3 - describes, compares, and/or contrasts diverse characteristics of living things.

Record Evidence #1 Score:	Record Evidence #2 Score:	Record Evidence #3 Score:

Indicator 4: ES.4.1.1 - explores earth materials.

Record Evidence #1 Score:	Record Evidence #2 Score:	Record Evidence #3 Score:

Indicator 5: ES.5.1.1 - understands cause and effect.

Record Evidence #1 Score:	Record Evidence #2 Score:	Record Evidence #3 Score:

How many of the pieces of evidence did you find difficult to rate? (circle one) 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15

In making their judgments about a student's skill level, the following rubric is to be used by the raters.

Skill Performance Rubric

1	2	3	4	5
Student displays **LITTLE OR NO** mastery of essential knowledge or performs this skill in 0 – 9% of the trials or probes	Student displays **LIMITED** mastery of essential knowledge or performs this skill In 10 – 29% of the trials or probes	Student displays **PARTIAL** mastery of essential knowledge or performs this skill in 30 – 69% of the trials or probes	Student displays **NEAR** mastery of essential knowledge or performs this skill in 70-89% of the trials or probes	Student displays **COMPLETE** mastery of essential knowledge or performs this skill in 90 – 100% of the trials or probes

No half points may be given.

Guidelines for rater scoring are as follows:

- Rate each piece of evidence separately using the rubric definitions and point values, and record your rating in the appropriate cell on your Rater Worksheet. Pay special attention that the evidence numbers on the Evidence Labels correspond to the evidence numbers on the Rater Worksheet. It is important that the ratings be recorded consistently in the same cells for the same pieces of evidence across raters.
- A rater is not to discuss their ratings or the contents of an evidence file with other raters. The ratings are to be done independently based on whatever evidence is in the Evidence Folio.
- When trials or probes are used as evidence of skill level, a minimum of 5 trials is required. If fewer than 5 trials are used, the missing trials should be counted as incorrect in calculating the rubric percentage of trials in which the skill is demonstrated. For example, if only three trials are presented in a piece of evidence and the student demonstrated the skill on two of the trials, one would still use 5 trials as the base in calculating a percent for use in assigning a point value on the rubric. In this example, evidence is available to demonstrate that the student exhibited the skill only two times. This would be 40 percent of the required minimum of 5 trials and the student should be awarded a rubric score of 3.
- As it is anticipated that many will use either a 5-trial or a 10-trial design for data collection, the following table shows the conversion for the number of trials correct to the value on the 5-point rubric scale shown above.

Rubric Score	1	2	3	4	5
# Correct					
5 trials	0	1 (20%)	2-3 (40-60%)	4 (80%)	5 (100%)
10 trials	0	1-2	3-6	7-8	9-10

- The level and type of support provided the student in responding is to be described on the Evidence Label for each piece of evidence. If the teacher is completing the task or leading the student to a correct response, it cannot be considered appropriate support for assessment and a rating of "1" on the skill rubric score scale should be made. Note also that hand over hand assistance is not considered appropriate support for the Alternate

Assessment. If hand over hand assistance is required for a student to complete the task correctly, the data can only receive a rating of "1" on the skill rubric score scale.

- The teacher may use cues/prompts to direct the student's attention to the task to elicit a response, unless the target skill in the indicator calls for student attention, e.g., "responds to stimuli." The teacher is not to lead the student to an answer or response with skill related cues/prompts. If the evidence file indicates they did so, a rating of "1" on the skill rubric score scale should be made.
- If one or more pieces of evidence is missing for an indicator (i.e., the folio does not have the required 15 pieces of evidence), check with the person who is managing the folio to see if evidence has been misplaced. Otherwise, assign a rating of "0" on the skill rubric score scale for all missing evidence pieces.

VI. Entering the Rater Scores

Once a rater has completed their ratings and recorded them on the Rater Worksheet, the worksheet should be placed in an envelope. Each district needs to develop procedures for getting these envelopes to the person designated to enter the data.

People made responsible for data entry need to register for an Alternate Entry type Userid. This may be done by going to the CETE website www.cete.us and clicking on the "Need a new account?" link. Adequate information to complete the registration process will be found on screen.

The Score Entry system is implemented so that an Alternate Entry userid can be used to enter or edit any data from the district that created the userid, but it is limited to only that activity on our website.

To enter scores, the assigned district/building personnel should do the following:

- Log onto the CETE website at www.cete.us using an Alternate Entry type Userid and Password.
- Click on "Alt. Assessment" in the left column menu list.
- Click on "Score Entry" when the options list appears.
- The opening page for score entry will allow the user to identify the building in which the student whose data is being entered was tested. The user will choose the building from a drop-down list and click on continue.
- Having established which building is of interest, a table of Alternate Assessments given in the building is presented. One sees the student name, building, grade, subject and a "Scores" link for each assessment. There is also room in this table for the names of raters whose data has already been entered. Click the "Scores" link for a particular test to begin data entry; click the name of a rater to edit their data.
- If editing data that has already been entered, clicking the rater's name will bring up a screen for verification of the rater code. Clicking continue on this screen brings up the data entry screen.

- If entering data, the next screen asks the user to choose the rater's code from a drop-down list and to identify the rater from a second drop-down list. Choose the rater code that appears on the Rater Worksheet. Next choose the rater's name from the name list. One of the options on the name list is "Add this rater". When the rater whose data is being entered is not on the drop-down list (and there will be no names when the score entry process begins), choose "Add this rater". Otherwise, choose the appropriate name from the list. In either case, click the Continue button.
- If a rater name is to be added, the next screen will have a text box for entry of the rater name and a Continue button. Enter the name; click the Continue button to bring up the data entry screen.
- Whether editing old data or entering new data, the data entry page should now be visible. The data entry form is laid out in a pattern similar to the Rater Worksheet, allowing the user to quickly enter the 15 ratings or locate numbers needing to be edited. If one or more pieces of evidence were not assessed, choose the "0 - Did not rate" option. For rapid entry, you may enter the rating numbers and tab from field to field rather than using the drop-down lists and mouse clicks.
- When finished entering the 15 score values, provide the rater's answer to the question about difficult questions at the bottom of the Rater Worksheet and then click on the "SAVE" button at the bottom of the screen. The values saved will be displayed for review and editing. At this point you may make corrections and SAVE again, click on the "DELETE" button to eliminate the entry or click on the "Back to Student List" link near the top of the window to end entry from the current Rater Worksheet.
- The list of students will appear again. If new data was being entered, the rater name will now appear in the student table as one of the raters entered. Once three raters have been entered for a given assessment, no more data may be added. Only three raters are allowed. If an incorrect set of data is entered for an assessment, it may be deleted by clicking the rater name as if to edit, and then using the delete button on the data entry screen.

VII. Locally Maintaining Student Evidence Folios

Student Evidence Folios are to be maintained and stored locally after the assessment is completed. The Rater Worksheets for each of the three raters should be added to the Folios. Folios should be retained for three years.

Part 2:
Alternate Assessment Reliability and Validity Information

Alternate Assessment Reliability Information

Science Score Reliability

1. The correlation coefficients among all raters' total rating scores for 1001 Alternate Assessment students were:
    a. rater1 vs. rater 2, .966
    b. rater1 vs. rater 3, .935
    c. rater 2 vs. rater 3, .968

2. The percent perfect agreement among all raters across the 15 indicators ranged from a low of 92.0 percent to a high of 95.2 percent of 1001 ratings per indicator. This percent agreement was over 98.5 percent when the criterion was "within one scale point."

3. The correlation between the average rating for the three local ratings and a fourth external "expert" reviewer of a sample of 36 student data folios was .890.

Science Alternate Assessment Performance Classification Reliability

As described in the Standard Setting section of the main part of this Technical Manual, score ranges on the KAA were established through standard setting procedures to classify students into five performance level categories (*Academic Warning*, *Approaches Standard*, *Meets Standard*, *Exceeds Standard*, and *Exemplary).* The five performance levels are used to provide feedback to students, parents, and teachers and serve as the basis of accountability decisions. To help provide context for all of these uses it is important to provide estimates of the consistency and accuracy of these categorizations. Consistency tells us how likely it is that a student categorized in a particular performance category would be categorized in that same category if they take another form of the test. Accuracy tells us if we knew the category to which a student truly belonged, what is the probability they would be so categorized when they took the test. Since consistency estimates contain two sources of error (one for each observed classification decision) but accuracy decisions contain only one (the hypothetical true classifications have no error), accuracy estimates are usually higher.

Procedures to estimate classification consistency and accuracy for the KAMM assessments mirrored those used for the general science assessment and KAMM science test forms with one exception. As the score ranges for classifying students into the performance level categories are the same for students at all grade levels, only one set of values is presented based on all Alternate Assessment student data as one group. The overall test classification consistency across all categories is .84. The classification accuracy coefficient is .88.

Because the most critical decision for school accountability is whether a student is correctly classified as being at or above the 'Meets Standard' performance level, it also is important to examine the classification consistency and accuracy for this binary classification decision. The classification consistency for this binary decision is .98. The classification accuracy coefficient also is .98. The reason these coefficients are so high is that the vast majority of Alternate Assessment students had scores above the lower cut-point for the 'Meets Standard' performance level.

Science Alternate Assessment Validity Information

As part of the Standard Setting (cut scores) process, a sample of student data folios were reviewed by a panel of "expert" judges, i.e., individuals who had been trained by KSDE and who had served as trainers in implementing and scoring the Alternate Assessment for other local personnel. In addition to scoring the sample of student data folios and making judgments on the performance level of the student based on the folios' evidence, the "expert" judges were also requested to supply validity judgments on the quality of the data in each folio in terms of: 1) the general **overall clarity** of the evidence samples in the Evidence Folio in allowing the judge to make ratings with confidence, 2) the **overall appropriateness** (fit as a measure of the indicator) of the assessment procedures, and 3) the **overall compliance** with Alternate Assessment requirements. The results from the review of student folios by the external panel of experts are presented below.

Independent Expert Panel Judgments

At the conclusion of the spring 2008 testing, a sample of Alternate Assessment folios was collected by KSDE. A four-day meeting was held in July, 2008 by KSDE staff for the purpose of examining completed student folios with the goal of making suggestions for improving the Alternate Assessment materials and procedures. A panel of 20 educators with knowledge and experience working with Alternate Assessment students and who had previously been trained in the Alternate Assessment procedures attended the meeting. .During the review, a sample of 42 student Science Alternate Assessment folios were reviewed and formally evaluated. As part of their review, these expert judges were asked to respond to the following questions addressing the quality (validity) of the student science data folios.

1. How would you rate the overall clarity of the evidence samples in the Evidence Folio in allowing you to make your ratings with confidence?
    a. I had no problems in making judgments for any of the evidence samples.
    b. I had difficulty in making judgments for a limited number of the evidence samples (1 – 2), but the rest were sufficiently clear.
    c. I had difficulty in making judgments for several of the evidence samples (3 – 7), but over half were sufficiently clear.
    d. I had difficulty in making judgments for more than half of the evidence samples (8 – 13) with only a few being sufficiently clear.
    e. I had difficulty in making judgments for almost all of the evidence samples (14 – 15).

2. How would you rate the overall appropriateness (fit as a measure of the indicator) of the assessment procedures used to collect data for each piece of evidence and each indicator?
    a. All assessment procedures were very appropriate.
    b. The vast majority were appropriate with only 1 or 2 procedures being questionable across the 3 evidence pieces for the 5 indicators.
    c. The majority of procedures were appropriate, but there were several (3 – 7) that were questionable across the 3 evidence pieces for the 5 indicators.
    d. More than half of the procedures (8 – 12) were questionable with only a few being sufficiently appropriate.
    e. Almost all of the procedures (14 – 15) were questionable.

3. How would you rate the overall compliance with Alternate Assessment requirements as evidenced by the information in the student's folio and its presentation?
    a. Perfectly compliant.
    b. Highly compliant with few irregularities.
    c. Generally compliant, but with some irregularities.
    d. Much of the information had irregularities.
    e. Almost all of the information had irregularities.

Science Judgments

In science, 42 student data folios were reviewed by an external expert judge. The expert judges responses to the questions stated above are provided in the tables below. To summarize the responses to the science data folios, 66.6 percent of the folios were judged to be sufficiently clear that there were no or very limited problems in allowing them to make their ratings with confidence. When judging the appropriateness of the assessment procedures, 78.5 percent of the folios were judged to have assessment procedures where all or a vast majority were appropriate to the indicator skill being measured. When addressing overall compliance with the Alternate Assessment procedures, 73.8 percent of the folios were judged to be perfectly or highly compliant with an additional 14.3 percent being judged as "generally compliant, but with some irregularities."

# of ratings	**How would you rate the general overall clarity of the evidence samples in the Evidence Folio in allowing you to make your ratings with confidence?**
20 (47.6%)	I had no problems in making judgments for any of the evidence samples.
8 (19.0%)	I had difficulty in making judgments for a limited number of the evidence samples (1 – 2), but the rest were sufficiently clear.
8 (19.0%)	I had difficulty in making judgments for several of the evidence samples (3 – 7), but over half were sufficiently clear.
3 (7.1%)	I had difficulty in making judgments for more than half of the evidence samples (8 – 13) with only a few being sufficiently clear.
3 (7.1%)	I had difficulty in making judgments for almost all of the evidence samples (14 – 15).
42	

# of ratings	**How would you rate the overall appropriateness (fit as a measure of the indicator) of the assessment procedures used to collect data for each piece of evidence and each indicator?**
19 (45.2%)	All assessment procedures were very appropriate.
14 (33.3%)	The vast majority were appropriate with only 1 or 2 procedures being questionable across the 3 evidence pieces for the 5 indicators.
6 (14.3%)	The majority of procedures were appropriate, but there were several (3 – 7) that were questionable across the 3 evidence pieces for the 5 indicators.
1 (2.4%)	More than half of the procedures (8 – 12) were questionable with only a few being sufficiently appropriate.
2 (4.8%)	Almost all of the procedures (14 – 15) were questionable.
42	

# of ratings	How would you rate the overall compliance with Alternate Assessment requirements as evidenced by the information in the student's folio and its presentation?
22 (52.4%)	Perfectly compliant.
9 (21.4%)	Highly compliant with few irregularities.
6 (14.3%)	Generally compliant, but with some irregularities.
1 (2.4%)	Much of the information had irregularities.
4 (9.5%)	Almost all of the information had irregularities.
42	

Changes in KAA for 2008

As part of the review process, the expert judges were asked for every student's folio to "Identify any deficiencies or problems you observed or had difficulty with during your review that you think should be brought to the attention of KSDE as they attempt to improve the Alternate Assessment process." The comments were summarized and reported to KSDE staff. KSDE staff then used these comments to target the further review of the student evidence files to gain feedback on areas where improvement in training and in the Manuals could be made to improve the reliability and validity of the KAA implementation. The training and associated manuals for KAA have been changed for the 2008 implementation to provide more direction and standardization of the assessment process.