Kansas Assessments
in Reading and Mathematics

2006

TECHNICAL MANUAL

for the

Kansas General Assessments

Kansas Assessments of Multiple Measures (KAMM)

Kansas Alternate Assessments (KAA)

Prepared by:

Patrick M. Irwin, Andrew J. Poggio, Xiangdong Yang,
Douglas R. Glasnapp, and John P. Poggio,

Center for Educational Testing and Evaluation
The University of Kansas

February 2007

Table of Contents

The Kansas Assessments in Reading and Mathematics

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 2006 Kansas Assessments for Reading and Mathematics for grades 3-8 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.

Information is provided addressing the technical quality of assessments developed to measure reading and mathematics learning outcomes specific to three distinct populations of Kansas students. Included is information on the Kansas General Assessments, the Kansas Assessments of Multiple Measures (KAMM) and the Kansas Alternate Assessments (KAA). The Kansas General Assessments are intended for administration to students in general or regular education classes whose educational programs are not regulated by IEP's. The KAMM and KAA are intended for administration to students with moderate and severe disabilities. The main body of this report addresses technical aspects focusing on scores from the Kansas General Assessments. Information on the KAMM tests and the KAA is found in Appendix A and Appendix B, respectively.

The remainder of this report is organized by first presenting an overview of the 2006 Kansas Assessment Program to provide a context for reviewing information. Next, information on the test development procedures aimed at maximizing the validity of the assessments as measures of the targeted indicators in the state's Curricular Standards are presented. Then, results from various psychometric analyses are presented in the sequence in which they were conducted for decision-making. The first psychometric results provide information from the differential item functioning (DIF) analyses. These analyses were conducted initially to identify any items that potentially needed to be dropped from the scoring of a test form due to the differential functioning of an item across gender or ethnic groups. Next, the test form equating analyses for Reading and Mathematics general assessment test forms are presented. The equating results are followed by a discussion of the standard setting analyses and procedures implemented to determine score ranges for classifying students into one of five performance levels defined by the state. Information on score and performance classification reliability follows the section on standard setting. The final section presents evidence from a variety of validity studies providing information on both internal and external sources of score validity. Information addressing the technical aspects of the KAMM and KAA are provided in Appendix A and Appendix B of the report.

Section 1

INTRODUCTION AND ORIENTATION

This technical manual provides information on the psychometric properties of the 2006 Kansas Assessments in Reading and Mathematics. The purposes of these assessments are to:

(1) provide aggregate state accountability and yearly progress information toward meeting the Kansas Curriculum Standards in the tested areas as required by the *No Child Left Behind* federal mandate;
(2) provide building and district information to support school improvement evaluation needs as appropriate; and
(3) report on the performance of students to support instructional planning for individuals and groups as judged appropriate by local educators.

As background, new Kansas assessments in Reading and Mathematics were planned and developed, then administered for the first time in spring 2006. WestEd served as the contractor for the development of test items based on test specification provided by the Kansas State Department of Education (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 in grades 3-8 (Reading and Mathematics), 10 (Mathematics) and 11 (Reading) participated in the assessments. Students to have been tested included regular education students, gifted students, students with disabilities, and English language learners (ELL). Some students at the designated grade levels were exempted from participating in the state assessment programs based on guidelines set out by KSDE. Exclusion of students from an assessment is considered the exception, and the rules governing exclusion are not permissive. The presumption is that all students were to be tested unless specifically and justifiably excluded.

The spring 2006 administration of the Kansas assessments 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 the content areas. 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 2006 administration incorporated important changes from prior Kansas assessments administered in the 2000 – 2005 testing cycle. Curriculum standards and targets for the assessments were changed, test specifications revised, and assessed grade levels expanded to include all students in grades 3-8 and one grade level in high school. In effect, no comparison to past student, building, district or state performance should be made.

To achieve a long term assessment and accountability system projected to be in place for a minimum of five academic years, between three to five different parallel forms of both the Reading and Mathematics general assessment tests were created and administered at each grade level. The tests were distributed and administered such that score equating across forms could occur using an equivalent random groups design. In subsequent years according to a specified plan, different intact forms will be cycled through the assessment to afford comparisons of performance over time at the school district and state level. To assure comparability of scores across the different forms of the tests in Reading and Mathematics, the score scale values on which trend information will be reported in subsequent years have been statistically "equated" across test forms during the baseline year (2006). Thus, while the "percent correct" metric has been chosen as the scale for reporting, the percent correct score values have been "adjusted" to achieve comparability in the interpretation of performance levels across different forms of the tests at each grade. Equating provides for necessary and appropriate adjustments among test forms at a grade for their different difficulties and score variability. Information on equating is provided in a later section of this technical report (see Section 4).

For both Reading and Mathematics general assessment test forms, the item format was multiple-choice with one correct answer to be selected from four response options provided to a question. In each Reading general assessment test form, reading selections representing different text types (narrative, expository, technical, or persuasive selections) were included based on those text types identified as appropriate in the grade level test specifications.

In addition to the general assessment test forms, separate assessments were developed by the state and made available in lieu of the general assessments for administration to students with moderate and severe disabilities. Single forms measuring targeted learning outcomes in each of the reading and mathematics content areas identified as appropriate for students with moderate disabilities were developed for each grade level, 3 – 8 and high school. This assessment is referred to as the Kansas Assessment of Multiple Measures (KAMM). The item format used in these assessments was multiple-choice with three response options. For students with severe disabilities, the Kansas Alternate Assessment (KAA) system was developed. The latter assessment is individualized for a student. A student's score is based on the independent ratings of information available in evidence files targeting learning outcomes from the state's Curriculum Standards selected by local IEP teams as a portion of the student's instructional goals for the year. Information on the technical adequacy of these assessments is found in Appendix A and Appendix B.

The Kansas assessments are planned and created to point 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 have been called upon to provide information to contribute to ongoing school accreditation status, and results from the reading and mathematics 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, cutscores 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 5 of this report provides additional details on the procedures put in place to set the specific test score criteria used to classify students into one of the performance categories established by the state.

As a final important aspect of the Kansas 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 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. Approximately 60 percent of the eligible students across the grades tested in 2006 took the reading and mathematics tests online using KCA. Studies addressing issues of mode comparability have been on-going and continue as part of the program. Results of initial studies may be found in Poggio, Glasnapp, Yang, & Poggio (2005) and Poggio, Glasnapp, Yang, Beauchamp, & Dunham (2005). These studies are not included as part of this Technical Manual.

Section 2

TEST DEVELOPMENT AND CONTENT REPRESENTATION

The content of the Kansas General Assessments is derived from the Kansas Curricular Standards (see http://www.ksde.org/Default.aspx?tabid=1678 for the Curricular Standards in all subject areas). 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 2006 Kansas tests measured targeted indicators in the Curricular Standards for Reading and Mathematics in grades 3-8 and high school (grade 10 for Mathematics and grade 11 for Reading).

Test Specifications

Test Specifications provide the blueprint to be followed in 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 at http://www.ksde.org/Default.aspx?tabid=420.

Item Type

The multiple choice item type is utilized exclusively on the Kansas General Assessments in Reading and Mathematics. For all multiple choice items appearing on any general assessment test form, students select the one best answer from among four choices provided.

Item Development

Beginning with the 2006 assessment cycle, KSDE contracted with WestEd, a third party, to supply Reading and Mathematics 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, the Kansas Curricular Standards, and 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.

Item Delivery and Tryouts

All Reading and Mathematics 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. From CETE reviews, where gaps or shortages in the item pool were identified based on the table of specifications, specific requests were made for additional items at the indicator level so that multiple operational test forms at a grade level in a content area could ultimately be constructed.

All Kansas schools were invited to participate in pilot testing. Due to the large number of items to be tried out across the two content areas of Reading and Mathematics, piloting was performed in three waves, commencing in fall 2004 and continuing through fall 2005. All pilot tests were administered via the KCA delivery mode. Students participating in the fall field tryouts were administered items one grade level below the grade of the student. All items in the item pools supplied to CETE were piloted. As pilot items were administered via computer (KCA), random selection and inclusion on a pilot test form was possible for each student taking a pilot test set of items. Thus, pilot test items (passage sets in reading) were randomly distributed to test takers ensuring that each test item was administered to a random group of students representative of the student population subgroups in Kansas. The number of students responding to an item ranged from a minimum of 250 students to a maximum for a few items of 1500 students.

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. Traditional, or classical item analysis was the method used to evaluate the item pilot data.

For multiple choice items, the classical test theory item difficulties and item discriminations were obtained by computing item means and item-test correlation coefficients. Item means and item-test correlations were represented by p-values and point-biserials, respectively. The p-value indicates the proportion of examinees responding correctly to an item, ranging from 0 to 1.0. The point-biserial gives a measure of the relationship between performance on an item and performance on the test as a whole and can range from -1.0 to 1.0. 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.

In this classical item analysis scheme, item means and item-test correlations are dependent on the sample of examinees that took the various pilot tests. If the group of examinees responding to an item has been well prepared in the concepts assessed, item means will be fairly high and the items will appear to be easy. If the examinees responding to an item do not possess the content knowledge or skills required by it, item means will be fairly low and items will appear to be difficult. If performance on an item does not relate well to performance on the test as a whole, item-test correlations will be low or possibly negative. Care was taken by CETE in randomizing pilot test items into sets and ensuring that an adequate number of item responses were collected on each item.

Across all grade levels assessed and over the two content areas of Reading and Mathematics, several thousand 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:

- $r_{pb} < .20$ for the keyed (correct) response
- $p > .95$ or $p < .25$ for the keyed response
- $r_{pb} > 0$ for any distractor (incorrect answer choice)
- $p > .25$ for a distractor for the high ability group OR $p > .15$ and $r_{pb} > .055$ for the low ability group
- the Kelly discrimination index for an item is less than .20

Each item that was flagged based on the criteria listed above was individually reviewed by CETE and KSDE. During these reviews, items were either accepted or rejected for the final pool of items. For items aligning to an indicator that had sufficient coverage in order to construct multiple test forms, the decision to accept or reject was the only one made for the particular item. Flagged items that aligned to indicators where coverage was an issue for the creation of multiple forms were examined more closely. Items found to be easily correctable or were judged to be conducive to a minor edit or modification with little or no effect on the original intent of the item (that is, no effect on indicator alignment or little effect on the item's characteristics) were retained on a case by case basis. Any poorly functioning item retained was done so based on a judgment that the item was an appropriate (valid) measure of important grade level content, but that students were performing poorly on the item due to lack of instructional opportunity to learn the content.

Test Form Development

As a basis for an equating design, sufficient groundwork on test development was needed to ensure that test forms were constructed to be classically parallel. In order for all test forms to mirror test specifications based on the Kansas indicators, the same number of items were initially selected for inclusion on each form. Content area forms within a grade level were constructed to be parallel and have the same number of items per indicator and as a total. In Reading, sufficient passages and corresponding items were available to construct 3-4 forms at a grade level whereas in Mathematics, a sufficient number of items were available to build five operational forms at each grade level

As a result of pilot testing, sufficient items were available to configure five parallel forms in Mathematics and three or four forms in Reading at each of the following grade levels; 3, 4, 5, 6, 7, 8, 10 (Mathematics only), and 11 (Reading only). For the Mathematics forms, items surviving the review of the pilot data were compiled at each grade level and grouped by measured student learning outcome classification (standard, benchmark, indicator, sub-indicator, cognitive complexity). Items were ordered on the basis of item difficulty from low to high (value) and placed on one of five forms. In some cases, more items existed in the pool for a given indicator than called for by the test specifications, so not all items were used during form construction due to the randomization process. After all forms were initially constructed in this manner at a grade level, content and statistical reviews of each form were conducted. All items corresponding to an indicator across forms at a grade level were examined to ensure adequate content coverage. In places where there was overlap on a form or content gaps, items were deliberately moved across forms in an attempt to ensure content representation and reduce content overlap within a form. Statistical reviews were then executed, whereby average difficulty values were calculated at the test and indicator level across forms. Items were moved across forms to ensure statistical similarity in terms of difficulty at the indicator and overall form level with consideration given to content representation. In the end, all five forms were pre-equated based on pilot data to have average difficulty values within 2 hundredths (.02) of one another at the test level and within 6 hundredths (.06) at the indicator level.

A review of the piloted item pool for Reading was conducted in a manner similar to the Mathematics item reviews. Pilot data were examined and poorly functioning items were flagged and reviewed. Surviving passages and items were compiled at each grade level, grouped by measured student learning outcome classification (standard, benchmark, indicator, sub-indicator). Passages were identified by type or genre (Narrative, Technical, Persuasive, Expository) based on the test specifications provided by KSDE. Additionally, the passages were examined based on their readability, word count, and indicator coverage. Passages and items were cast onto forms balancing difficulty at the text type and item indicator level. After all forms were constructed at a grade level, content and statistical reviews of each form were conducted. All items corresponding to an indicator across forms were examined to ensure content coverage. Statistical reviews were then performed, whereby average difficulty values were calculated at the test, passage, and indicator level across forms. Passages and items were moved across forms to ensure statistical similarity in terms of difficulty at the indicator and overall form level with consideration given to content representation. The lengths of passages and test questions were also controlled to minimize the structural differences between forms. In Reading, sufficient passages and corresponding items were available to construct 3-4 forms at a grade level. Pre-equating based upon pilot data allowed passage and items to be distributed across forms such that mean item difficulties were within .05 across forms at the passage type level.

For the spring 2006 administration, all operational test forms were administered on KCA using random assignment of test form for purposes of equating test scores across forms (see Section 4). Due to delays in the development process, only one form of a Reading test and one form of a Mathematics test was readied at each grade to be printed in time for distribution into the field. Thus, only one form of any grade level test was made available to be administered in the traditional paper and pencil modality.

Following the administration of the first operational forms of the Kansas Assessments in Reading and Mathematics in spring 2006, analysis work commenced employing classical and IRT methods. Traditional item analysis studies were conducted on each test form to reconfirm the pilot test results that items selected for operational use were functioning adequately and as expected. As sufficient numbers of students in impacted subgroups do not exist in Kansas for examining differential item functioning (DIF) during the pilot testing phase of item development and selection, DIF analyses were performed on all items across all content area forms using spring administration test data. A Bias/Equity Review Committee was formed to review all items flagged as showing DIF (see Section 3 of this report). Test form equating was performed (see Section 4) following the DIF studies. Before AYP reporting could occur, standard setting activities needed to be implemented to establish score ranges on the tests that would define levels of test score performance needed for students to be classified into one of the five performance level categories established by the state (Exemplary, Exceeds Standard, Meets Standard, Approaches Standard, and Academic Warning). See Section 5 of this report for descriptions of the standard setting activities implemented. Based on the standards recommended by KSDE and approved by the Kansas State Board of Education, final results for the Kansas Assessments in Reading and Mathematics were reported in September 2006.

Section 3

DIFFERENTIAL ITEM FUNCTIONING (DIF) ANALYSES

Differential item functioning (DIF) is an important step in test construction. It refers to an empirical analysis of item responses to identify items on which examinees from different gender or ethnic groups have different probabilities or likelihoods of success, after they have been matched on the ability (or test total scores) of interest. DIF provides a necessary, but not sufficient, condition for item bias. Commonly, logical judgmental reviews of DIF items by panels representing impacted gender and ethnic groups need to be conducted before any judgments can be made about whether an item shows any bias, insensitivity, or offensiveness toward any gender or ethnic group.

There is currently no single industry standard for conducting studies of DIF in terms of either methodology or criteria used for making decisions. The literature contains several proposed procedures, each different in the way it statistically handles item data and the indices it produces as criteria for identifying DIF items. A strategy of DIF analysis adopted this year was to select one statistical procedure of DIF analysis as the primary method and to augment it by a different procedure on some subsets of DIF comparison groups where the appropriateness of the primary method might become a concern. A comparative study of the results from both procedures was then conducted and a decision made based on the consistency across procedures.

Several implementation issues essential for appropriate DIF analysis were considered for the assessment this year. As with any statistical procedure, sample sizes of the comparison groups have direct impact on the power of the DIF procedure. With very small samples of reference or focal groups, results from the DIF analysis might not be trustworthy. Based on the sample size recommendations in the literature, sample sizes for both reference and focal groups were examined before the DIF analysis.

Procedures for identifying DIF may be over-sensitive to different curriculum/instructional approaches that could influence performance given the content of an item. This effect is particularly important in Kansas where ethnic groups involved in the DIF analyses are largely congregated in a few districts, but where results would typically be compared to a random sample of White test takers across the entire state. The sampling plan was developed to handle this issue appropriately.

Procedures

Samples

Taking the above into account, the DIF analysis procedures and criteria put in place emphasized sufficient sample sizes and curriculum matching as a basis for making decisions and recommendations. In 2006, analyses were conducted for each test form using gender and racial/ethnic groups. To control for the effects of different curriculum/instructional approaches, samples of White test takers were drawn from schools that had minority groups. Separate samples of Whites were drawn for each minority group.

In 2006, there were on average four or five test forms per grade, which led to smaller sample sizes for minority groups than previous years. The sample size issue becomes particularly relevant for Asian American and Native Americans (which mostly number less than 100 in Mathematics and 120 in Reading). Such sample sizes are consistently less than 200, as suggested by the literature as the minimal sufficient sample size for conducting DIF studies. Therefore, for racial/ethnic group comparison, DIF studies were conducted only on African Americans and Hispanic Americans, using sampled White as the reference group.

Items

The reading items and math items from general assessment test forms at all grade levels were analyzed for DIF. In Mathematics, there were five test forms per grade with an equal number of items across test forms. For each of the seven grades tested (Grades 3, 4, 5, 6, 7, 8, and 10), the number of items on each test form at a given grade was 70, 73, 73, 86, 84, 86, and 84, respectively. Thus, the total number of mathematics items involved in the DIF analyses ranged from 350 at grade 3 to 430 at grade 8. In Reading, there were four test forms for Grades 3, 4, 6, 7, and 11, respectively, and three test forms for Grades 5 and 8, respectively. Again, all test forms at a given grade had an equal number of items. For each grade (Grades 3, 4, 5, 6, 7, 8, and 11), the number of items on each test form at a grade were 58, 74, 74, 80, 84, 83, and 81, respectively. The number of items involved in the DIF analyses in Reading ranged from 232 (grade 3) to 336 at grade 7. All items were in the multiple choice format and thus were scored dichotomously.

Statistical Methods

The main procedure used was the Mantel-Haenszel (MH) technique. For DIF analyses on minority groups, SIBTEST procedure was also conducted. Results from both procedures were very consistent and only the results from the MH technique are reported here. The criteria used in these analyses were (1) the absolute ETS delta value larger than 1.5 and (2) the absolute ETS delta value statistically significantly larger than 1.0. Using a significant level of .01, the second criterion is equivalent to a MH chi-squared value of 12.7866. Items with negative delta values created a disadvantage for the focal group while positive values created an advantage for the focal group in comparison to the reference group.

Results

A sample output for a Mathematics DIF analysis at Grade 3, Form 405 is given in Table 3.1 below. Twenty items were deleted arbitrarily for space conservation. From Table 3.1, item 2 appeared to show DIF for the Female versus Male comparison. This item has an absolute Delta value of 1.66 and MH chi-squared value of 18.66. This item seemed to advantage Females.

Tables 3.2 through 3.8 give summaries of items flagged by the Mantel-Haenszel procedure for each DIF comparison by form at each of the seven grade levels for Mathematics, respectively. In each of the tables, information about the flagged DIF items for each specific comparison performed on each form at a given grade is grouped into four parts. The test form number is given in the first column of each table (under the title *Form*). In the second part (under the title *DIF Group*), both the reference and focal group in each of the three comparisons performed on a test form, as well as their corresponding sample sizes, are given. It should be noted that different samples of White were drawn for Hispanic/White and Black/White comparison, respectively. In the third part (under the title *DIF Items*), ID numbers for items that are showing DIF are presented in the table. Specifically, the ID number for each item is a unique number in the CETE test system that makes it possible to track all changes and decisions made for the item. For each item ID number, a "+" or "-" sign indicates the direction of the DIF that the item shows. As mentioned earlier, items with "-" were seen to disadvantage the focal group while items with "+" advantaged this group in comparison to the reference group. The last part of each table gives the total counts of the number of flagged items for each test form (*Total*). Items that advantage or disadvantage the focal group were tallied separately.

As an example, Table 3.2 presents the results of DIF analyses for the five mathematics test forms at Grade 3 (i.e., Forms 405, 664, 665, 666 and 669). For each form, DIF analyses were conducted for each of the three comparisons (i.e., Male vs. Female, White vs. Hispanic, and White vs. Black). Table 3.2 shows that 14 items were flagged at Grade 3 from a total of 1050 comparisons (350 items and 3 group comparisons), most of them (10 items) with negative DIF. Five out of the 14 items were flagged for gender comparison, and the remaining items were flagged for ethnic comparisons. For the mathematics DIF analyses across all grade levels, 8,295 comparisons were made and a total of 39 items were flagged showing positive DIF and 45 items were flagged showing negative DIF. The number of flagged items represent one percent (1%) of the total number of statistical comparisons made.

For DIF analyses in Reading, Tables 3.9 through 3.15 give the corresponding summaries for each of the seven grade levels. Table 3.9 shows that six items were flagged at Grade 3 reading, with four items showing negative DIF. All but one item were flagged for ethnic comparisons. Across grades, 20 items were flagged showing positive DIF and 45 items were flagged showing negative DIF from a total of 5937 comparisons made. The number of flagged items represent slightly over one percent (1.09%) of the total number of statistical comparisons made.

Judgmental Review of DIF Items

As tests should be free bias, examinees of equal standing with respect to the construct of the test should, on average, earn the same test score irrespective of group membership (AERA/ APA/NCME, 1999). At various points during the test development, administration, and review process for the Kansas assessments, various efforts were made to eliminate potential bias against groups of examinees on the basis of irrelevant factors or characteristics. These efforts focused on a combination of professional judgments about the appropriateness and freedom from bias of program materials and the gathering and interpretation of statistical information about differential item functioning. It has been suggested that the construct of bias is multidimensional, (Berk, 1982) and that judgmental reviews and statistical methods of bias detection should complement each other. According to this view, each method may contribute its own separate strengths to the analysis of potential bias. Statistical analysis is strongest in detecting test items that produce larger than expected group differences in performance but are also susceptible to random errors expected to occur in the comparison process. In contrast, professional reviewers may focus on aspects of the bias construct (e.g., stereotyping) that it are highly desirable to eliminate from test materials but that might have either no negative effect on examinee performance or no locally detectable effect but only a more subtle, cumulative effect over an entire test or set of tests (Tittle, 1982). There is consensus in the field of educational measurement that this combination of professional judgment and statistical analysis is a necessary practice within any testing program. These two applications for identifying potential bias in a test are best conceptualized not as separate activities, but rather as important complementary components.

Equity Review Committee

An equity review committee was established by the Kansas State Department of Education to review potentially biased items on the Kansas Assessments. The committee, composed of minority educators from across the state, was formed to judgmentally review test items for sensitivity and fairness that were flagged as showing differential item functioning (DIF) during the statistical DIF analyses. The Equity Review Committee consisted of 19 Kansas individuals (4 male, 15 female). The table below details the committee members' representation of specific impacted groups.

Impacted group	Male	Female
Hispanic	1	4
Female	0	6
African American	1	5
Asian	2	0
Total	**4**	**15**

Sub-committees were formed on the basis of these gender or ethnic groupings. The Equity Review Committee convened on June 5, 2006 in Topeka, Kansas to review items that evidenced DIF in the statistical analysis of the items for students belonging to the respective ethnic or gender groupings. Each sub-committee member was provided sets of these items from the Reading and Mathematics content area tests; sub-committees reviewed items only evidencing DIF for students in their same ethnic or gender grouping.

An overview of the bias review process was presented by CETE and KSDE staff to start the proceedings. After the training session, committee members divided into groups and began the judgmental procedure. Panelists were directed to review each item flagged for DIF in the statistical analyses for students in their particular gender or ethnic group independently in terms of fairness, focusing specifically on content, language, offensiveness, or stereotypes that may have been present in the respective items. After the independent item review was completed within each sub-committee, panelists remained in their groups and engaged in a discussion regarding each item under review. The review criteria presented to the committee during the training session required the sub-committees to reach consensus regarding each item. For items that the review committees detected bias present in the item, a description or explanation of the source of the bias was required. KSDE was supplied with the item feedback from each sub-committee and made the final decision regarding an item's deletion or retention. A total of five flagged items from the Kansas Reading and Mathematics Assessments were judged to contain content or language that was biased against members of certain gender or ethnic groups. These items were deleted from the scoring of the form on which they occurred and will not appear on future test forms. Of those items, three appeared on various Reading forms at assorted grade levels and two appeared on Mathematics forms.

Table 3.1. Example DIF output for Mathematics

MANTEL-HAENSZEL DIP ANALYSIS FOR DIF, G3 Math,form405,Male(1,reference) vs. Female(0, focal)

NUMBER OF ITEMS = 70 & CHK = .000

ITEM	P-1	PB-1	P-0	PB-0	P-1+0	PB-1+0	CHI-I	APLHA-I	DELTA-I	CHI-E	APLHA-E	DELTA-E
1	.95	.18	.95	.13	.95	.16	.45	.89	.27	.68	.87	.32
2	.94	.25	.96	.28	.95	.26	18.43	.49	1.66	16.61	.52	1.56
3	.91	.28	.93	.24	.92	.26	7.70	.70	.84	6.75	.71	.79
4	.83	.43	.86	.41	.85	.42	10.08	.72	.76	11.10	.72	.78
5	.93	.25	.93	.33	.93	.29	1.04	.86	.35	.96	.87	.33
6	.87	.39	.88	.36	.87	.37	5.34	.78	.59	3.96	.81	.51
7	.90	.41	.88	.39	.89	.40	1.75	1.17	-.37	2.49	1.20	-.44
8	.87	.46	.84	.44	.85	.45	4.64	1.26	-.54	5.32	1.27	-.56
9	.83	.42	.84	.40	.84	.41	1.91	.87	.33	1.89	.87	.32
10	.85	.38	.85	.40	.85	.39	.06	.97	.07	.06	.97	.07
11	.86	.40	.86	.37	.86	.39	.00	1.01	-.01	.01	1.02	-.04
12	.75	.33	.71	.32	.73	.33	4.30	1.18	-.38	4.28	1.17	-.38
13	.88	.48	.88	.44	.88	.46	.96	.89	.28	1.18	.88	.31
14	.68	.17	.67	.22	.67	.19	.12	1.03	-.06	.17	1.03	-.07
15	.77	.35	.76	.40	.77	.38	.24	1.05	-.10	.16	1.04	-.09
16	.84	.35	.81	.32	.82	.33	3.53	1.19	-.41	3.11	1.17	-.38
17	.90	.43	.87	.47	.88	.45	7.16	1.36	-.72	7.92	1.38	-.76
18	.70	.36	.70	.39	.70	.37	.12	.97	.07	.10	.97	.06
19	.92	.35	.93	.41	.92	.38	2.85	.79	.54	3.67	.77	.61
20	.90	.42	.90	.44	.90	.43	.48	.91	.21	.76	.90	.26
21	.69	.43	.70	.40	.69	.41	4.26	.85	.38	3.13	.87	.32
22	.95	.38	.96	.39	.95	.38	.75	.84	.40	.67	.85	.37
23	.73	.42	.74	.43	.74	.43	3.91	.85	.39	2.59	.88	.31
24	.96	.31	.95	.31	.95	.31	1.31	1.23	-.49	1.51	1.25	-.52
25	.92	.34	.91	.32	.91	.33	.60	1.11	-.24	1.18	1.15	-.33
26	.82	.46	.79	.47	.80	.47	5.07	1.24	-.50	4.75	1.22	-.48
27	.88	.44	.86	.46	.87	.45	1.48	1.15	-.32	.97	1.12	-.26
28	.80	.41	.81	.40	.81	.40	2.81	.86	.36	3.43	.85	.40
29	.96	.28	.95	.31	.95	.30	.19	.92	.21	.11	.93	.16
30	.83	.51	.78	.50	.81	.50	19.16	1.53	-.99	17.38	1.49	-.93
31	.89	.44	.88	.41	.89	.43	.21	1.06	-.14	.33	1.07	-.17
32	.91	.38	.85	.38	.88	.38	27.98	1.81	-1.40	28.03	1.80	-1.38
33	.89	.40	.88	.40	.89	.40	.30	1.07	-.16	.50	1.09	-.20
34	.84	.34	.87	.37	.86	.35	9.66	.73	.74	9.65	.73	.73
35	.68	.51	.69	.50	.69	.51	2.76	.87	.33	2.25	.88	.29
36	.83	.45	.81	.44	.82	.44	1.22	1.11	-.25	1.63	1.13	-.29
37	.91	.37	.89	.41	.90	.39	.85	1.12	-.27	.95	1.13	-.29
38	.62	.40	.65	.37	.64	.38	7.18	.82	.47	6.26	.83	.43
39	.80	.44	.81	.42	.80	.43	1.83	.88	.29	1.00	.91	.22
40	.91	.44	.91	.45	.91	.45	.74	.89	.28	.57	.90	.25
41	.52	.24	.46	.18	.49	.21	11.88	1.26	-.55	12.84	1.27	-.56
42	.53	.39	.55	.37	.54	.38	4.56	.86	.36	3.92	.87	.33
43	.87	.30	.85	.33	.86	.32	.45	1.07	-.17	.46	1.07	-.17
44	.86	.39	.83	.36	.84	.37	5.20	1.25	-.52	4.98	1.24	-.51
45	.80	.35	.81	.28	.81	.31	1.40	.90	.25	1.43	.90	.25
46	.76	.49	.72	.53	.74	.51	3.76	1.19	-.40	3.74	1.18	-.39
47	.95	.23	.96	.25	.96	.24	.74	.86	.36	.82	.85	.38
48	.94	.24	.92	.29	.93	.27	7.16	1.44	-.86	6.75	1.42	-.83
49	.84	.34	.83	.37	.83	.35	1.27	1.12	-.26	1.44	1.12	-.27
50	.94	.17	.94	.22	.94	.19	1.40	.84	.41	1.15	.85	.37

** SUMMARY DATA **

	MEAN	SD	CASES	KR-20
GROUP-1	58.281	9.527	1951.	.913
GROUP-0	57.760	9.831	1932.	.916
TOTAL	58.022	9.683	3883.	.914

Table 3.2. Summary of Differentially Functioning Items for Grade 3 Mathematics Forms

Form	DIF Group				DIF Items		Total
	Reference	N	Focal	N			
405	Male	1951	Female	1932	8922+		
	White	1066	Hispanic	485			
		772	Black	199	8529-	8288-	
							1 +, 2 -
664	Male	1988	Female	1908	8771-	8398+	
	White	1056	Hispanic	485			
		720	Black	211	4535-	4478-	
							1 +, 3 -
665	Male	1994	Female	1884			
	White	1008	Hispanic	482	8866-	4661+	
		761	Black	208			
							1 +, 1 -
666	Male	2016	Female	1883			
	White	1071	Hispanic	494	7265-	9407-	
		691	Black	208	9407-		
							0 +, 2 -
669	Male	1947	Female	1935	8093+	4537-	
	White	886	Hispanic	484	4533-		
		660	Black	219	4533-		
							1 +, 2 -
Total							4 +, 10 -

Table 3.3. Summary of Differentially Functioning Items for Grade 4 Mathematics Forms

Form	DIF Group				DIF Items			Total
	Reference	N	Focal	N				
595	Male	2246	Female	2191				
	White	1149	Hispanic	538				
		979	Black	249				
								0 +, 0 -
670	Male	2242	Female	2224	8170-	4619-	7401-	
	White	1149	Hispanic	528	8840-			
		951	Black	221	8840-	7159+	7401-	
								1 +, 4 -
671	Male	2251	Female	2168				
	White	1175	Hispanic	516				
		878	Black	236				
								0 +, 0 -
672	Male	2270	Female	2175	9440-			
	White	1316	Hispanic	524	4617-	7078-		
		1040	Black	234				
								0 +, 3 -
673	Male	2276	Female	2179	9370-			
	White	1251	Hispanic	539				
		974	Black	235				
								0 +, 1 -
Total								1 +, 8 -

Table 3.4. Summary of Differentially Functioning Items for Grade 5 Mathematics Forms

Form	DIF Group				DIF Items			Total
	Reference	N	Focal	N				
406	Male	2221	Female	2220				
	White	1147	Hispanic	489				
		936	Black	215				
								0 +, 0 -
674	Male	2278	Female	2156				
	White	1067	Hispanic	508	9533-			
		819	Black	227				
								0 +, 1 -
675	Male	2257	Female	2147				
	White	1118	Hispanic	467				
		902	Black	222				
								0 +, 0 -
676	Male	2195	Female	2168	4423+			
	White	1067	Hispanic	485	8069+	14431+	9528-	
		877	Black	243	7124+	6487+		
								5 +, 1 -
678	Male	2265	Female	2161				
	White	1076	Hispanic	482	11433-	11260-		
		851	Black	219	9506-			
								0 +, 3 -
Total								5 +, 5 -

Table 3.5. Summary of Differentially Functioning Items for Grade 6 Mathematics Forms

Form	DIF Group				DIF Items		Total
	Reference	N	Focal	N			
479	Male	2410	Female	2202			
	White	1761	Hispanic	517			
		1433	Black	298	6695+		
							1 +, 0 -
680	Male	2349	Female	2253			
	White	1553	Hispanic	477	6666-		
		1401	Black	269			
							0 +, 1 -
681	Male	2263	Female	2194	9117+	4441+	
	White	1598	Hispanic	453	4463-	6646+	
		1416	Black	301	4066+		
							4 +, 1 -
686	Male	2376	Female	2194	8893+	6510-	
	White	1662	Hispanic	474			
		1309	Black	304			
							1 +, 1 -
687	Male	2441	Female	2149	8962+	6453-	
	White	1455	Hispanic	464			
		1295	Black	295	6661+		
							2 +, 1 -
Total							8 +, 4 -

Table 3.6. Summary of Differentially Functioning Items for Grade 7 Mathematics Forms

Form	DIF Group				DIF Items						Total
	Reference	N	Focal	N							
596	Male	2486	Female	2333	8702-						
	White	2095	Hispanic	434							
		1715	Black	290	4059+	14432-	6828+	6976+	8388+	6199+	
											5 +, 2 -
682	Male	2546	Female	2276	8894-						
	White	1990	Hispanic	448							
		1647	Black	298	4597+	8894-	11401+	4082+	6826-		
											3 +, 2 -
683	Male	2462	Female	2346	14914-						
	White	1984	Hispanic	426	11317-						
		1652	Black	278	14462+	11413+	6942-				
											2 +, 3 -
684	Male	2443	Female	2362	10699+	8355+					
	White	1995	Hispanic	420							
		1818	Black	303	9056-						
685	Male	2433	Female	2355							
	White	1740	Hispanic	429							
		1688	Black	289	7199+						
											1 +, 0 -
Total											14 +, 8 -

Table 3.7. Summary of Differentially Functioning Items for Grade 8 Mathematics Forms

Form	DIF Group				DIF Items			Total
	Reference	N	Focal	N				
586	Male	2544	Female	2512				
	White	2401	Hispanic	487				
		1809	Black	275	7369+			
								1 +, 0 -
688	Male	2561	Female	2564				
	White	2129	Hispanic	470				
		1833	Black	329	3805+			
								1 +, 0 -
690	Male	2564	Female	2460				
	White	2113	Hispanic	450				
		1612	Black	315	7486+	7033-		
								1 +, 1 -
691	Male	2580	Female	2452				
	White	1978	Hispanic	468				
		1741	Black	289	6325-	6592+	6755-	
								1 +, 2 -
700	Male	2568	Female	2453				
	White	2128	Hispanic	470	9199-			
		1761	Black	286	7485+			
								1 +, 1 -
Total								5 +, 4 -

Table 3.8. Summary of Differentially Functioning Items for Grade 10 Mathematics Forms

Form	DIF Group				DIF Items			Total
	Reference	N	Focal	N				
590	Male	2409	Female	2395	9131-			
	White	2198	Hispanic	353				
		1825	Black	267				
								0 +, 1 -
702	Male	2444	Female	2319				
	White	2010	Hispanic	378				
		1651	Black	232				
								0 +, 0 -
719	Male	2471	Female	2320	8219-			
	White	2258	Hispanic	403				
		1739	Black	253	4048+	14949+	9242-	
								2 +, 2 -
720	Male	2465	Female	2366	4113-			
	White	2117	Hispanic	408				
		1755	Black	255	6894-			
								0 +, 2 -
721	Male	2448	Female	2356				
	White	2318	Hispanic	360				
		1720	Black	251	6355-			
								0 +, 1 -
Total								2 +, 6 -

Table 3.9. Summary of Differentially Functioning Items for Grade 3 Reading Forms

Form	DIF Group				DIF Items		Total
	Reference	N	Focal	N			
386	Male	2255	Female	2201			
	White	1357	Hispanic	527	19600-		
		819	Black	208			
							0 +, 1 -
522	Male	2313	Female	2113			
	White	1363	Hispanic	526	19698-	19705+	
		929	Black	217			
							1 +, 1 -
558	Male	2236	Female	2194	19763+		
	White	1338	Hispanic	535	19799-		
		999	Black	242	19803-		
							1 +, 2 -
559	Male	2182	Female	2229			
	White	1238	Hispanic	504			
		989	Black	241			
							0 +, 0 -
Total							2 +, 4 -

Table 3.10. Summary of Differentially Functioning Items for Grade 4 Reading Forms

Form	DIF Group				DIF Items				Total
	Reference	N	Focal	N					
404	Male	2601	Female	2515					
	White	1551	Hispanic	580	20386-	20403-			
		1366	Black	253					
									0 +, 2 -
561	Male	2647	Female	2499	19996+	20008-	20020+		
	White	1621	Hispanic	543	19999+	20016-	20030-	20057-	
		1161	Black	252					
									3 +, 4 -
562	Male	2574	Female	2542					
	White	1631	Hispanic	570	20150-				
		1218	Black	264					
									0 +, 1 -
563	Male	2564	Female	2529					
	White	1515	Hispanic	569	20229-	10627-			
		1331	Black	269	20213-				
									1 +, 2 -
Total									4 +, 9 -

Table 3.11. Summary of Differentially Functioning Items for Grade 5 Reading Forms

Form	DIF Group				DIF Items				Total
	Reference	N	Focal	N					
388	Male	3642	Female	3459	20445+	20472-	12929-		
	White	2478	Hispanic	742	20491-	12929-			
		2128	Black	365					
									1 +, 3 -
565	Male	3583	Female	3518	20281-	20653+	10601-		
	White	2507	Hispanic	732	20289-	20320+	20322-	20665-	
		2127	Black	355					
									2 +, 5 -
566	Male	3540	Female	3522					
	White	2509	Hispanic	747	20737-	20748-			
		1920	Black	334					
									0 +, 2 -
Total									3 +, 10 -

Table 3.12. Summary of Differentially Functioning Items for Grade 6 Reading Forms

Form	DIF Group				DIF Items		Total
	Reference	N	Focal	N			
401	Male	2928	Female	2733	20577-	20580-	
	White	2122	Hispanic	594	20634-		
		1733	Black	371			
							0+, 3-
569	Male	2907	Female	2694			
	White	2086	Hispanic	582	20826-	20843-	
		1589	Black	334			
							0+, 2-
571	Male	2914	Female	2731			
	White	2273	Hispanic	579			
		2011	Black	383			
							0+, 0-
572	Male	2932	Female	2694	20884-	20949+	
	White	2075	Hispanic	556	20922+		
		1749	Black	362			
							2+, 1-
Total							2+, 6-

Table 3.13. Summary of Differentially Functioning Items for Grade 7 Reading Forms

Form	DIF Group				DIF Items		Total
	Reference	N	Focal	N			
410	Male	2950	Female	2872			
	White	2351	Hispanic	601	22118-	22178-	
		2075	Black	372	22187+		
							1+, 2-
573	Male	2974	Female	2823	21168+		
	White	2405	Hispanic	569	21168+		
		1996	Black	358			
							1+, 0-
574	Male	3052	Female	2825			
	White	2700	Hispanic	605	21235-		
		2210	Black	353			
							0+, 1-
575	Male	2997	Female	2844	21418-	21467+	
	White	2468	Hispanic	578	21378-		
		1947	Black	337			
							1+, 2-
Total							3 +, 5 -

Table 3.14. Summary of Differentially Functioning Items for Grade 8 Reading Forms

Form	DIF Group				DIF Items			Total
	Reference	N	Focal	N				
470	Male	4144	Female	4133	21854-	21885+	21904-	
	White	4250	Hispanic	858	21859-			
		3416	Black	468				
								1+, 3-
577	Male	4271	Female	4063	12950-			
	White	4060	Hispanic	805	22212-	22213-	22225-	
		3490	Black	515				
								0+, 4-
578	Male	4187	Female	4130				
	White	4350	Hispanic	881				
		3477	Black	513				
								0+, 0-
Total								1+, 7-

Table 3.15. Summary of Differentially Functioning Items for Grade 11 Reading Forms

Form	DIF Group				DIF Items				Total
	Reference	N	Focal	N					
480	Male	2883	Female	2794	21704+				
	White	2754	Hispanic	407	21610-				
		2112	Black	311	21634-				
									1 +, 2 -
581	Male	2897	Female	2777	21590+				
	White	2760	Hispanic	457	21590+				
		2235	Black	320					
									1 +, 0 -
582	Male	2867	Female	2716	23960+	22392-			
	White	2705	Hispanic	417	22329+				
		2031	Black	289					
									2 +, 1 -
583	Male	2866	Female	2732	21797-				
	White	2597	Hispanic	406	21769-				
		2039	Black	316	21789+	21800-	21808-	21809-	
									1 +, 5 -
Total									5 +, 8 -

Section 4

TEST EQUATING

An important property of test equating is equity (Kolen and Brennan, 1995; Lord, 1980). Simply stated, equity requires that it should be a matter of indifference to examinees at every ability level whether they respond to form X or form Y, for example, of a test. Two additional properties are symmetry and identical test specifications. Without these three properties or assumptions, a test form cannot be said to be satisfactorily equated even if sophisticated methods were applied.

With newly developed Kansas Assessments in 2006, scores from parallel test forms administered to different groups needed to be equated to ensure the equitability of scores for every examinee. As detailed in Section 2 under forms development, care was taken to configure test forms that were pre-equated based on pilot data to ensure that test forms were constructed to be classically parallel, an important prerequisite as a basis for equating scores across multiple test forms. This section summarizes the description the equating design, and the methods, as well as issues in equating multiple forms of the Kansas Assessments.

Procedures

Equating Design and Data Configuration

The spring 2006 administration of the Kansas Assessment test forms in Reading and Mathematics allowed buildings to select the mode of administration (paper-and-pencil or computer) for individual students. Thus, a building at each tested grade for each subject (Reading or Mathematics) voluntarily elected to test none, part, or all of its students on the computer. Due to the constraints imposed by deadlines for printing paper-and-pencil (P&P) test forms, only one form of each grade level test was made available for administration in the P&P mode. All other forms for a grade level content area were available only on the computer (KCA). Items were configured from pilot test data to make up to five parallel forms available in Mathematics and up to four parallel forms available in Reading dependent on the grade level. All test forms were made available on KCA and were randomly assigned to students when students were registered for KCA. Thus, the basic configuration for test administration in the state without manipulation is as follows.

P&P Selected Group 1	**KCA** Selected Group 2				
Form A	**Form A**	**Form B**	**Form C**	**Form D**	**Form E**
	Random G1	Random G2	Random G3	Random G4	Random G5

The above configuration effectively and efficiently put in place a randomized, equivalent groups design for equating test form scores using only the KCA tested students. The potential problem is the volunteer-nature of the KCA group and that it may not sufficiently reflect the complete distribution of ability and performance of all students in the state. To the extent that the KCA score distribution differs from the complete state distribution, the score equating in one or both score distribution tails may contain greater error.

From the approximately 33,000-35,000 regular education students taking the test at each grade level, approximately one-third took the test using the P&P mode and received the single P&P form available at a grade level. For the remaining two-thirds of the students who were administered the test using KCA, approximately 3,500-5,000 students took each of the Mathematics test forms at each grade level. In Reading, fewer test forms were available at the assessed grade levels resulting in approximately 4,500-8,000 students taking each form. These numbers are more than adequate for the random groups equating method.

Tables 4.1 and 4.2 below show the percentages of students administered the different KCA forms of the Mathematics and Reading assessments across schools in Kansas. For the values in both tables, percentages of students taking each form were obtained for each school, and these percentages were summarized across schools. In addition, the table provides percentages of students taking each form by gender, race, and educational classifications. The bolded numbers in the table refer to the characteristics of students across school on the base form at each grade level. Across content areas and all grade levels, percentages based on demographic information support the equivalence of groups obtained through this data collection design. In other word, data in Table 4.1 and 4.2 suggest the equivalence of the KCA groups responding to each form, at all grades, for each tested content area.

Table 4.1. Number of KCA Kansas Schools and Percentages of Students Taking Different Mathematics Test Forms

Grade	N of Schools	Form		Gender		Race		Education	
			Total	Female	Male	White	Minority	Regular	Sped
3	**534**	**405**	**20.4**	**49.8**	**50.2**	**76.3**	**23.7**	**91.0**	**9.0**
		664	20.9	49.0	51.0	76.1	23.9	90.4	9.6
		665	20.4	48.6	51.4	76.0	24.0	91.1	8.9
		666	21.1	48.3	51.7	76.3	23.7	90.4	9.6
		669	20.4	49.8	50.2	75.1	24.9	90.6	9.4
4	**577**	**595**	**20.3**	**49.4**	**50.6**	**76.6**	**23.4**	**89.3**	**10.7**
		670	20.6	49.8	50.2	76.9	23.1	89.1	10.9
		671	20.6	49.1	50.9	76.6	23.4	89.5	10.5
		672	20.6	48.9	51.1	76.6	23.4	89.2	10.8
		673	20.3	48.9	51.1	76.2	23.8	89.6	10.4
5	**557**	**406**	**20.8**	**50.0**	**50.0**	**78.6**	**21.4**	**87.3**	**12.7**
		674	20.3	48.6	51.4	77.8	22.2	88.0	12.0
		675	20.4	48.8	51.2	78.4	21.6	88.0	12.0
		676	20.6	49.7	50.3	77.6	22.4	88.2	11.8
		678	20.7	48.8	51.2	77.9	22.1	89.1	10.9
6	**454**	**479**	**20.7**	**47.7**	**52.3**	**76.7**	**23.3**	**88.5**	**11.5**
		680	20.4	49.0	51.0	77.7	22.3	88.0	12.0
		681	20.8	48.1	51.9	78.1	21.9	88.6	11.4
		686	20.5	48.0	52.0	77.5	22.5	88.2	11.8
		687	20.9	46.8	53.2	78.1	21.9	89.1	10.9
7	**380**	**596**	**20.5**	**48.4**	**51.6**	**79.9**	**20.1**	**87.5**	**12.5**
		682	20.7	47.2	52.8	79.5	20.5	87.8	12.2
		683	20.3	48.8	51.2	80.2	19.8	88.2	11.8
		684	20.7	49.2	50.8	80.0	20.0	88.5	11.5
		685	20.6	49.2	50.8	79.9	20.1	87.7	12.3
8	**388**	**586**	**20.3**	**49.7**	**50.3**	**80.2**	**19.8**	**88.6**	**11.4**
		688	21.2	50.0	50.0	79.7	20.3	87.6	12.4
		690	20.3	49.0	51.0	80.5	19.5	88.2	11.8
		691	20.4	48.7	51.3	80.1	19.9	88.7	11.3
		700	20.6	48.9	51.1	80.5	19.5	88.9	11.1
10	**323**	**590**	**20.5**	**49.9**	**50.1**	**82.1**	**17.9**	**88.0**	**12.0**
		702	20.6	48.7	51.3	83.1	16.9	89.1	10.9
		719	21.1	48.4	51.6	81.7	18.3	88.7	11.3
		720	21.5	49.0	51.0	82.2	17.8	90.2	9.8
		721	20.3	49.0	51.0	82.7	17.3	88.6	11.4

Table 4.2. Number of KCA Kansas Schools and Percentages of Students Taking Different Reading Test Forms

Grade	N of Schools	Form	Total	Gender		Race		Education	
				Female	Male	White	Minority	Regular	Sped
3	**514**	**386**	**26.0**	**49.4**	**50.6**	**77.9**	**22.1**	**90.6**	**9.4**
		522	25.8	47.7	52.3	77.5	22.5	91.6	8.4
		558	26.3	49.5	50.5	76.6	23.4	90.8	9.2
		559	25.2	50.5	49.5	77.5	22.5	91.3	8.7
4	**548**	**404**	**25.6**	**49.2**	**50.8**	**77.7**	**22.3**	**89.8**	**10.2**
		561	26.0	48.6	51.4	78.7	21.3	89.0	11.0
		562	25.6	49.7	50.3	78.0	22.0	89.5	10.5
		563	25.4	49.7	50.3	77.6	22.4	89.7	10.3
5	**534**	**388**	**33.7**	**48.7**	**51.3**	**79.2**	**20.8**	**87.7**	**12.3**
		565	34.0	49.5	50.5	79.2	20.8	88.8	11.2
		566	33.7	49.9	50.1	78.8	21.2	88.7	11.3
6	**458**	**401**	**26.1**	**48.3**	**51.7**	**77.3**	**22.7**	**88.6**	**11.4**
		569	25.7	48.1	51.9	78.5	21.5	88.1	11.9
		571	25.5	48.4	51.6	78.0	22.0	88.5	11.5
		572	25.4	47.9	52.1	77.7	22.3	88.3	11.7
7	**384**	**410**	**25.6**	**49.3**	**50.7**	**78.5**	**21.5**	**87.7**	**12.3**
		573	24.9	48.7	51.3	79.2	20.8	88.0	12.0
		574	26.5	48.1	51.9	78.9	21.1	87.7	12.3
		575	26.1	48.7	51.3	79.2	20.8	87.7	12.3
8	**383**	**470**	**33.8**	**49.9**	**50.1**	**79.6**	**20.4**	**88.0**	**12.0**
		577	33.8	48.8	51.2	79.5	20.5	88.1	11.9
		578	34.0	49.7	50.3	78.8	21.2	88.2	11.8
11	**324**	**480**	**25.7**	**49.2**	**50.8**	**83.3**	**16.7**	**90.0**	**10.0**
		581	25.9	48.9	51.1	82.0	18.0	89.5	10.5
		582	25.9	48.6	51.4	82.2	17.8	89.9	10.1
		583	25.6	48.8	51.2	82.5	17.5	89.8	10.2

Statistical Procedures

Using random student score samples from the KCA test form, results from both classical and IRT test equating procedures were examined. For classical equating, linear and equipercentile methods were employed. For IRT test equating, IRT observed score equating procedures were used. Each form was separately calibrated using the 3-parameter (3-PL) model and observed score frequency distributions were obtained by summing the compound binomial distribution across all values of theta. Then, the equipercentile method of equating was used on the obtained observed score frequency distributions. In both Mathematics and Reading, the total score levels are expressed in the percent correct metric. The test form given in both the KCA and the P&P mode (Form A) served as the base form in all equating analyses. Scores from all other forms were transformed onto the base form score percent correct scale. Criteria for selecting the best method for equating two specific sets of scores are presented below.

A major issue in 2006 involved equating scores between the P&P test form and the corresponding KCA form (Form A). As the assignment of test taking mode for a student was not random but rather a local decision made by districts or schools, the possibility exists that the assignment of students to KCA or P&P was related to or determined by characteristics of the student. Consequently, the two populations (students who take P&P Form A and students who take the KCA test forms) might be different in terms of proficiency for a given subject at a given grade. Thus, the effects of test mode and population ability differences are intertwined. In a scenario with small mode effects where any difference in P&P and KCA student scores reflects primarily population ability differences, one need not equate. Rather, there would be an assumption of score value equivalency for the same two scores in both populations. This situation has been evidenced to some extent by data from two prior studies in Kansas on mode effects and from other studies and reviews found in the testing literature. Based on prior studies where the mode effect size has been judged small, Kansas made the decision to treat the P&P test form as equivalent to the same KCA administered test form, thus no adjustment was made in the scores for either set of data. Kansas continues to study the comparability issue and the need for conversion tables that might adjust scores obtained under different modes of testing.

Equating Criteria

Because both classical and IRT test equating methods were implemented, comparison between several competing methods was necessary. Thus, equating methods were evaluated and deliberate decisions made as to which method produced the most reasonable conversion of scores for students taking different forms of the test. To assist in selecting the best equating conversion, the following criteria were used.

1. *Fidelity to the equated data*

   An equating conversion that provides the closest approximation to the base form distributional moments given the best score transformation will be used. When there is no difference in form difficulty, the distributional moments of the equated scores will approximate those of the base form.

2. *Minimal impact across score levels for the majority of the data*

   In the random groups' design, examinee groups are assumed equal in ability. Thus, the mean difference between base and to-be-equated forms gives a reasonable indication of the direction and magnitude of transformation from non-equated scores. If the mean difference is negative in value when base scores are subtracted from raw to-be-equated scores, then the to-be-equated form is more difficult and should be converted to higher scores at the majority of the scale points. The opposite holds if the value is positive. If the magnitude of the mean difference between raw scores on these forms is small, equating methods that suggest radical conversions may not be justified by this difference in form to form difficulty.

3. *Parsimony*

   When two equating conversions are similar to each other, the simpler conversion will be used. The standard error for the equipercentile equating at each score level will be used to judge the degree of similarity between equating conversions.

4. *Smoothed distributional properties*

   An equating conversion that provides fewer gaps at the top or bottom of the percent correct scale will be chosen.

These criteria were used simultaneously, with the favored, and subsequently adopted, methods meeting all or most criteria.

Results

Table 4.3 on the following page shows a descriptive summary of the equating samples obtained in Reading. The bolded numbers in the table refer to the characteristics of the base form at each grade level. The base form at a grade level was the form administered in both the KCA and P&P mode. The number of items on a test form varies at certain grade levels as items were dropped after the test administration window due to classical item analyses, differential item functioning analyses, or printing errors. Table 4.3 below presents total scores on forms in terms of average correct and average percent correct. Also included is reliability information for each of the test forms. Table 4.3 shows that all the forms across grade levels had sufficient reliability for equating purposes.

Table 4.3. Descriptive statistics for equating samples for Reading by test form

Grade	Form	N of Items	N	Reliability (α)	Mean Raw Score	SD of Mean Score	Mean Percent Correct	SD Mean % Correct
3	**386**	**58**	**15997**	**0.90**	**44.64**	**8.844**	**76.98**	**15.187**
3	**386**	**58**	**4479**	**0.88**	**45.19**	**7.900**	**77.91**	**13.613**
3	522	58	4476	0.89	44.43	8.570	78.00	13.613
3	558	58	4475	0.91	45.38	8.982	77.95	13.647
3	559	58	4446	0.88	44.50	8.266	77.98	13.590
4	**404**	**74**	**13504**	**0.92**	**56.95**	**10.948**	**76.97**	**14.764**
4	**404**	**74**	**5142**	**0.91**	**58.11**	**10.012**	**78.54**	**13.498**
4	561	74	5169	0.92	59.38	9.891	78.46	13.653
4	562	74	5136	0.91	57.94	10.132	78.55	13.523
4	563	74	5117	0.92	58.76	10.997	78.53	13.598
5	**389**	**74**	**13038**	**0.92**	**56.98**	**11.356**	**77.03**	**15.299**
5	**388**	**74**	**7129**	**0.91**	**57.79**	**10.574**	**78.12**	**14.267**
5	565	74	3177	0.91	57.32	9.949	78.12	14.335
5	566	74	7098	0.88	57.19	9.021	78.01	14.483
5	5651	74	3952	0.89	58.30	9.077	78.22	14.232
6	**401**	**79**	**11885**	**0.93**	**58.86**	**12.926**	**74.54**	**16.280**
6	**401**	**79**	**5708**	**0.92**	**60.28**	**12.029**	**76.30**	**15.233**
6	569	80	5645	0.92	60.61	12.009	76.15	15.430
6	571	80	5683	0.92	61.13	11.692	76.24	15.320
6	572	80	5659	0.92	60.67	12.295	76.18	15.338
7	**4101**	**84**	**11368**	**0.93**	**61.89**	**13.501**	**73.71**	**16.008**
7	**410**	**84**	**5902**	**0.92**	**63.62**	**12.507**	**75.74**	**14.885**
7	573	84	5871	0.93	61.35	13.368	75.64	14.694
7	574	84	5936	0.93	61.30	13.880	75.85	14.873
7	575	83	5902	0.94	62.45	13.811	75.74	14.850
8	**677**	**80**	**10722**	**0.94**	**59.43**	**14.111**	**74.49**	**17.529**
8	**470**	**83**	**8378**	**0.93**	**63.66**	**12.903**	**76.35**	**16.080**
8	577	83	8410	0.94	63.35	13.340	76.34	16.083
8	578	83	8397	0.92	60.47	12.560	76.38	16.102
11	**592**	**77**	**9614**	**0.93**	**60.19**	**12.465**	**78.25**	**16.104**
11	**480**	**80**	**5766**	**0.93**	**60.45**	**12.479**	**77.79**	**14.538**
11	581	81	5748	0.93	63.08	12.271	77.71	14.656
11	582	81	5699	0.92	62.96	11.864	77.72	14.626
11	583	79	5709	0.92	60.73	11.877	77.65	14.745

Table 4.4 shows a descriptive summary of the equating samples in Mathematics. Again, bolded numbers in the table correspond to characteristics of the base forms at each grade level.

Table 4.4. Descriptive statistics for equating samples for Mathematics by test form

Grade	Test ID	N of Items	N	Reliability (α)	Mean Raw Score	SD of Mean Score	Mean Percent Correct	SD Mean % Correct
3	**405**	**70**	**14657**	**0.93**	**55.91**	**11.362**	**79.89**	**16.194**
3	**405**	**70**	**3949**	**0.92**	**57.79**	**9.905**	**82.56**	**14.163**
3	664	70	3912	0.92	57.91	10.072	82.73	13.946
3	665	69	3895	0.92	57.95	9.300	82.67	14.025
3	666	70	3913	0.91	58.26	9.503	82.67	14.006
3	669	70	3891	0.91	58.17	9.306	82.64	14.002
4	**595**	**73**	**12005**	**0.92**	**54.36**	**11.601**	**74.50**	**15.840**
4	**595**	**73**	**4502**	**0.92**	**56.42**	**10.884**	**77.29**	**14.927**
4	670	72	4479	0.91	54.95	10.750	77.49	14.667
4	671	73	4431	0.92	57.74	10.930	77.40	14.840
4	672	72	4459	0.92	55.03	11.440	77.47	14.742
4	673	73	4470	0.92	57.57	10.886	77.50	14.709
5	**406**	**73**	**12449**	**0.92**	**53.37**	**11.789**	**73.12**	**16.149**
5	**406**	**73**	**4499**	**0.91**	**54.98**	**10.957**	**75.31**	**15.026**
5	674	73	4446	0.91	53.79	10.816	75.64	14.701
5	675	73	4415	0.91	53.16	10.961	75.60	14.734
5	676	73	4379	0.91	54.74	10.920	75.59	14.865
5	678	73	4436	0.92	54.25	11.543	75.63	14.750
6	**479**	**86**	**11615**	**0.95**	**62.22**	**15.308**	**72.42**	**17.743**
6	**479**	**86**	**4691**	**0.94**	**64.46**	**14.492**	**74.99**	**16.834**
6	680	86	4620	0.94	64.22	14.619	75.31	16.521
6	681	86	4572	0.94	64.01	14.348	75.42	16.394
6	686	86	4584	0.94	63.12	14.190	75.23	16.460
6	687	86	4610	0.93	62.13	13.572	75.41	16.626
7	**597**	**82**	**10732**	**0.94**	**53.15**	**15.159**	**64.85**	**18.419**
7	**596**	**84**	**4941**	**0.94**	**56.10**	**14.491**	**66.79**	**17.528**
7	682	84	4853	0.94	54.89	15.258	66.71	17.352
7	683	84	4835	0.94	55.94	14.980	66.60	17.273
7	684	84	4846	0.95	57.31	15.785	66.59	17.310
7	685	84	4827	0.94	55.85	15.418	66.59	17.236
8	**586**	**85**	**10384**	**0.95**	**54.77**	**16.371**	**64.44**	**19.254**
8	**586**	**85**	**5185**	**0.94**	**56.56**	**15.496**	**66.54**	**18.219**
8	688	86	5151	0.94	57.77	15.409	67.15	17.938
8	690	86	5056	0.94	57.91	15.486	67.23	17.902
8	691	86	5062	0.94	58.13	15.054	67.12	17.947
8	700	86	5053	0.95	58.70	15.841	67.19	17.872
10	**590**	**84**	**11106**	**0.95**	**47.63**	**17.347**	**56.74**	**20.613**
10	**590**	**84**	**4966**	**0.95**	**48.30**	**16.953**	**57.51**	**20.176**
10	702	84	4816	0.95	51.76	16.811	57.95	20.031
10	719	84	4852	0.94	50.52	16.271	57.83	20.096
10	720	83	4881	0.94	49.75	16.269	57.81	20.054
10	721	84	4848	0.95	50.58	17.086	57.91	20.091

All forms at a grade level were constructed to the same content and statistical specifications, but items were dropped after test administration either because of errors in printing, or more commonly, due to both the classical and differential item functioning analyses. Table 4.4 presents total scores on forms in terms of average correct and average percent correct. Reliability information shows that all coefficients are greater than .90, indicating sufficient reliability for equating purposes.

Parts of an equating output from Grade 7 Mathematics are provided on the following pages. For this equating, Form 683 was to be equated to Form 596 on the total score scale. Both Forms 683 and 596 have 84 items. Employing the criteria listed in the previous paragraphs, the four moments of the equated scores from several competing methods were compared to the base form. Though moments from IRT observed scores are not typically computed, they were calculated to facilitate comparison between the various approaches. Table 4.5 shows that the linear and equipercentile methods yielded equated scores with the closest approximation of the first two moments of the distribution for the base form, as expected. Equating utilizing the IRT observed score approach provided moments most dissimilar to that of the distribution for the base form.

Table 4.5. Moments of the Equated Form 683 by Equating method

Test Form/Method	Mean	SD	Skewness	Kurtosis
Old Form: 596 n= 4692 **New Form:** 683 n= 4687 ;				
Raw Scores				
Form596	56.1032	14.4951	-0.4121	2.4631
Form683	55.9471	14.9796	-0.3358	2.2831
Form 683 equated to base form 596				
Unsmth:	56.1032	14.4901	-0.4116	2.4607
S=0.01:	56.1018	14.4969	-0.4136	2.4671
S=0.05:	56.1009	14.4983	-0.4139	2.4673
S=0.10:	56.1003	14.4996	-0.4138	2.4664
S=0.20:	56.0980	14.5121	-0.4130	2.4553
S=0.30:	56.0901	14.5416	-0.4083	2.4249
S=0.40:	56.0819	14.5606	-0.4033	2.4032
S=0.50:	56.0728	14.5740	-0.3982	2.3861
S=0.75:	56.0458	14.5934	-0.3841	2.3531
S=1.00:	56.0163	14.6006	-0.3697	2.3293
Linear:	56.1032	14.4951	-0.3358	2.2831
Irt tru:	55.7448	14.5605	-0.4006	2.4216
Irt obs:	55.7557	14.5399	-0.3941	2.4007

The mean difference between base Form 596 and Form 683 was .16. Figure 4.1 illustrates the conversions for each method graphically across the total score distribution. The Figure indicates that the equating methods other than the IRT and True Score approaches provided reasonable conversions across score levels, particularly in the area of the distribution where the majority of the data congregated.

Figure 4.1. Total Grade 7 Mathematics Form 683 equated to base Form 596

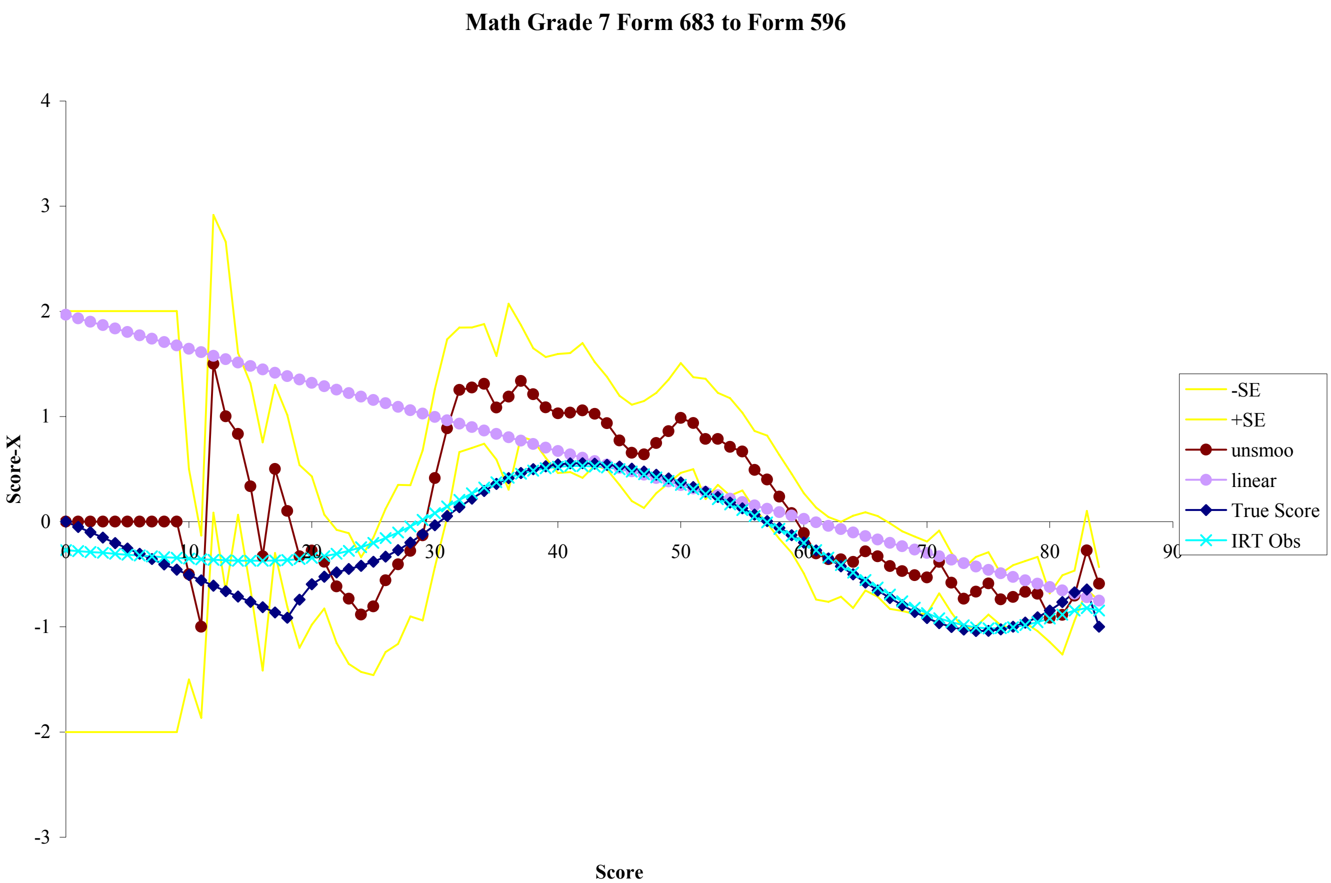


Table 4.6 shows a portion of the respective conversion tables for all competing methods (between raw score values of 56 through 84). Most conversions showed reasonable progression of equated scores through the raw score scale.

Table 4.6. Conversion table for various methods

Raw Score	TRUE Score Equiv	IRT Obs Equ Score	unsmoo	s=0.01	s=0.05	s=0.10	s=0.20	s=0.30	s=0.40	s=0.50	s=0.75	s=1.00	linear	Freq
56	56.07	56.05	56.49	56.52	56.49	56.44	56.38	56.38	56.37	56.36	56.29	56.21	56.15	113
57	57.00	56.99	57.40	57.39	57.35	57.33	57.29	57.31	57.31	57.30	57.25	57.18	57.12	98
58	57.94	57.93	58.24	58.24	58.21	58.21	58.21	58.24	58.25	58.25	58.21	58.14	58.09	103
59	58.87	58.87	59.08	59.06	59.07	59.1	59.13	59.17	59.19	59.19	59.16	59.10	59.06	109
60	59.80	59.80	59.89	59.89	59.94	59.98	60.04	60.10	60.13	60.14	60.12	60.06	60.03	96
61	60.73	60.73	60.70	60.73	60.82	60.88	60.96	61.03	61.07	61.08	61.07	61.03	60.99	99
62	61.65	61.66	61.64	61.65	61.74	61.8	61.89	61.97	62.00	62.02	62.02	61.99	61.96	109
63	62.58	62.59	62.64	62.63	62.68	62.73	62.82	62.90	62.94	62.97	62.98	62.94	62.93	129
64	63.50	63.52	63.62	63.66	63.65	63.68	63.76	63.84	63.89	63.91	63.93	63.90	63.90	120
65	64.42	64.44	64.72	64.68	64.63	64.63	64.70	64.78	64.83	64.86	64.88	64.86	64.86	118
66	65.35	65.37	65.67	65.66	65.61	65.6	65.65	65.73	65.77	65.80	65.83	65.82	65.83	110
67	66.27	66.31	66.58	66.59	66.58	66.57	66.60	66.67	66.72	66.75	66.78	66.78	66.80	97
68	67.20	67.24	67.53	67.53	67.56	67.54	67.56	67.62	67.67	67.70	67.73	67.74	67.77	105
69	68.14	68.18	68.49	68.5	68.53	68.51	68.51	68.57	68.62	68.65	68.69	68.69	68.73	108
70	69.08	69.13	69.47	69.52	69.5	69.48	69.48	69.53	69.57	69.60	69.64	69.65	69.70	115
71	70.04	70.08	70.62	70.53	70.47	70.45	70.44	70.49	70.52	70.55	70.59	70.61	70.67	101
72	71.00	71.04	71.42	71.43	71.43	71.42	71.41	71.45	71.48	71.50	71.54	71.56	71.64	99
73	71.97	72.01	72.27	72.33	72.38	72.39	72.38	72.41	72.43	72.45	72.49	72.52	72.61	98
74	72.96	72.99	73.33	73.33	73.35	73.36	73.35	73.37	73.39	73.41	73.45	73.48	73.57	112
75	73.96	73.99	74.41	74.35	74.33	74.33	74.33	74.33	74.35	74.36	74.40	74.43	74.54	94
76	74.97	74.99	75.26	75.31	75.31	75.3	75.30	75.30	75.31	75.32	75.35	75.39	75.51	81
77	76.00	76.00	76.28	76.3	76.28	76.28	76.28	76.27	76.27	76.28	76.31	76.35	76.48	86
78	77.04	77.02	77.33	77.31	77.25	77.26	77.25	77.24	77.23	77.23	77.26	77.30	77.44	58
79	78.09	78.04	78.31	78.24	78.22	78.23	78.23	78.21	78.19	78.19	78.21	78.26	78.41	44
80	79.15	79.08	79.09	79.13	79.2	79.21	79.21	79.18	79.16	79.15	79.17	79.22	79.38	47
81	80.24	80.12	80.11	80.14	80.2	80.2	80.19	80.15	80.12	80.11	80.12	80.17	80.35	41
82	81.33	81.15	81.30	81.3	81.24	81.22	81.20	81.14	81.10	81.08	81.09	81.16	81.31	33
83	82.36	82.18	82.73	82.58	82.55	82.54	82.52	82.48	82.46	82.45	82.45	82.50	82.28	11
84	83.00	83.16	83.41	83.86	83.85	83.85	83.84	83.83	83.82	83.82	83.82	83.83	83.25	4

Table 4.7 shows the same conversion table when transformed onto the percent correct metric for the same sample of score values. In this percent correct metric, differences in distributional smoothness throughout the scale are not immediately apparent. However, using these conversion tables in accord with the moments output from the various methods and the graphical representation of the conversions for each method is useful. Given a consideration of the multiple criteria discussed for making equating decisions, it appeared that the unsmoothed method with a smoothing parameter applied (s=.05) gave the most reasonable conversion.

Table 4.7. Conversion table for various methods expressed in percent correct metric

Raw Score	TRUE Score Equiv	IRT Obs Equ Score	unsmoo	s=0.01	s=0.05	s=0.10	s=0.20	s=0.30	s=0.40	s=0.50	s=0.75	s=1.00	linear	Freq
56	66.75	66.73	67.25	67.29	67.25	67.19	67.12	67.11	67.11	67.09	67.01	66.92	66.85	113
57	67.86	67.85	68.33	68.32	68.28	68.25	68.21	68.22	68.23	68.22	68.16	68.07	68.00	98
58	68.98	68.97	69.33	69.33	69.3	69.3	69.30	69.33	69.35	69.34	69.30	69.21	69.15	103
59	70.08	70.08	70.33	70.31	70.32	70.35	70.39	70.44	70.46	70.47	70.43	70.36	70.31	109
60	71.19	71.19	71.30	71.29	71.35	71.41	71.48	71.55	71.58	71.59	71.57	71.50	71.46	96
61	72.29	72.30	72.26	72.3	72.41	72.48	72.58	72.66	72.70	72.71	72.70	72.65	72.61	99
62	73.40	73.40	73.38	73.39	73.49	73.57	73.68	73.77	73.81	73.84	73.84	73.79	73.76	109
63	74.49	74.51	74.57	74.56	74.62	74.68	74.79	74.88	74.93	74.96	74.97	74.93	74.91	129
64	75.59	75.61	75.73	75.78	75.77	75.8	75.90	76.00	76.05	76.08	76.10	76.08	76.07	120
65	76.69	76.72	77.05	77	76.94	76.95	77.02	77.12	77.18	77.21	77.24	77.22	77.22	118
66	77.79	77.83	78.18	78.17	78.1	78.09	78.15	78.25	78.30	78.34	78.37	78.36	78.37	110
67	78.90	78.94	79.26	79.28	79.27	79.25	79.28	79.37	79.43	79.46	79.50	79.50	79.52	97
68	80.01	80.05	80.39	80.39	80.42	80.4	80.42	80.50	80.56	80.59	80.64	80.64	80.67	105
69	81.12	81.17	81.54	81.55	81.58	81.56	81.56	81.64	81.69	81.72	81.77	81.78	81.83	108
70	82.24	82.30	82.70	82.76	82.74	82.72	82.71	82.77	82.82	82.85	82.90	82.92	82.98	115
71	83.38	83.43	84.07	83.96	83.89	83.87	83.86	83.91	83.95	83.99	84.03	84.05	84.13	101
72	84.52	84.58	85.02	85.03	85.03	85.02	85.01	85.05	85.09	85.12	85.17	85.19	85.28	99
73	85.68	85.73	86.03	86.11	86.17	86.18	86.17	86.20	86.23	86.25	86.30	86.33	86.43	98
74	86.86	86.90	87.30	87.3	87.33	87.33	87.32	87.34	87.37	87.39	87.44	87.47	87.59	112
75	88.05	88.08	88.59	88.51	88.49	88.49	88.48	88.49	88.51	88.53	88.57	88.61	88.74	94
76	89.26	89.27	89.60	89.65	89.65	89.65	89.64	89.64	89.65	89.67	89.70	89.75	89.89	81
77	90.48	90.47	90.81	90.83	90.81	90.81	90.81	90.80	90.80	90.80	90.84	90.89	91.04	86
78	91.71	91.69	92.06	92.04	91.97	91.97	91.97	91.95	91.94	91.94	91.98	92.03	92.19	58
79	92.97	92.91	93.23	93.14	93.12	93.13	93.13	93.10	93.09	93.08	93.11	93.17	93.35	44
80	94.23	94.14	94.15	94.2	94.29	94.3	94.30	94.26	94.23	94.22	94.25	94.31	94.50	47
81	95.52	95.38	95.37	95.4	95.47	95.48	95.47	95.41	95.38	95.36	95.38	95.45	95.65	41
82	96.82	96.61	96.78	96.79	96.72	96.7	96.67	96.59	96.54	96.52	96.54	96.62	96.80	33
83	98.04	97.83	98.48	98.31	98.27	98.26	98.24	98.19	98.16	98.15	98.16	98.21	97.95	11
84	98.81	99.00	99.30	99.83	99.82	99.82	99.81	99.79	99.78	99.78	99.78	99.80	99.11	4

Equating Decisions

The equating method selected for each form in Reading and Mathematics is summarized in this section of the report. A total of 20 equating analyses were performed in Reading and 28 in Mathematics, each subjected to the criteria previously listed. With past Kansas Assessments, equating analyses were performed at the knowledge/application sub-test level in Mathematics and at the text-type sub-test level in Reading. Equating analyses in 2006 were only conducted at the total score level. For all forms, a value of zero on the raw score scale converted to a value of zero regardless of equating method adopted. Additionally, equated scores that had a negative value were set to the minimum score value of zero. Similarly, equated scores that were greater than the top score on the base form were set to a percent correct value of 100%.

All methods selected across grade levels in both content areas required the use of conversion tables. These conversion tables, similar to the samples detailed in Tables 4.6 and 4.7, are not provided in this report in a consideration for space. After comparing all possible methods employing the criteria set forth in the previous section, decisions were made for each form individually. For Reading, the smoothed equipercentile equating method was chosen exclusively for all 20 equating decisions. The smoothing parameter selected for each decision was .01. All forms were equated at the raw score level and subsequently expressed in the percent correct metric.

Table 4.8 details the equating decisions adopted for Mathematics. The smoothed equipercentile equating method was selected for all 28 equating decisions. The smoothing parameters varied across forms.

Table 4.8. Summary of equating decisions for Mathematics

	Form			Equating
Grade	Base	Equated	Scale	Decision
3	405	664	Total	Smoothed Equipercentile (s=0.01)
3		665	Total	Smoothed Equipercentile (s=0.01)
3		666	Total	Smoothed Equipercentile (s=0.01)
3		669	Total	Smoothed Equipercentile (s=0.05)
4	595	670	Total	Smoothed Equipercentile (s=0.05)
4		671	Total	Smoothed Equipercentile (s=0.05)
4		672	Total	Smoothed Equipercentile (s=0.05)
4		673	Total	Smoothed Equipercentile (s=0.05)
5	406	674	Total	Smoothed Equipercentile (s=0.10)
5		675	Total	Smoothed Equipercentile (s=0.10)
5		676	Total	Smoothed Equipercentile (s=0.05)
5		678	Total	Smoothed Equipercentile (s=0.05)
6	479	680	Total	Smoothed Equipercentile (s=0.05)
6		681	Total	Smoothed Equipercentile (s=0.05)
6		686	Total	Smoothed Equipercentile (s=0.10)
6		687	Total	Smoothed Equipercentile (s=0.05)
7	596	682	Total	Smoothed Equipercentile (s=0.05)
7		683	Total	Smoothed Equipercentile (s=0.05)
7		684	Total	Smoothed Equipercentile (s=0.05)
7		685	Total	Smoothed Equipercentile (s=0.05)
8	586	688	Total	Smoothed Equipercentile (s=0.10)
8		690	Total	Smoothed Equipercentile (s=0.05)
8		691	Total	Smoothed Equipercentile (s=0.10)
8		700	Total	Smoothed Equipercentile (s=0.10)
10	590	702	Total	Smoothed Equipercentile (s=0.10)
10		719	Total	Smoothed Equipercentile (s=0.05)
10		720	Total	Smoothed Equipercentile (s=0.10)
10		721	Total	Smoothed Equipercentile (s=0.10)

Summary

The 2006 Kansas Assessments tests forms in Reading and Mathematics were built, by grade level and content area, to the same specifications as articulated by KSDE and were developed to the same statistical specifications. The reliability coefficients for all forms were acceptable for the purpose of equating. Further, data collected from the spring 2006 administration show that groups administered various test forms appeared to be random.

Several equating methods were considered including IRT and classical methods, and thus certain criteria were used to select the equating method for a particular test form that would provide for the most equitable scores for the Kansas students administered the assessments. Methods that best fit the data through the criteria listed were selected.

While test forms were constructed to the same content and statistical specifications, a handful of items were not retained after the spring 2006 administration of the tests. Some items were lost after item analyses, a few items were dropped because of printing errors, and a few items were dropped due to their functioning differentially for different subgroups of students. Items that were removed from forms will not be replaced with scored items in future administrations as the equating relationships between particular test forms and the base form have accounted for the item loss.

Section 5

STANDARD SETTING

Standard Setting for the 2006 Kansas Assessments in Reading and Mathematics was conducted in late spring through summer 2006. CETE implemented various standard setting methodologies to recommend performance level standards for the Kansas Assessments, including the general and KAMM assessments in Reading and Mathematics as well as the Kansas Alternate Assessment (KAA). For the general Reading and Mathematics assessments, three standard setting methodologies were implemented: Bookmark, Borderline Group, and Contrasting Groups. For the KAMM tests, the Bookmark standard setting procedure was used to recommend cut scores. For the KAA, the Body of Work method was utilized to make performance category recommendations.

The following pages document the standard setting processes conducted for each instrument. Part 1 provides an overview of the Bookmark standard setting workshop for Reading and Mathematics, including a description of the participants, the process and procedures, and evidence of the successfulness of the activity. Part 2 describes the implementation of the Borderline Group and Contrasting Groups standard setting methods. Part 3 highlights the process and participants used for recommending cut scores on the Kansas Alternate Assessment (KAA). Part 4 details the "super committee" panels that synthesized the judgments made by educators in the standard setting processes and made the final recommendations to the Kansas State Department of Education (KSDE).

Part 1: Bookmark Standard Setting

The Bookmark standard setting procedure (Mitzel, Lewis, Patz, & Green, 1996) was implemented to recommend performance level cut scores for the general and KAMM assessments for the content area tests for grades 3-8 (Mathematics and Reading) and grade 10 (Mathematics) and grade 11 (Reading). The Bookmark procedure consisted of training, orientation, and three rounds of judgments by participants. The Bookmark standard setting for the Kansas general Mathematics and KAMM assessments was held in conjunction on June 19 and 20, 2006, in Salina, Kansas. The standard setting for the Kansas general Reading and KAMM assessments was held in conjunction June on 20 and 21, 2006, in Salina, Kansas. Each content area Standard Setting lasted two days, with one-half day devoted to Table Leader training and approximately one-and-one-half days devoted to the actual procedure.

The Bookmark standard setting was conducted to recommend cut scores on each grade-level test (general Reading, general Mathematics, KAMM Reading, KAMM Mathematics) that separate students into five performance levels defined by the state of Kansas: *Unsatisfactory, Basic, Proficient, Advanced, and Exemplary* (performance levels were later renamed: *Academic Warning*, *Approaches Standard*, *Meets Standard*, *Exceeds Standard*, and *Exemplary*).

Bookmark Standard Setting Roles

CETE Staff

CETE worked with staff from the Kansas State Department of Education (KSDE) to design and organize the standard setting. The CETE standard setting team was comprised of John Poggio, Ph.D., Doug Glasnapp, Ph.D., Patrick Irwin, Ph.D., and Andrew Poggio. Dr. Poggio and Dr. Glasnapp are co-Directors of CETE. Dr. Irwin and Andrew Poggio are Research Assistants for CETE. During the standard setting, John Poggio, Patrick Irwin, and Andrew Poggio were responsible for facilitating the standard setting meeting, training participants, and monitoring the participant results database. Prior to the meeting, the entire CETE group prepared all materials related to the standard setting.

Additionally, four graduate students working for CETE on assistantships attended the meeting and served as technical staff working mainly in the operations room, taking responsibility for the entry of the judgment data into the optical scanner. These graduate students also implemented quality assurance procedures, such as double-checking the data. CETE support staff also played an integral role in planning and facilitating the standard setting meeting. Amy Tackkett, Ronda Consolver, and Wendy Coonrod were primary points of contact with KSDE and assumed a variety of roles before and during the standard setting meeting, including creating the registration process and conducting it on-site.

Table Leaders

There were 21 tables of Bookmark standard setting panelists in each content area (two tables at each grade level for the general assessments and one table per grade level for the KAMM). Every table had a Table Leader who was chosen prior to the standard setting. Table Leaders were chosen based on years of experience as a respective grade-level educator as well as by information received through the participant nomination process. Table Leaders were active, voting participants in the standard setting process. They received additional training on the Bookmark method in a session prior to the arrival of the rest of the participants (duration: approximately 4 hours). The primary role of the Table Leaders was to lead the table discussions, monitor the group discourse, facilitate discussion, collect materials, and maintain the schedule.

Participants

Participants in the Bookmark standard setting meeting for the general and KAMM assessments were selected by CETE from a pool of 550 nominees. These nominations were made by Kansas educators (e.g., administrators, school faculty, district coordinators). The participants were selected based on factors such as grade and content area of primary instruction, geographic location, school size, and years of experience teaching in Kansas. Every effort was made to select a representative group of participants. Within each content area for the general assessments, participants were divided into two groups that were balanced as well as possible in terms of relevant demographic characteristics. For the KAMM, one table per content area, which was similarly balanced, was used. The number of participants in each content area is reported in

Table 5.1. Participants were asked to complete an evaluation following the standard setting. Using these evaluations, demographic information about the participants was summarized, specifically the academic position (administrator, coordinator, classroom teacher), teaching experience, district location, and building status (in terms of SES) of the panelists.

Table 5.1. Number of Participants in Each Content Area on General and KAMM Assessments

Content Area	General	KAMM
Reading	106	44
Mathematics	103	39

Bookmark Materials

Ordered Item Booklets

The Ordered Item Booklets (OIBs) for each grade level and content area were comprised of items from an operational form of the respective assessment administered for the first time in Spring 2006. For the general Reading and Mathematics assessments, the form used to create the OIBs was the "base" form; that is, the form that was available for computerized or traditional delivery. For the KAMM forms, the single operational form in existence was used to create the OIBs. The items for each grade level assessment form utilized were ordered in terms of difficulty using a 2-parameter logistic (2PL) Item Response Theory (IRT) model. The OIBs were ordered from the easiest to the hardest item. Table 5.2 summarizes the number of score points in each OIB by test and content area.

Table 5.2. Number of Score Points in Ordered Item Booklets by Test Form and Content Area

Level	Test/Content Area	Number of Score Points in OIB
Grade 3	General Reading	58
	General Mathematics	70
	KAMM Reading	30
	KAMM Mathematics	40
Grade 4	General Reading	74
	General Mathematics	73
	KAMM Reading	37
	KAMM Mathematics	40
Grade 5	General Reading	74
	General Mathematics	73
	KAMM Reading	45
	KAMM Mathematics	40
Grade 6	General Reading	79
	General Mathematics	86
	KAMM Reading	47
	KAMM Mathematics	40
Grade 7	General Reading	84
	General Mathematics	82
	KAMM Reading	46
	KAMM Mathematics	40
Grade 8	General Reading	80
	General Mathematics	85
	KAMM Reading	48
	KAMM Mathematics	40
Grade 10	General Mathematics	84
	KAMM Mathematics	40
Grade 11	General Reading	77
	KAMM Reading	49

Standard Setting Day 1, Mathematics: Table Leader Training and Orientation

Training

Table Leaders were trained on the morning of the first day of the Kansas Bookmark standard setting for the general and KAMM Mathematics assessments. During this session, which lasted approximately four hours, Table Leaders were given an overview of the standard setting process and were trained specifically on the Bookmark method. The Table Leader training went into detail regarding the type of behavior expected of Table Leaders during the Bookmark standard setting.

Standard Setting Day 1, Mathematics: Participant Training and Orientation

Orientation

CETE welcomed the panelists to the Kansas Mathematics general and KAMM assessments standard setting in the afternoon of Day 1. All participants were checked in and given a packet of materials that included the Powerpoint slides used during the training session. The meeting took place in a large conference space in Salina, Kansas, a central location in the state that allowed for participation from educators from all areas. All of the tables (21 in all) were located in one large room. For the general assessment, two tables of panelists per grade level were utilized for the purpose of comparison. Tables were arranged by CETE so that any two tables at a grade level were not in proximity to one another. Upon registration and check-in, participants made their way to their tables and the training session commenced. John Poggio of CETE ran the training session, which lasted approximately three-and-one-half hours. The orientation and training consisted of a brief overview of the Kansas testing program followed by an overview of the standard setting process. A description of the review procedures that would follow the Bookmark meeting was also provided to panelists. Participants were trained on the use of the OIBs during the orientation session as well.

A majority of the orientation session was devoted to training on the Bookmark method, specifically how to place a bookmark. In the training materials provided, several explanations of bookmark placement were described. It was explained that for the Kansas assessments, four cut scores or "bookmarks" would need to be set to provide for the five performance level categories: *Unsatisfactory, Basic, Proficient, Advanced,* and *Exemplary.* The mechanics of bookmark placement were then described, with an explanation that all items preceding the bookmark define the knowledge, skills, and abilities that an *Advanced* student, for example, is expected to do. Participants were instructed to examine each item in terms of its content and the knowledge and skill requirements of the item and make a judgment about the type of content a student would need to know in order to be considered barely *Advanced.*

The final topic covered in the training session was using a response probability to determine an item's association with a given performance category. This explanation linked the use of the 2/3 or .67 mastery probability to bookmark placement.

Round 1

All participants were provided materials via mail to be reviewed prior to the Bookmark standard setting meeting. Panelists were instructed to review and consider Kansas' Performance Level Descriptors (PLDs) that define the five performance categories. Round 1 began with an opportunity for tables to review and discuss the PLDs. Then, participants spent approximately one-and-a-half hours taking the operational grade level test in Mathematics via the OIB. Panelists were asked to familiarize themselves with the test, specifically studying each item in terms of what each item measured, the knowledge, skills, and abilities required by the item, and why each item is more difficult than the items preceding it. The panelists were directed to place their bookmarks, starting with *Unsatisfactory,* keeping in mind the Performance Level Descriptors. Participants were instructed during training and reminded by the Table Leaders that bookmark placement is an individual activity. A dual-mode strategy for bookmark recording was utilized. For efficient and accurate database entry, a scannable recording form was developed by CETE and used for this meeting. Panelists were asked to bubble the bookmark placement for the appropriate performance category on the rating form. Also, a traditional hand-recording form was collected from each panelist as they recorded their judgments round by round. This form was a backup that was also used for accuracy checks. After the appropriate Round 1 recordings were made, the meeting adjourned for the day.

Standard Setting Day 2, Mathematics: Rounds 2 and 3

Round 2

Round 2 commenced with Round 1 results returned to each table. Each participant was provided with the minimum, maximum, and median cut score recommendation for each performance category for the table. Panelists were asked to discuss those items that fell between the first and last placed bookmark per performance category; that is, they were instructed to discuss those items for which there was disagreement among the group. After discussion, the participants were asked to place their Round 2 bookmarks. Again, participants were reminded that bookmark placement is an individual activity.

Round 3

At the beginning of Round 3, the two tables of grade-level general assessment participants were combined and descriptive results (minimum, maximum, median) regarding cut score recommendations for each performance category were provided to each table and also as a combination of the two tables. Relevant cumulative frequency distributions were administered to each table as well. John Poggio of CETE addressed the entire room of participants, presenting a grade-level sample of aggregated impact data based on the Round 2 bookmarks. It was emphasized to participants that the impact data was being presented as a “reality check.” The Table Leaders from the two general assessment groups then facilitated discussion among the panelists on their bookmark placements. For KAMM tables, impact data was also presented and followed by a discussion. CETE staff answered process-related questions, while KSDE staff was present and available for any policy-related questions concerning the impact data that arose. After discussion, panelists placed their final bookmarks. Participants were again reminded that bookmark placement is an individual activity. Upon completion of the Round 3 bookmark placements, the participation of the panelists in the Bookmark standard setting for the Kansas Mathematics assessments had concluded. Prior to leaving the meeting, participants were asked to complete an evaluation of the standard setting meeting.

The Bookmark standard setting meeting for the Kansas general and KAMM assessments in Mathematics and Reading took place over three days in June 2006 (June 19-21), with Mathematics occurring on June 19 and 20 and Reading on June 20 and 21. In essence, as the standard setting for Mathematics was finishing, the participants for the Reading standard setting were arriving. The methodology and procedures followed for both content areas were almost exclusively the same. Unless noted, the process for the Bookmark standard setting for Reading was the same as the process for Mathematics. Additional detail will be provided below only where there were deviations between the two implementations.

Standard Setting Day 1, Reading: Table Leader Training and Orientation

Training

Table Leaders for Reading were trained on the morning of the first day of the Reading standard setting, which was the second day of the Mathematics standard setting meeting. Andrew Poggio of CETE led the Table Leader training session. The presentation was identical to the orientation for the Mathematics Table Leaders, and lasted approximately four hours during which the participants were given an overview of the standard setting process and trained specifically on the Bookmark method. They were given a synopsis of the meeting activities as well as their responsibilities and tasks. The Table Leader training went into detail regarding the type of behavior expected of Table Leaders during the Bookmark standard setting.

Standard Setting Day 1, Reading: Participant Training and Orientation

Orientation

Participants were welcomed to the Kansas Reading general and KAMM assessments standard setting in the afternoon of day 1. The process and procedures for registration, orientation, and training for Reading panelists were identical to those for the Mathematics participants described above.

Round 1

The same procedures were implemented for Round 1 of Reading as Round 1 of Mathematics. The Performance Level Descriptors for Reading were sent to all participants prior to the standard setting meeting as with Mathematics. The PLDs were reviewed and discussed by participants prior to taking the grade-level Reading assessment. Participants were reminded that bookmark placement is an individual activity, and then they were asked to place their Round 1 bookmarks. The standard setting meeting adjourned for the day after this activity was completed.

Standard Setting Day 2, Reading: Rounds 2 and 3

Round 2

The same process was followed for Round 2 of the Reading standard setting as the process for Round 2 of Mathematics.

Round 3

The same process was followed for Round 3 of the Reading standard setting as the process for Round 3 of Mathematics. Upon completion of the Round 3 bookmark placements, the participation of the panelists in the Bookmark standard setting for the Kansas Reading assessments had concluded. Prior to leaving the meeting, participants were asked to complete an evaluation of the standard setting meeting.

Evaluations of the Kansas Bookmark Standard Settings

Quality Control Procedures

CETE adhered to several quality control procedures to foster accuracy of the standard setting materials and the results presented during the standard setting. Prior to the meeting, the ordering of the items in the OIBs was checked along with the accuracy and completeness of test information, training materials, and impact data tables. During the meeting, data were collected by a dual-mode recording strategy. Data were entered via the scannable form, and then checked for accuracy by CETE staff. Results from each round computed by SPSS were checked for accuracy before being returned to tables. Any results that appeared to be questionable were investigated further.

Effectiveness of Training

At the end of the respective standard setting activities, participants were asked to respond to an evaluation of the standard setting meeting. Evidence of the successfulness of the standard setting may be found in the participants' self-reported understanding of training, perceived validity of the activity, and confidence in their bookmark placements. During the Kansas general and KAMM assessments standard setting, participants were trained during the orientation session before placing bookmarks. Table Leaders spent an extra half-day in training before the start of the workshop.

The majority of participants indicated that the training was adequate in preparation for the judgment tasks. The majority of participants agreed that the training was useful in the standard setting.

Perceived Validity

Another indication of the successfulness of the standard setting may be found in the participants' perceived validity of the meeting. A majority of the participants understood the purpose of the standard setting and were confident that the process and methods used in the standard setting can contribute to valid and reliable cut score recommendations.

Part 2: Borderline Group and Contrasting Groups Standard Settings

As various standard settings procedures will invariably produce different cut score recommendations, multiple methods were implemented by CETE to recommend performance level standards for the Kansas general Reading and Mathematics assessments. Two additional student-centered methods were utilized, specifically the Contrasting Groups (Livingston and Zieky, 1982) and Borderline Group (Livingston and Zieky, 1982) methods.

A memo was sent to all Kansas building administrators, teachers, and test coordinators calling for participation and detailing the standard setting activities on April 13, 2006. As Kansas offers two modes of delivery for the general assessments, teachers were asked to make individual student ratings online if the assessment was taken via computer or make the ratings on their students' answer sheets if the assessment was taken via paper-and-pencil in the appropriate location on the scannable answer sheet. If answer sheets for paper-and-pencil tested students had already been returned, student ratings could be entered on-line as described for computerized tested students. For both testing modes, teachers making the student judgments were provided the Kansas Performance Level Descriptors (PLDs) and were asked to place their current students into an appropriate category: *Unsatisfactory, Basic, Proficient, Advanced, Exemplary* (for the Contrasting Groups method) or between categories (for the Borderline Groups method).

Participants

Professional judgments were solicited from teachers for students they had taught in the relevant subject area (Reading, Mathematics) during the 2005-06 academic year. Because making independent ratings of students in terms of the Kansas performance categories is a time-consuming task, guidelines for participation were outlined in the April 13th letter to building administrators from CETE. Specifically, teachers and students were requested as follows:

- Where the school district is an odd number, the following grades were requested in the respective building as available: Grades 3, 5, 7, and 11 to rate students in Reading, and Grades 4, 6, and 8 to rate students in Mathematics; **OR**,
- Where the school district is an even number, the following grades were requested in the respective building as available: Grades 4, 6, and 8 to rate students in Reading, and Grades 3, 5, 7, and 10 to rate students in Mathematics.
- Private schools participating in this activity were to use their Building Number as a referent for "odd or even" and follow the rules above for grade and content selection.

Further, the number of students to be rated depended on the number of students at a grade in a building. Participation in terms of students to be rated was based on the following guidelines:

a) For buildings with fewer than 50 students at the grade, it was requested that all students taking the general assessment be rated. **OR**,
b) For buildings with 51 or more students at the grade taking the general assessment, it was requested that a random sample of 50 students reflecting the make-up of the entire grade-level be rated. Sampling intact classes was deemed acceptable as long as that class could be considered representative.

Across all grade levels, for the Contrasting Groups standard setting method, there were 33,732 student ratings made by teachers for Reading and 33,435 student ratings in Mathematics. For the Borderline Group standard setting, 12,850 students were rated across all grade levels in Reading and 12,485 students were rated in Mathematics.

Contrasting Groups

Participating teachers were provided the Kansas PLDs and were asked to place their current students into the appropriate category. Teachers were instructed to make this evaluation independently.

Borderline Group

If teachers were uncomfortable placing a student into a single performance category, they were provided the option to place students between two categories or on the border between two categories. These ratings were used to apply the Borderline Group standard setting method.

Part 3: Kansas Alternate Assessment Standard Setting

As an overview, the following identify the major activities required to implement the Kansas Alternate Assessment (KAA).

1. Students meeting the criteria for receiving the Alternate Assessment must be registered. Criteria for participation are found on the KS Assessment/KS Alternate Assessment links at www.kansped.org.
2. Five indicators are to be selected by the local IEP team as the focus of the assessment for each content area being assessed in any one year. At least one indicator must be selected for assessment from <u>each</u> of the Extended Curriculum Standards areas. This applies to both the Reading and the Mathematics Extended Standards. In Reading, there are three (3) Standards' areas: 1) Reading, 2) Literature, and 3) Communication for Social Interaction. At least one indicator from each of these three areas must be selected for a student's assessment in Reading. The other two indicators may both be selected from one Standards' area or spread across two areas. In Mathematics, there

are four (4) Standards' areas: 1) Number and Computation, 2) Algebra, 3) Geometry, and 4) Data. At least one indicator from each of these four areas must be selected for a student's assessment in Mathematics. The remaining indicator may be selected from any of the four Standards' areas.

3. Three pieces of evidence are to be collected during the assessment window for each indicator and submitted to a student's folio using state identified procedures and forms. This will result in 15 separate pieces of evidence in Reading and 15 in Mathematics if a student is assessed in each area.
4. Each piece of evidence submitted is to be <u>independently</u> rated by three local raters using a required rubric on the amount/extent of the skill demonstrated related to an indicator.
5. Local score ratings are transmitted to CETE and serve as the data elements for scoring and reporting.

All district/buildings were asked to maintain and store all Alternate Assessment evidence files (student data folios). Following the completion of the Spring 2006 assessments, CETE randomly identified a 10 percent sample from the Alternate Assessment student population in the state (from the state's student ID database) and requested from each district in which a student sampled was tested that the sampled student's complete data folio be sent to CETE. Thus, a random sample of student data folios was gathered by the state's testing contractor for use in standard setting and in other studies (validity) as appropriate.

These latter student data folios were used to implement the expert judgmental review standard setting procedure commonly labeled as the "Body of Work" methodology (Kingston, Kahl, Sweeney, and Bay, 2001). The method requires that expert judges review the "Body of Work" (in this case the students' data folios) and categorize the work into one of the performance levels being used by the state based on state's descriptions of the performance levels made available to them. These judgments are used to classify the student data folios into categories of performance based on the external and independent review and opinion of the expert panel. Once classified, the data folios in the performance level categories form the basis to examine and compare the mean <u>local</u> ratings on the folios across categories much in the way one would conduct a Contrasting Groups analysis for standard setting.

Body of Work Data Collection Procedures

Procedurally for Kansas, the expert panel (n = 18) to participate in the "Body of Work" review was selected by KSDE from a list of individuals who had been trained by KSDE and who had served as trainers in implementing and scoring the Alternate Assessment for other local personnel. From the pool of randomly sampled student data folios, samples representing students across grades and content (Reading and Mathematics) were drawn and sent for review by the independent panel of trained expert judges. Expert review and performance level classification judgments were obtained for 127 data folios in Reading and 132 in Mathematics. Student scores based on the local ratings for each of the sampled data folios were then linked to the expert judges' performance level classifications, and mean local rating scores were computed for the data folios placed by the experts in each performance level category. These means served as one piece of information supplied to a "super committee" panel that was convened to make final recommendations on cut scores to the State Department and State Board of Education.

Part 4: Super Committee Standard Setting Meetings

For each assessment type (general, KAMM, Alternate), a panel of participants was convened to examine all of the recommended cut scores for each assessment type and grade level and to synthesize the information in an attempt to make a final recommendation to the Kansas State Department of Education (KSDE). For the general Mathematics and Reading assessments, the super committee members were presented with cut score recommendations from the Bookmark, Contrasting Groups, and Borderline Group methods. The KAMM super committee panels were presented the cut score recommendations from the Bookmark standard setting. The KAA super committee members reviewed the KAA cut scores from the Body of Work standard setting. Each super committee was presented with an explanation of the method(s) implemented, recommended cut scores for the assessments from the various methods used, student performance or impact data, and, where available, participant information and standard setting evaluation results.

The general Mathematics, general Reading, KAMM Mathematics, KAMM Reading, and KAA cut score recommendations were each reviewed by a unique super committee. For the general assessment content areas, two panels comprised the committee. One panel reviewed primarily cut score recommendations from the elementary grades levels (3, 4, 5) while the other focused primarily on the cut score recommendations on the assessments at the middle and high school grade levels (6, 7, 8, 10/11). For the KAMM Mathematics, KAMM Reading, and KAA, the respective super committees reviewed the cut score recommendations at all grade levels. The general Reading assessment super committee met in Lawrence, Kansas on July 13, 2006, and the general Mathematics assessment super committee met on the July 14, 2006, in Lawrence. The KAMM and KAA super committees met on August 2, 2006, in Lawrence, Kansas.

Super Committee Meetings

The super committee meetings were conducted by CETE with representatives from KSDE in attendance. Additionally, the general assessment super committee meetings were video recorded. Dr. John Poggio and Dr. Doug Glasnapp of CETE directed the meetings with Dr. Patrick Irwin, Andrew Poggio, and Dr. Jonathan Templin, CETE research staff, assisting.

The super committee meetings began with an orientation that introduced the task at hand and the goals of the meetings. Dr. Poggio conducted the orientation for the general assessment super committee panels and Andrew Poggio conducted the overview for the KAMM and KAA super committee panels. Participants were provided with an explanation of the standard setting methods used to collect the cut score recommendations they would be reviewing, information on the panelists that provided the scores and evaluation results where available, as well as the Kansas Performance Level Descriptors (PLDs).

After the orientation, the super committee panels reviewed all of the information provided for a given assessment. At this point, the panels were separated on the basis of grade level for the general assessments. For the KAMM assessments and the KAA, the specific assessment committees convened after the group orientation. After reviewing the cut score recommendations from the various standard setting methods implemented previously, considering impact data, reviewing any other information provided for a given assessment, and engaging in a group discussion of the information provided, the panels were asked to make individual cut score recommendations for the performance categories on a given test at each grade level being considered. Each super committee member's individual cut score recommendations were visually displayed to the panel. The range of the individual cut score recommendations served as the basis for a subsequent round of discussion where the goal was to determine a single cut score recommendation at each performance category, or at least to reduce the range at each performance level. In most cases, after group discussion, a single score point was recommended while in a few cases, a very small range of scores was selected. This was done for each of the performance categories within a grade level. The panel members then considered the next grade level and repeated the process of reviewing materials, making individual recommendations, engaging in group discussion, and then coming to group consensus. After reviewing each of their assigned assessments, the super committee members were presented each recommended cut score across all grade levels visually and were asked to consider and discuss further, if necessary, these scores and make any final adjustments. The final cut score recommendations from the super committees were given to KSDE along with impact data. KSDE made the cutscore recommendations to the Kansas State Board of Education and were subsequently approved. The cut scores, as approved by KSBE, are provided below.

Part 5:
KANSAS ASSESSMENTS
PERFORMANCE LEVEL CUTSCORES

General Reading Assessment
% correct

Grade	Academic Warning	Approaches Standard	Meets Standard	Exceeds Standard	Exemplary
3rd	0-54	55-66	67-79	80-88	89-100
4th	0-56	57-67	68-80	81-88	89-100
5th	0-56	57-67	68-79	80-87	88-100
6th	0-51	52-63	64-78	79-87	88-100
7th	0-49	50-62	63-76	77-86	87-100
8th	0-49	50-63	64-78	79-88	89-100
High School	0-53	54-67	68-80	81-88	89-100

General Mathematics
% correct

Grade	Academic Warning	Approaches Standard	Meets Standard	Exceeds Standard	Exemplary
3rd	0-57	58-69	70-84	85-92	93-100
4th	0-53	54-62	63-79	80-88	89-100
5th	0-53	54-61	62-77	78-87	88-100
6th	0-52	53-62	63-78	79-89	90-100
7th	0-43	44-55	56-70	71-83	84-100
8th	0-44	45-57	58-72	73-85	86-100
High School	0-37	38-49	50-67	68-81	82-100

Kansas Assessment of Multiple Measures (KAMM) Mathematics
% correct

Grade	Academic Warning	Approaches Standard	Meets Standard	Exceeds Standard	Exemplary
3rd	0-34	35-56	57-81	82-94	95-100
4th	0-34	35-54	55-81	82-94	95-100
5th	0-34	35-56	57-81	82-91	92-100
6th	0-34	35-54	55-79	80-91	92-100
7th	0-34	35-51	52-74	75-86	87-100
8th	0-34	35-49	50-74	75-84	85-100
High School	0-31	32-44	45-64	65-79	80-100

Kansas Assessment of Multiple Measures (KAMM) Reading
% correct

Grade	Academic Warning	Approaches Standard	Meets Standard	Exceeds Standard	Exemplary
3rd	0-39	40-57	58-88	89-93	94-100
4th	0-38	39-54	55-84	85-90	91-100
5th	0-37	38-51	52-82	83-90	91-100
6th	0-38	39-47	48-75	76-83	84-100
7th	0-34	35-40	41-71	72-82	83-100
8th	0-35	36-46	47-76	77-85	86-100
High School	0-37	38-48	49-81	82-87	88-100

Alternate Assessment Pre-Reading
% correct

Grade	Academic Warning	Approaches Standard	Meets Standard	Exceeds Standard	Exemplary
3rd – 4th	0-39	40-54	55-64	65-75	76-100
5th – 11th	0-39	40-56	57-68	69-80	81-100

Alternate Assessment Reading
Rubric Scores Range 0-5.00

Grade	Academic Warning	Approaches Standard	Meets Standard	Exceeds Standard	Exemplary
All	0 – 2.99	3.00 – 3.74	3.75 - 4.24	4.25 – 4.79	4.80 – 5.0

Alternate Assessment Mathematics
Rubric Scores Range 0-5.00

Grade	Academic Warning	Approaches Standard	Meets Standard	Exceeds Standard	Exemplary
All	0 – 2.99	3.00 – 3.74	3.75 – 4.24	4.25 – 4. 79	4.80 – 5.0

Section 6

RELIABILITY ANALYSES

Score Reliability

Information on the reliability of test scores for each general assessment test form was provided in Section 4, Tables 4.3 for Reading test form scores and Tables 4.4 for Mathematics test form scores. The score reliability estimates reported in the tables are Cronbach alpha coefficients. The coefficient values range from a low of .88 to a high of .94 across all the Reading grade level forms and from .91 to .95 across all the Mathematics grade level forms. The overall general standard errors of measurement on the percent correct score scale range from 3.65 to 4.70 for scores on the Reading general assessment test forms and from 3.95 to 4.60 for scores on the Mathematics general assessment test forms.

Score reliability information on the KAMM assessment test forms is provided in Appendix A. Reliability information for scores resulting from the KAA is provided in Appendix B.

Performance Classification Reliability

In addition to percent correct scores, test scores from the 2006 Kansas Assessments in Reading and Mathematics were used to classify students into one of five performance categories: *Academic Warning*, *Approaches Standard*, *Meets Standard*, *Exceeds Standard*, and *Exemplary*. For each subject at a given grade level, cutscores that were used to identify performance categories were determined via standard setting procedure, as discussed in the previous section for standard setting (see Section 5 of this report).

Reliability analyses on the consistency and accuracy of performance classifications provide important information toward properly understanding and interpreting student performance categories. This is similar to the role of score reliability in helping practitioners interpret test results by assisting in the understanding of the consistency of student performance. As stated in standard 2.15 in the Standards for Educational and Psychological Testing (AERA/APA/NCME, 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).

There are two important indices used in reliability analysis for classification: classification consistency and classification accuracy. Classification consistency refers to the extent to which the classifications agree on the basis of two independent administrations of the test (or, two parallel forms of the test). Classification accuracy refers to the extent to which the actual classifications that are based on observed cut scores approximate those that are based on “true” cut scores. Because repeated measurements are not easily or regularly obtained in a practical assessment program, it has been customary to estimate such reliability indices on classification based on test scores obtained from a single test administration by imposing a psychometric model on such scores.

Procedure

Sample

In 2006, there were on average 4-5 test forms per grade across the content areas of Reading and Mathematics, which lead to smaller sample sizes for each form than in previous years (fewer forms in operational use prior to 2006). As grade-level scores from all test forms in the two content areas were equated to the percent correct scale of the base form and the same set of cut scores were used (within a grade-level content area) to classify the performance of all students, reliability analyses for classification was conducted for the base form only. Table 6.2 below presents the test form number and sample sizes for the base test form across subject and grade.

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. Classification consistency can subsequently 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.

Results

Sample output from the reliability analyses for classification is presented in Table 6.1 below. The first part of the output consists of two data matrices, one for classification consistency and the other for classification accuracy. The number in each cell of the matrix represents the joint probability of falling in the particular category in the given row and column for a randomly sampled student from the testing population. For the consistency matrix, performance categories for both rows and columns are based on observed test cut scores. Classification consistency is simply the sum of the joint probabilities in the diagonal cells of this matrix. For the classification accuracy matrix, performance categories (column) are based on observed test cut scores, while row performance categories are based on “true” cut scores. The classification accuracy is calculated as the sum of the joint probabilities in the diagonal cells of this matrix. It should be noted that the consistency matrix is symmetric whereas the accuracy matrix is not.

The third matrix from the output gives various indices of classification reliability, specifically at the overall level or conditioned on a particular performance category. Four indices, namely classification accuracy, classification consistency, percent agreement by chance, and Cohen's kappa, were given in each of the four columns, respectively. Cohen's kappa is indeed a rescaled version of classification consistency, corrected by the percent agreement by chance. Also presented in the output is the reliability of being classified into a particular performance category or below versus above it, as shown in the last data matrix of the output. Besides for the four indices given in the third matrix, two errors of classification (false positive and false negative) were also calculated for each comparison.

Table 6.2 presents a summary of the classification consistency and classification accuracy indices for the base form across testing grade and content area. For Mathematics, classification consistency values range from .59 (grade 5) to .72 (grade 10). Classification accuracy, on the other hand, is consistently higher than classification consistency yet maintains the same pattern of variability across grades. Classification accuracy coefficients range from .69 (grade 5) to .80 (grade 10). For Reading, classification consistency values range from .54 (grade 4) to .74 (grade 11). The same pattern between classification and accuracy were found for Reading. It is interesting to note that both reliability indices increase as grade level increases, which suggests that performance classification is more reliable for higher grades, likely due to increased test length.

Among the performance categories that students were classified into, three of the classifications are of particular interest. The classification of being placed in the category of *Academic Warning* versus above, which will indicate whether a student is at risk relative to learning expectations; the classification of falling in or below the category of *Approaches Standard* versus above, which will signify whether a student demonstrates acceptable performance relative to AYP criteria; and the classification of being placed in the category of *Exemplary* versus below, which will indicate whether a student's performance is at the highest level. Therefore, classification reliability at those cut points is important as they have implications for the state, schools, and students.

Tables 6.3 and 6.4 below present the relevant information for both Mathematics and Reading. In both tables, in addition to classification accuracy and consistency, probabilities of misclassifications occurring at a given cut point are also provided. While a false positive indicates the probability of misclassifying a student into a given category they are not in, a false negative refers to the probability of not classifying a student into their true category. For both Mathematics and Reading, the reliabilities of classification at a given cut point are generally high whereas probabilities of misclassifications are low. For math, all but one classification accuracy coefficient is over .90. The majority of values of misclassification probability are below .04. At lower grades of Mathematics, a trend of decreasing classification reliability values was observed. For upper grades, consistent high reliabilities were observed across classification categories. For Reading, a similar decreasing trend was observed across the assessed grade levels; however, the lowest classification accuracy coefficient had a value of .87 (.81 for classification consistency).

Summary

Reliability analyses for performance category classification were conducted for the base form of each grade level in both content areas utilizing test scores from the base form with the four-parameter compound binomial true score model. Results showed that classification reliabilities were acceptable. For both Mathematics and Reading, reliabilities of classification at a given cut point are generally high whereas probabilities of misclassifications are low.

Table 6.1. Sample Output for Classification Reliability Analyses

Misclassification Analysis

Grade 3
Content Area: Math
Test Form: f405k

Consistency Matrix

	Level 1	Level 2	Level 3	Level 4	Level 5
Level 1	0.04165	0.01352	0.00205	0.00002	0.00000
Level 2	0.01352	0.03683	0.02612	0.00174	0.00004
Level 3	0.00205	0.02612	0.09928	0.04546	0.00610
Level 4	0.00002	0.00174	0.04546	0.10449	0.06667
Level 5	0.00000	0.00004	0.00610	0.06667	0.39429

Accuracy Matrix

	Level 1	Level 2	Level 3	Level 4	Level 5
Level 1	0.04259	0.00714	0.00018	0.00000	0.00000
Level 2	0.01418	0.04745	0.01785	0.00016	0.00000
Level 3	0.00048	0.02351	0.12287	0.03607	0.00134
Level 4	0.00000	0.00015	0.03758	0.14118	0.06281
Level 5	0.00000	0.00000	0.00053	0.04098	0.40295

Overall Indices & Conditional on Level

	Acc	Con	PC	K
Overall	0.76	0.68	0.31	0.53
Level 1	0.85	0.73	0.00	0.73
Level 2	0.60	0.47	0.01	0.47
Level 3	0.67	0.55	0.03	0.54
Level 4	0.58	0.48	0.05	0.45
Level 5	0.91	0.84	0.22	0.80

Indices by Cut Point

	Acc	FP	FN	Con	PC	K
1 / 2345	0.98	0.01	0.01	0.97	0.89	0.71
12 / 345	0.96	0.02	0.02	0.94	0.77	0.74
123 / 45	0.92	0.04	0.04	0.89	0.57	0.75
1234 / 5	0.89	0.06	0.04	0.85	0.50	0.71

Table 6.2. Classification Indices for Base Form across Grade and Subjects

Subject	Grade	Base Form	Sample Size	Classification Consistency	Classification Accuracy
Mathematics					
	3	405	3949	0.68	0.76
	4	595	4502	0.60	0.70
	5	406	4499	0.59	0.69
	6	479	4691	0.67	0.76
	7	596	4941	0.64	0.74
	8	586	5185	0.66	0.75
	10	590	4966	0.72	0.80
Reading					
	3	386	4479	0.54	0.64
	4	404	5142	0.59	0.69
	5	388	7129	0.63	0.73
	6	401	5708	0.61	0.71
	7	410	5902	0.68	0.77
	8	577	8410	0.69	0.78
	11	582	5699	0.74	0.81

Table 6.3. Classification Indices by Cut Points in Mathematics

Grade	Cut Point*	Classification Accuracy	Classification Consistency	False Positive	False Negative
3					
	1 / 2345	0.98	0.97	0.01	0.01
	12 / 345	0.96	0.94	0.02	0.02
	123 / 45	0.92	0.89	0.04	0.04
	1234 / 5	0.89	0.85	0.06	0.04
4					
	1 / 2345	0.96	0.94	0.01	0.03
	12 / 345	0.93	0.90	0.03	0.04
	123 / 45	0.90	0.87	0.05	0.04
	1234 / 5	0.90	0.86	0.06	0.04
5					
	1 / 2345	0.95	0.93	0.02	0.03
	12 / 345	0.92	0.89	0.03	0.05
	123 / 45	0.90	0.86	0.04	0.05
	1234 / 5	0.91	0.88	0.06	0.03
6					
	1 / 2345	0.96	0.95	0.01	0.02
	12 / 345	0.94	0.92	0.03	0.03
	123 / 45	0.92	0.89	0.04	0.04
	1234 / 5	0.93	0.90	0.04	0.04
7					
	1 / 2345	0.94	0.92	0.03	0.03
	12 / 345	0.92	0.89	0.04	0.04
	123 / 45	0.92	0.89	0.04	0.04
	1234 / 5	0.95	0.92	0.04	0.02
8					
	1 / 2345	0.94	0.92	0.02	0.03
	12 / 345	0.93	0.90	0.04	0.03
	123 / 45	0.93	0.90	0.04	0.03
	1234 / 5	0.95	0.93	0.03	0.02
10					
	1 / 2345	0.93	0.91	0.04	0.03
	12 / 345	0.94	0.91	0.03	0.03
	123 / 45	0.96	0.94	0.03	0.02
	1234 / 5	0.97	0.96	0.02	0.01

* 1 = Academic Warning, 2 = Approaches Standard, 3 = Meets Standard, 4 = Exceeds Standard, 5 =Exemplary

Table 6.4. Classification Indices by Cut Points in Reading

Grade	Cut Point*	Classification Accuracy	Classification Consistency	False Positive	False Negative
3					
	1 / 2345	0.96	0.94	0.01	0.03
	12 / 345	0.92	0.89	0.03	0.05
	123 / 45	0.96	0.81	0.06	0.08
	1234 / 5	0.90	0.87	0.09	0.01
4					
	1 / 2345	0.98	0.97	0.01	0.01
	12 / 345	0.96	0.94	0.02	0.03
	123 / 45	0.90	0.86	0.05	0.05
	1234 / 5	0.86	0.81	0.07	0.07
5					
	1 / 2345	0.97	0.96	0.01	0.02
	12 / 345	0.95	0.93	0.02	0.03
	123 / 45	0.91	0.87	0.04	0.05
	1234 / 5	0.89	0.85	0.06	0.04
6					
	1 / 2345	0.97	0.96	0.01	0.02
	12 / 345	0.95	0.93	0.02	0.03
	123 / 45	0.91	0.87	0.04	0.06
	1234 / 5	0.88	0.84	0.07	0.05
7					
	1 / 2345	0.98	0.98	0.01	0.01
	12 / 345	0.96	0.95	0.01	0.02
	123 / 45	0.93	0.90	0.03	0.04
	1234 / 5	0.89	0.85	0.05	0.06
8					
	1 / 2345	0.98	0.97	0.01	0.01
	12 / 345	0.96	0.95	0.02	0.02
	123 / 45	0.93	0.90	0.03	0.04
	1234 / 5	0.9	0.87	0.06	0.04
10					
	1 / 2345	0.99	0.99	0.00	0.01
	12 / 345	0.98	0.97	0.01	0.01
	123 / 45	0.95	0.93	0.02	0.03
	1234 / 5	0.90	0.85	0.04	0.06

* 1 = Academic Warning, 2 = Approaches Standard, 3 = Meets Standard, 4 = Exceeds Standard, 5 =Exemplary

Section 7

EVIDENCE FOR THE VALIDITY OF INFERENCES FROM TEST SCORES

Part 1: INTERNAL EVIDENCE FOR THE VALIDITY OF KANSAS GENERAL ASSESSMENT SCORES

Item Intercorrelations and Factor Structure

The analysis of test structure based on examinee response data is an important part of the test development and evaluation. Results from such analyses provide empirical evidence for identifying the targeted domain being measured, understanding the internal structure of the domain, and justifying test score interpretations, as well as singling out construct-irrelevant sources of variance in test scores.

Evaluating test structure involves both verification of the hypothesized unidimensionality/multidimensionality of the domain and identifying the nature of the underlying construct(s) measured by the items. There are two different definitions of test dimensionality in the literature. One is the strict dimensionality of the test (McDonald, 1981), which refers to the minimum number of latent abilities required to produce monotonicity and strict local independence. The other is the essential dimensionality of the test (Stout, 1987), which refers to the minimum dimensionality required to produce monotonicity and essential local independence. Compared to strict dimensionality, the definition of essential dimensionality states that relatively fewer dominant dimensions are required to produce monotonicity and local independence. In practice, it is common that the number of test dimensions is high in its strict sense, though relatively fewer dimensions are dominant and all other dimensions are minor.

Understanding a particular measurement construct through test structure assessment can be done either through post-hoc analyses of the pattern of item loading and the characteristics of the set of items that load on a given dimension, or through confirmation of the hypothetical structural complexity of the test. For example, if reading comprehension items are expected to be dependent on both general reading ability and the particular reading passage, then a bi-factor model is expected to show adequate fit to the data.

Common approaches to evaluating test structure include exploratory and/or confirmatory factor analyses (Muthen & Muthen, 1998), full-information item factor analysis (Bock, Gibbons, & Muraki, 1988) and nonlinear factor analysis (McDonald, 1967). It is well established in the literature that the tetrachoric correlation, rather than the product moment correlation, should be used for dimensionality assessment when analyzing dichotomous data in order to prevent the presence of a spurious difficulty factor (Hulin, Drasgow, & Parsons, 1983). Compared to other approaches of factor analysis, the full-information item factor analysis maximizes the likelihood function of the factor model that is based on the observed item response patterns and thus utilizes all of the information in the item responses. Both full-information item factor analysis and nonlinear factor analysis are based on multidimensional Item Response Theory (IRT) models and also have the advantage of incorporating into the analysis chance success through guessing with multiple-choice items, which is common in large scale assessment.

Procedure

Sample

The evaluation of test structure was based on samples of students who were administered the base form of the Kansas general assessments in a given subject via the computer mode. In 2006, parallel test forms of the Kansas general assessment were constructed at grades 3, 4, 5, 6, 7, and 8 for both Mathematics and Reading and at grade 10 for Mathematics and grade 11 for Reading. There were five test forms per grade for Mathematics and four test forms per grade for Reading (with the exception of grades 5 and 8 where only three test forms were available). The first four columns of Table 7.1 below present the base form number at each grade/subject, the number of items in the base form, and the sample sizes used in the analyses.

Method

The computer program TESTFACT 4.0 (Wood, Wilson, Gibbons, Schilling, Muraki, & Bock, 2002) was used to conduct the evaluation of test structure. This program calculates the sample tetrachoric correlation matrix and provides corrected values to item pairs that lack sufficient sample size to estimate the tetrachoric correlation. It also conducts the full-information item factor analysis for dichotomous data using maximum marginal likelihood estimation method with the Expectation-Maximization algorithm. For a confirmatory analysis of test structure, a bi-factor model can be specified in the program.

As guessing parameters need to be specified in the TESTFACT program in order to incorporate the guessing effect in the analyses, guessing parameters were first estimated by running a 3-PL IRT model using the BILOG-MG program. Those parameters were then imported into the syntax file of TESTFACT.

A combined approach was taken to set the criteria for determining the dimensionality of a test. Those criteria included: (a) an inspection of the scree plot of the first 10 largest eigenvalues of the correlation matrix and (b) investigating model fit improvement due to adding an additional factor to the model. To inspect the scree plot, both the absolute magnitudes of the eigenvalue and the ratio between successive factors were taken into account. A test was judged to be unidimensional if the first eigenvalue was at least five times as large as the second one and the subsequent eigenvalues were all similar in values. Using the Maximum Marginal Likelihood method, TESTFACT provides a chi-square statistical test for assessing the relative model fit improvement. To avoid a sample size effect, an additional descriptive measure of the model fit, the root mean square residual (RMSR), was also calculated based on the residual matrix (Joreskog & Sorbom, 1996; Muthen & Muthen, 1998; Steiger, 1990). Specifically, a 10% or greater reduction of the RMSR over the preceding model was judged as adequate for adding an additional factor into the model (Tate, 2003). When different criteria led to different conclusions, the chi-square statistical test was given the least attention since it is sensitive to sample size. The criterion based on analyzing the scree plot has higher priority over the RMSR criterion because the latter is more vulnerable to outliers of item-pair correlations.

Results

Inter-Item Correlations

Inter-item correlations were calculated for each form at each grade in both Mathematics and Reading. The tetrachric correlation, instead of the Pearson product-moment correlation, was used for the reasons discussed previously.

For space consideration, only descriptive statistics for inter-item correlations among items are presented in this report. The second part of Table 7.1 displays such information for each form in Mathematics and Reading, respectively. While the means and standard deviations of the inter-item correlations are similar in value, different distributional shapes of the inter-item correlations, reflected in the values of skewness and kurtosis, are observed across test forms, both within and between the subjects (Mathematics and Reading). Because different test specifications were used at different grade levels and subjects, direct comparison of such results is not tenable. In Mathematics, the average inter-item correlation ranges from .311 to .451, with an increasing trend toward higher grades. The distributions of inter-item correlations all have positive kurtosis. Figure 7.1 below depicts the distribution of inter-item correlations for Form 405 at grade 5, which has the largest value of kurtosis. As shown in the figure, the distribution is slightly positive skewed, with the majority of inter-item correlations falling between .2 and .6.

In Reading, the average inter-item correlations range from .336 to .385. There is no clear increasing or decreasing trend across grades. Similar to those of Mathematics, most of the inter-item correlation distributions show negative skewness, but smaller values in kurtosis. Figure 7.2 shows the distribution of inter-item correlations of Form 388 at grade 5. The distribution is similar to a normal distribution, with the majority of inter-item correlations falling between .2 and .6.

Table 7.1. Test Length, Sample Size, and Descriptive Statistics of Inter-Item Correlations for Each Form in Mathematics and Reading

Subject/ Grade	Form	Number of Items	Sample Size	Mean	Standard Deviation	Skew	Kurtosis	Min	Max
Math									
3	405	70	3894	0.342	0.105	0.295	1.352	-0.008	0.879
4	595	73	4448	0.335	0.112	-0.368	1.072	-0.127	0.790
5	406	73	4446	0.311	0.102	0.910	3.980	-0.093	0.923
6	479	86	4624	0.383	0.124	-0.175	0.660	-0.036	0.909
7	596	84	4844	0.393	0.122	-0.198	1.233	-0.119	0.909
8	586	85	5076	0.413	0.105	0.160	0.443	0.070	0.860
10	590	84	4862	0.451	0.119	-0.274	0.881	-0.175	0.947
Reading									
3	386	58	4476	0.340	0.133	-0.598	1.082	-0.138	0.753
4	404	74	5138	0.336	0.113	-0.033	0.122	-0.110	0.750
5	388	74	7127	0.363	0.110	0.082	-0.500	0.040	0.830
6	401	79	5699	0.345	0.141	-0.117	-0.182	-0.040	0.890
7	410	84	5891	0.351	0.140	-0.136	0.006	-0.100	0.780
8	577	83	8399	0.385	0.117	-0.046	0.412	-0.020	0.900
11	582	81	5682	0.355	0.127	0.074	-0.468	0.02	0.750

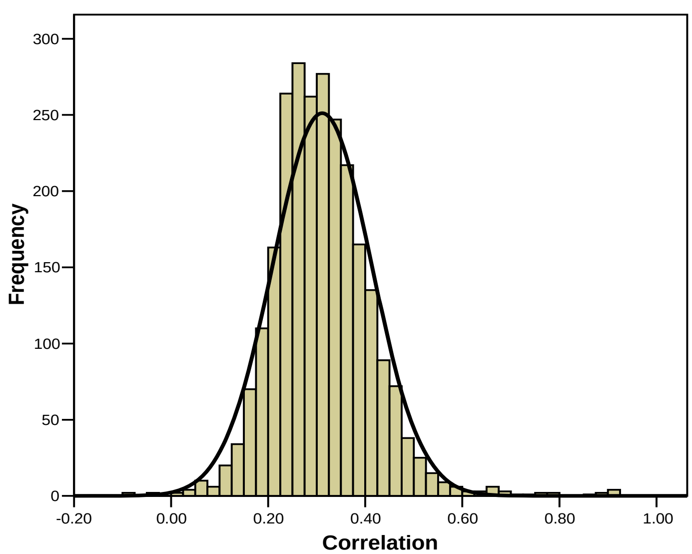


Figure 7.1. Distribution of Inter-item Correlations for Test Form 406 at Grade 5 of Mathematics

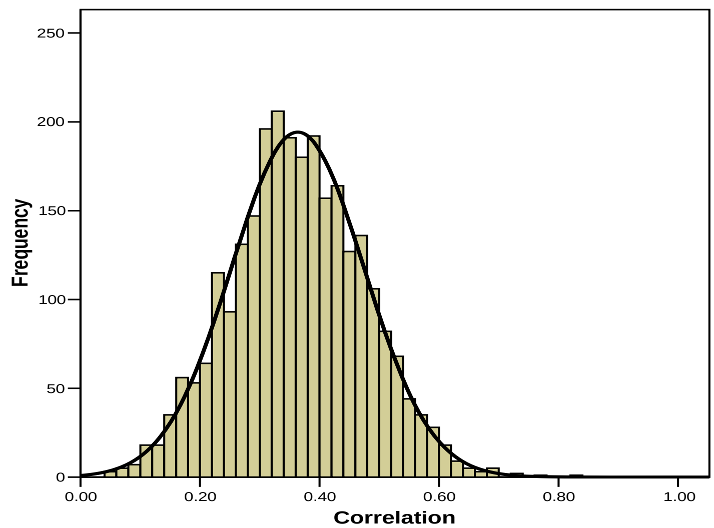


Figure 7.2. Distribution of Inter-item Correlations for Test Form 388 at Grade 5 of Reading

Test Dimensionality/Structures

All test forms in both Mathematics and Reading show strong evidence of essential unidimensionality. The sample outputs for Form 405 at grade 3 of Mathematics are shown in Figure 7.3 and Table 7.2, respectively. Figure 7.3 depicts the scree plot of the first 10 largest eigenvalues. The scree plot clearly shows that the test is essentially unidimensional in that the first eigenvalue is more than 20 times that of the second and any subsequent ones are similar in values. Table 7.2 presents two different sets of model fitting information. One is the chi-square test statistic of model improvement, which provides a statistical-testing approach to model selection. The other is the RMSR. Although the chi-square test statistics show that the two-factor models fit better (statistically) than the one-dimensional model, the improvement of RMSR is not substantial (less than 10%). According to the criteria specified in the procedure section above, the conclusion of unidimensionality is still tenable.

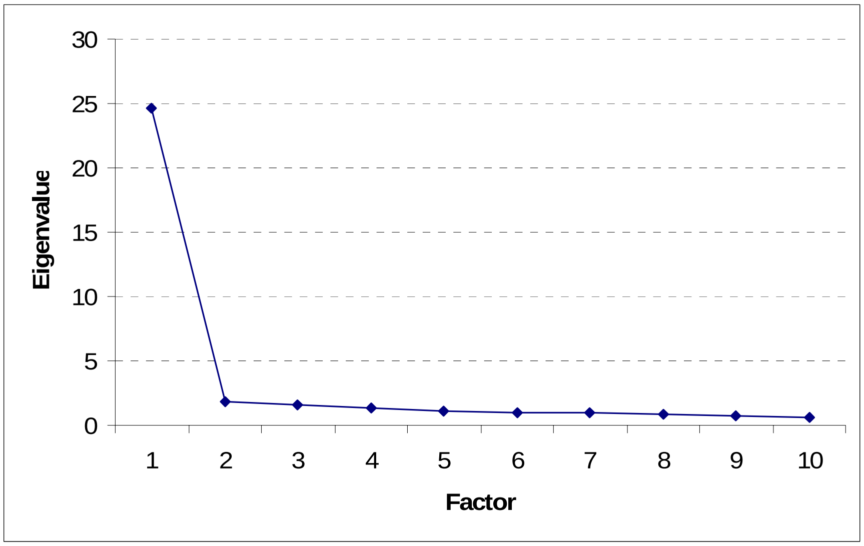


Figure7. 3. Scree Plot for Test Form 405 at Grade 3 of Mathematics

Table 7.2. Chi-square Test Statistics and RMSR for Models Fitted to Test Form 405 at Grade 3 of Mathematics

Statistics	Number of Factors	
	1	2
χ^2	140422.95	139213.3
df	3753	3684
$\Delta\chi2$	n/a	1209.67
Δdf	n/a	69
P	n/a	0.0000
RMSR	0.0785	0.0769
% Reduced in RMSR	n/a	2.04

Table 7.3 presents the model fit information for all test forms in Mathematics. The chi-square test of model fit improvement suggests that all test forms are multidimensional. However, according to the 10% reduction in RMSR rule, this is true only for Form 406 at grade 5. In Figure 7.4, the scree plot for Form 406 at grade 5 is provided. As shown, the test form is essentially unidimensional as the first eigenvalue (23.0557) is more than 9 times that of the second eigenvalue (2.46207). The rest of the factors are essentially trivial as they explain very little variation in the data (that is, the largest eigenvalue among the rest of the factors explains about 3% of the total variance). Based on those criteria, the test form for grade 5 Mathematics is essentially unidimensional as well.

Table 7.3. Chi-square Statistics and RMSR for Models Fitted to Test Forms of Mathematics

Grade	Test Form	Models / # of Factors	χ^2	df	$\Delta\chi 2$	Δdf [a]	P	RMSR	% Reduced in RMSR
3	405	One	140423.0	3753	n/a	n/a	n/a	0.079	n/a
		Two	139213.3	3684	1209.67	69	0.000	0.077	2.04
4	595	One	193859.3	4303	n/a	n/a	n/a	0.064	n/a
		Two	193368.4	4232	490.93	71	0.000	0.065	-1.56 [b]
5	406	One	210425.9	4301	n/a	n/a	n/a	0.072	n/a
		Two	208175.4	4230	2250.52	71	0.000	0.063	11.89
6	479	One	262457.6	4453	n/a	n/a	n/a	0.071	n/a
		Two	259684.0	4369	2773.6	84	0.000	0.070	1.40
7	596	One	307440.8	4679	n/a	n/a	n/a	0.070	n/a
		Two	303607.0	4598	3833.8	81	0.000	0.067	3.58
8	586	One	325461.3	4911	n/a	n/a	n/a	0.067	n/a
		Two	323466.1	4830	1995.15	81	0.000	0.064	3.30
10	590	One	352770.2	4693	n/a	n/a	n/a	0.083	n/a
		Two	350532.6	4610	2237.6	83	0.000	0.078	5.43

Note: a. changes of the degree of freedom, *Δdf*, in some forms are not consistent with the number of items in the form because some items were removed in order to get the model to converge correctly.

b. A prior distribution for item parameters was employed when fitting a two-factor model in order to get the model to converge properly.

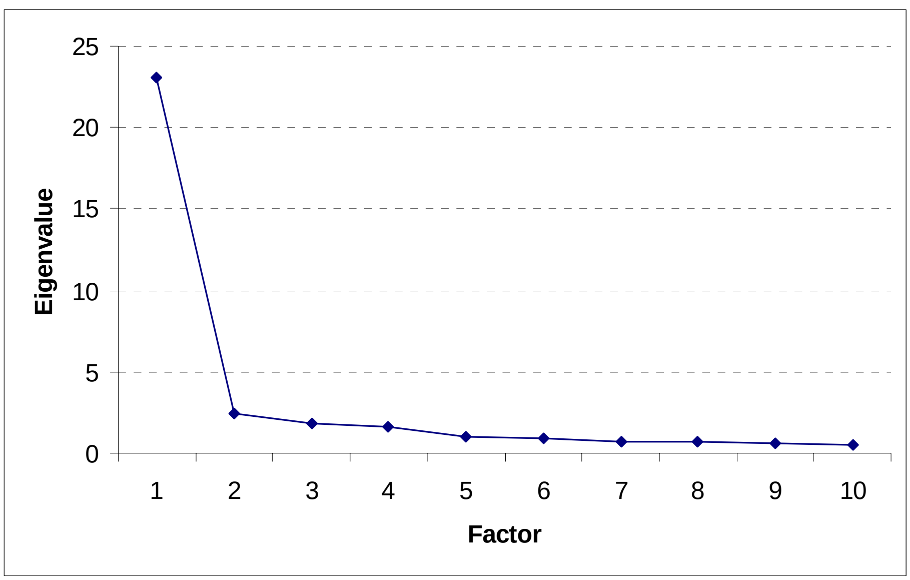
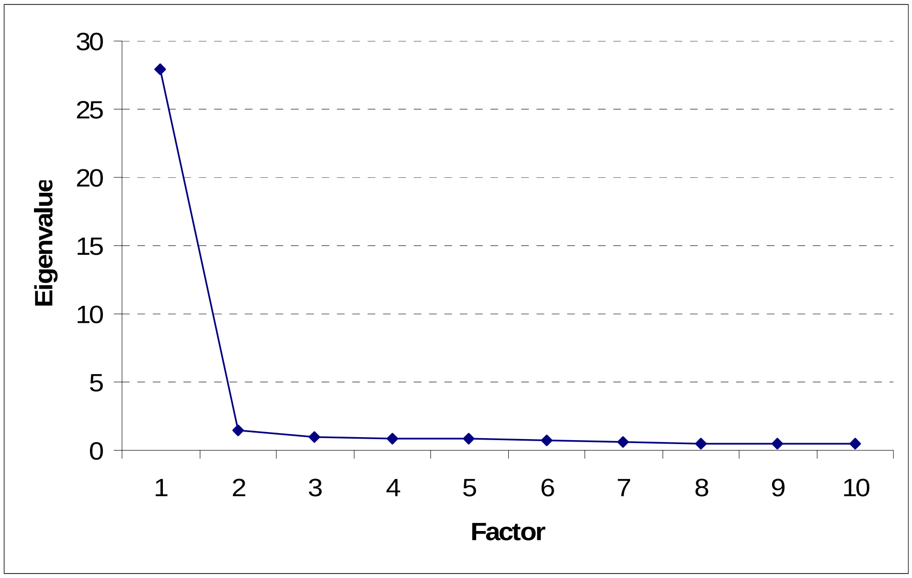


Figure 7.4. Scree Plot for Test Form 406 at Grade 5 of Mathematics

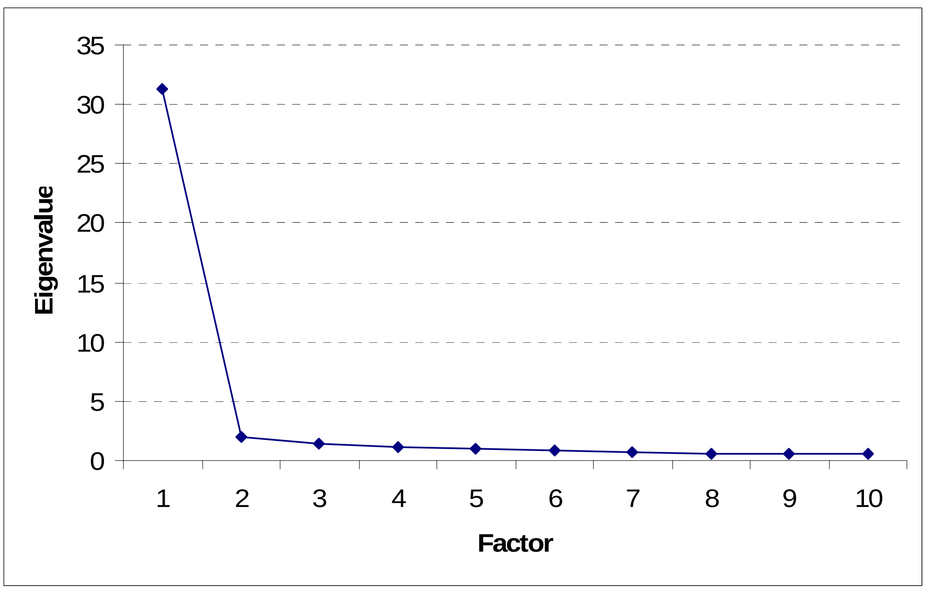
For Reading, Table 7.4 presents the corresponding model fit information for the test forms across all testing grades. According to the RMSR criteria, it seems that a second factor needs to be added for Forms 388 and 410, respectively. However, the scree plots for both test forms, as shown by Figure 7.5, clearly suggest that the test forms are essentially unidimensional.

Table 7.4. Chi-square Statistics and RMSR for Models Fitted to Test Forms of Reading

Grade	*Test Form*	*Models / # of Factors*	χ^2	*df*	$\Delta\chi 2$	Δdf	*P*	*RMSR*	*% Reduced in RMSR*
3	386	One	129973.8	4361	n/a	n/a	n/a	0.068	n/a
		Two	129377.9	4305	595.89	56	0.0000	0.064	5.76
4	404	One	220851.6	4989	n/a	n/a	n/a	0.077	n/a
		Two	220056.5	4916	795.16	73	0.0000	0.073	4.92
5	388	One	305007.9	6978	n/a	n/a	n/a	0.059	n/a
		Two	303917.3	6905	1090.68	73	0.0000	0.052	11.62
6	401	One	287486.0	5542	n/a	n/a	n/a	0.071	n/a
		Two	285891.1	5465	1594.87	77	0.0000	0.074	-3.52
7	410	One	333319.5	5722	n/a	n/a	n/a	0.063	n/a
		Two	331725.1	5639	1594.37	83	0.0000	0.055	12.62
8	577	One	454030.4	8232	n/a	n/a	n/a	0.070	n/a
		Two	451904.0	8150	2126.38	82	0.0000	0.066	6.83
11	582	One	292843.3	5519	n/a	n/a	n/a	0.072	n/a
		Two	291268.1	5439	1575.18	80	0.0000	0.065	8.81



Form 388



Form 410

Figure 7.5. Scree Plots for Test Form 388 (Upper Panel) and Test Form 410 (Lower Panel) in Reading

Part 2: CRITERION-RELATED EVIDENCE FOR THE VALIDITY OF KANSAS GENERAL ASSESSMENT SCORES

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 multidimensional 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 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, 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.

Three analyses documenting the relationship of Kansas Assessment scores to relevant variables external to the test were performed to provide sources of criterion-related validity evidence. Results from these studies are detailed below.

Study 1: Predictive validity study between formative and general assessments

The Kansas Assessment system includes a customized formative testing component that is designed to provide feedback regarding whether a student has mastered particular content standards during the course of instruction. For the content area of mathematics (grades 3-8 and 10), each assessed indicator (range of 12-15 indicators per grade level) at a grade level is featured by one standard-specific testlet that ranges from 4 to 13 items, as well as a longer, comprehensive formative assessment. For Reading (grades 3-8 and 11), testlets at each grade level are arranged by passage-type (Narrative, Expository, Technical, and Persuasive) and range from 11 to 23 items. The formative assessment feature of the Kansas Assessment program is delivered via computer, with score reports generated instantly upon student completion of an indicator (Math) or passage-type (Reading) testlet.

The fully-customized formative assessment program was introduced in Kansas prior to the spring testing window during the 2005-06 academic year. Students sitting for a formative assessment were logged into the system using a test session ticket based on their unique state identification number, thus affording the ability to link a student's formative assessment performance to that student's respective general content area assessment (summative) result.

To evaluate the impact of Kansas' formative assessment system within the context of the statewide testing program-specifically the relationship between the components of the formative assessment system and performance on the summative statewide assessments, individual student data were matched on the basis of formative and summative assessment results.

For formative Mathematics, a total score was calculated for students that completed every standard-specific testlet at the respective grade levels. Correlations between the formative aggregate and the General Assessment equated total scores were obtained for each grade level. Table 7.5 below details the relationship between formative and summative assessment performance for Mathematics, overall and by specific General Assessment test form.

Table 7.5. Formative assessment correlated with General. ALL forms, then split by forms.

Math 2006

Gr	All Forms	P&P		A (Computer)		B		C		D		E	
	r (n)	*r*	*n*	*r*	*n*	*r*	*n*	*r*	*n*	*r*	*n*	*r*	*n*
3	.71 (2358)	.70	284	.67	432	.72	396	.75	424	.71	416	.73	406
4	.78 (1405)	.74	206	.78	235	.73	244	.84	250	.79	234	.78	236
5	.76 (1362)	.54	129	.81	252	.76	244	.76	247	.81	242	.80	248
6	.77 (1262)	.50	44	.78	231	.76	233	.75	250	.77	244	.85	260
7	.81 (1175)	.90	47	.83	222	.79	227	.83	223	.77	227	.81	229
8	.87 (1337)	.83	8	.86	261	.88	269	.89	269	.89	260	.85	270
10	.82 (830)	.83	15	.86	166	.86	158	.82	164	.76	156	.82	171

The values of the coefficients across forms are high, offering evidence that the predictive utility of the customized formative assessments is strong. The correlations between formative assessment performance and summative assessment performance for all forms range from .71 to .87. By General Assessment form, within each assessed grade level, coefficients range from .67-.75 (grade 3), .73-.84 (grade 4), .54-.81 (grade 5), .50-.85 (grade 6), .77-.90 (grade 7), .83-.89 (grade 8), and .76-.86 (grade 10). As accountability systems depend on the use of information to improve student performance, the strong relationship observed between the customized formative testing component of the state assessment program and General Assessment performance is encouraging.

Also observed from the Table is that the number of students completing all formative testlets available tends to decrease as grade level increases. Descriptive statistics reveal, however, that the Mathematics assessments increase in difficulty as grade level increases. Another trend seen from the results is that students who sit for the General Math Assessment via the paper-and-pencil testing mode participate significantly less in the formative program in terms of exhausting the full set of indicator testlets.

For formative Reading, a total score was calculated for students that completed every text-type testlet at assessed grade levels. Correlations between the formative aggregate and the General Assessment equated total score were obtained for each grade level. Table 7.6 below details the relationship between formative and summative assessment performance for Reading, overall and by specific General Assessment test form.

Table 7.6. Formative assessment correlated with General. ALL forms, then split by forms.

Reading 2006

Gr	All Forms	P&P		A (Computer)		B		C		D	
	r (n)	*r*	*n*	*r*	*n*	*r*	*n*	*r*	*n*	*r*	*n*
3	.76 (2328)	.78	324	.76	494	.76	515	.76	501	.74	494
4	.76 (1730)	.81	221	.73	371	.75	359	.78	395	.74	384
5	.78 (859)	.72	129	.81	1418	.84	137	.77	250	.73	105
6	.80 (792)	.85	17	.76	189	.81	185	.79	215	.81	186
7	.81 (587)	.85	17	.81	144	.74	136	.84	149	.84	141
8	.82 (829)	.87	12	.83	263	.83	284	.79	270		
11	.83 (535)	.74	33	.85	118	.81	127	.88	126	.82	131

Similar patterns emerge from these data as with the Mathematics data, namely that the predictive validity of the formative assessments is strong. Correlations between formative assessment performance and summative assessment performance for all forms range from .76 to .83. By General Assessment form, within each assessed grade level, coefficients range from .74-.78 (grade 3), .73-.81 (grade 4), .72-.84 (grade 5), .76-.85 (grade 6), .74-.85 (grade 7), .79-.87 (grade 8), and .74-.88 (grade 10).

Trends similar to those observed with the Mathematics data appear in the Reading data as well. There is a substantial drop-off as grade level increases in the number of students that completed all formative Reading testlets. Also, the number of students that participate in the formative system that sit for the General Assessment via paper-and-pencil is considerably low.

Study 2: Relationship of test scores across years given mode of assessment

The Kansas Assessment program is a dual-mode system, with the Kansas Computerized Assessment (KCA) application having been an operational feature of the program since the spring 2003 administration. Advantages of the KCA system include immediate score reporting on student performance, the cost reduction related to printing and shipping, improvements in test security and score accuracy, continuous testing of students during the testing window, and the offering of assessment strategies designed to support instruction (formative testing). The number of students using online KCA testing has increased every year as the word has spread that it is educationally sound, instructionally supportive, and benefits Kansas. In 2005, 58% of Kansas students who were required to be tested were assessed on computers. In 2006, approximately 60% of Kansas students sitting for a General form of the assessments were tested via computer.

An analysis looking at the relationship of individual student test scores across years was conducted to provide another source of criterion-related validity evidence for the Kansas Assessments, specifically investigating differences in test performance across groups/settings in response to interventions or potential changes in task requirements. Through 2005, testing occurred annually in Kansas for students in grades 4, 7, and 10 (Math) and grades 5, 8, and 11 (Reading). Beginning in 2006, new testing mandates called for students in grades 3-8 and 10 (Math) and 11 (Reading) be tested yearly. For this study, performance data for individual students were matched across two years, 2005 and 2006, for students in the following configurations:

- 4th grade students in 2005 who were 5th graders in 2006 in Mathematics
- 7th grade students in 2005 who were 8th graders in 2006 in Mathematics
- 5th grade students in 2005 who were 6th graders in 2006 in Reading

For eligible students, information regarding the mode in which they were assessed in both years was also available, making it possible to examine the relationship between scores within testing mode or between testing mode across years and with the students serving as their own control in the matching design. Table 7.7 below details the number of students available for this study and the resulting coefficients obtained.

Table 7.7. 2005-2006 matched group design results

	Grade 4-5 Math		Grade 7-8 Math		Grade 5-6 Reading	
within mode	**n**	***r****	**n**	***r****	**n**	***r****
Paper2005 to Paper 2006	6613	0.71	5450	0.80	6503	0.72
KCA2005 to KCA2006	13298	0.70	16504	0.78	12639	0.71
across mode						
Paper 2005 to KCA2006	3552	0.71	3823	0.78	5272	0.74
KCA2005 to Paper2006	2907	0.75	2694	0.80	2113	0.77

*all values significant at the .01 level

Correlation coefficients appear stable for the matched groups across the conditions, providing a source of evidence that the meaning of the scores is the same for students under these different conditions. It should be noted that the scores being correlated were obtained a year apart, measured somewhat different constructs and instructional interventions occurred that potentially affected student performance differentially. Given the latter sources of variability affecting test scores, the coefficients observed are moderately high and in an expected range.

Study 3: Investigating the relationship between teacher ratings and student test performance

Another piece of criterion-related evidence documented to support the validity of the Kansas Assessments involved an analysis of teacher ratings of student performance, specifically the relationship between teacher ratings of students in terms of the Kansas performance categories and actual performance on the state assessments. The data used for this study were obtained as part of one of the standard setting procedures (Contrasting Groups method) that was implemented in 2006 for the purpose of identifying cutscores for the new tests.

During the spring 2006 testing window, teachers were asked to participate in an activity whereby, calling on their knowledge of their students and relying on their professional judgments, they were asked to place their students into the appropriate category (called *Unsatisfactory, Basic, Proficient, Advanced, Exemplary* at the time) based on the Kansas Performance Level Definitions. Professional judgments were solicited from teachers for students they had taught in the relevant subject area (Reading, Mathematics) during the 2005-06 academic year.

As Kansas offers two modes of delivery for the general assessments, for students that were assessed via computer, teachers were asked to make individual student ratings online if the assessment was taken via computer or make the ratings on their student's answer sheets if the assessment was taken via paper-and-pencil in the appropriate location on the scannable answer form. If answer sheets for paper-and-pencil tested students had already been returned, student ratings could be entered on-line as described for computerized tested students.

Table 7.8 below shows the relationship, by test form taken, between teacher rating of students in Reading and their actual performance on the Kansas Reading assessment. The number in parentheses indicates the total number of teacher ratings that were collected at a grade level for the mode assessed (total ratings combined across computer forms).

Table 7.8. Correlation between teacher ratings and student performance in Reading

Grade	All Forms (N)	Paper&Pencil A (N)	Computer A (N)	B	C	D
3	0.68 (7160)	0.69 (4175)	0.65 (2185)	0.71	0.68	0.62
4	0.62 (4411)	0.44 (1560)	0.72 (2851)	0.74	0.70	0.72
5	0.71 (7086)	0.71 (2924)	0.73 (4162)	0.71	0.68	0.68
6	0.71 (3760)	0.67 (854)	0.72 (2906)	0.73	0.74	0.69
7	0.69 (4027)	0.54 (920)	0.71 (3107)	0.75	0.73	0.75
8	0.70 (3596)	0.68 (368)	0.73 (3228)	0.69	0.68	
11	0.71 (3070)	0.62 (460)	0.73 (2610)	0.74	0.73	0.75
range	.62-.71	.44-.71	.65-.74	.69-.75	.68-.75	.62-.75

The coefficients across grade levels for all forms combined and across grade levels for the computerized forms are quite stable and strong. However, it is likely the case that the ratings of students assessed via the paper and pencil format are likely the more pure measure as far as teacher judgment is concerned as these ratings are less contaminated by feedback on student performance. As one of the benefits of computerized delivery system in Kansas is the immediate reporting of performance, teachers rating students via computer likely had viewed feedback on the student. This would not have been the case with students assessed using paper and pencil, which is the plausible explanation for more variability across grade levels for this group. The median coefficient value across grades of .67, however, indicates a robust relationship between test performance and teacher ratings in Reading.

For Mathematics, Table 7.9 below shows the relationship, by test form taken, between teacher rating of students in Math and their actual performance on the Kansas Mathematics assessment. The number in parentheses indicates the total number of teacher ratings that were collected at a grade level for the mode assessed (total ratings combined across computer forms).

Table 7.9. Correlation between teacher ratings and student performance in Mathematics

Grade	All Forms (N)	Paper&Pencil A (N)	Computer A (N)	B	C	D	E
3	0.66 (4112)	0.61 (1056)	0.67 (3056)	0.71	0.70	0.69	0.64
4	0.73 (7509)	0.70 (2116)	0.73 (5393)	0.74	0.77	0.75	0.74
5	0.64 (4286)	0.44 (1416)	0.74 (2870)	0.77	0.73	0.76	0.76
6	0.72 (5037)	0.62 (1045)	0.72 (3992)	0.74	0.75	0.75	0.76
7	0.78 (3016)	0.67 (362)	0.78 (2654)	0.80	0.82	0.77	0.81
8	0.75 (4765)	0.60 (937)	0.79 (3828)	0.80	0.78	0.75	0.79
10	0.80 (3720)	0.73 (477)	0.79 (3243)	0.82	0.80	0.79	0.83
range	.66-.80	.44-.73	.67-.79	.71-.82	.70-.82	.69-.79	.64-.83

The patterns for Math resemble those for Reading, whereby the ratings made for students assessed via paper and pencil are likely the more pure measure as far as teach judgment. The median correlation coefficient across grade levels for this group is .62.

Teacher judgments of students typically are contaminated with sources of irrelevant construct variance due to non-related information about the student that influences their ratings. While teachers were asked to make judgments of students according to the Kansas performance category definitions, the expectation is that other irrelevant factors would somewhat influence their ratings. Given these circumstances, the degree of relationships observed between those independent teacher ratings and student performance on the Reading and Mathematics tests are sufficiently high to offer a criterion-related source of evidence supporting the validity of scores from the Kansas assessments.

Summary

Validity is an evaluative judgment about the degree to which the test scores can be interpreted to mean what is claimed that they mean. Generally, there are about a half dozen different strategies for obtaining evidence for the validity of test scores (Messick, 1989). Other sections of this report describe the judgment of content in relation to the subject area domains and selection of items that have adequate psychometric characteristics as well as whether examinees' performance within the set of items on the test is consistent. While content representation, item quality, and internal structure are important aspects of tests, they do not ensure the validity of test scores. In order to examine the validity of test scores, it is important to determine the degree to which examinees' performance on a test correlates at expected levels 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 results of these analyses provide evidence to support the validity of 2006 Kansas Assessment scores.

REFERENCES

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

Berk, R.A. (Ed.). (1982). *Handbook of methods for detecting test bias*. Baltimore,MD: The Johns Hopkins University Press.

Bock, R. D., Gibbons, R., & Muraki, E. J. (1988). Full information item factor analysis. *Applied Psychological Measurement, 12*, 261 – 280.

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.

Hulin, C. L., Drasgow, F., & Parsons, C. K. (1983). *Item response theory: Applications to psychological measurement*. Homewood, IL: Dow Jones-Irwin.

Jőreskog, K. G. & Sőrbom, D. (1996). *LISREL 8: User's reference guide*. Chicago, IL: Scientific Software.

Kingston, N.M., Kahl, S.R., Sweeney, K.P., & Bay, L. (2001). The Bookmark procedure: Psychological perspectives. In G. J. Cizek (Ed.), *Setting performance standards* (pp. 249-282). Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Kolen, M.J. & Brennan, R.L. (1995). Test Equating: Methods and Practices. New York, NY: Springer-Verlag.

Livingston, S. A., & Zieky, M. J. (1982). *Passing scores: A manual for setting standards of performance on educational and occupational tests*. Princeton, NJ: Educational Testing Service.

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.

McDonald, R. P. (1967). Nonlinear factor analysis (Psychometric Monographs, No. 15). The Psychometric Society.

McDonald, R. P. (1981). The dimensionality of tests and items. British Journal of Mathematical and Statistical Psychology, 34, 100 – 117.

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher, 18,* 5-11.

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.

Muthén, L. K. & Muthén, B. O. (1998-2001). *Mplus User's Guide*. Los Angeles, CA: Muthén & Muthén.

Poggio, J., Glasnapp, D. R., Yang, X., & Poggio, A. J. (2005). A comparative evaluation of score results from computerized and pager & pencil mathematics testing in a large scale state assessment program. The Journal of Technology, Learning, and Assessment, 3 (6). Available from http://www.jtla.org.

Poggio, J., Glasnapp, D., Yang, X., Beauchamp, A., & Dunham, M. (2005). Moving from paper and pencil to online testing: Findings from a state large scale assessment program. A series of papers presented at the NCME annual meeting, Montreal, April.

Steiger, J. H. (1990). Structural model evaluation and modification: an interval estimation approach. *Multivariate Behavioral Research, 25*, 173 -180.

Stout, W. (1987). A nonparametric approach for assessing latent trait dimensionality. *Psychometrika, 52*, 589 – 617.

Tate, R. (2003). A comparison of selected empirical methods for assessing the structure of responses to test items. *Applied Psychological Measurement, 27*, 159 – 203.

Tittle, C.K. (1982). Use of judgmental methods in item bias studies. In R.A. Berk (Ed.), *Handbook of methods for detecting item bias*. Baltimore, MD: The Johns Hopkins University Press.

Wood, R.L., Wilson, D., Gibbons, R., Schilling, S., Muraki, Eiji, & Bock, D. (2002). *TESTFACT 4.0: test scoring and item factor analysis*. [computer program] Chicago: Scientific Software Inc.

Appendix A

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 IEP team used the KAMM eligibility criteria (available from KSDE) and determined the KAMM/modified assessment is the appropriate assessment for the student may take the KAMM. Less complexity of test items is the basis for the KAMM. KAMM compares to the general assessments in Reading and Mathematics 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 Reading, the number of total items across grade level tests range from 30 (grade 3) to 49 (grade 11) items. At all levels, there are fewer passages to read. At grades 3 and 4 there are two narrative and two expository passages. At grades 5, 6, and 7 there are two narrative, two expository, and one technical passage. At grades 8 and HS there are two narrative, two expository, one technical passage, and one persuasive passage. In Mathematics, all grade level KAMM forms have 40 items and calculator use was allowed for all portions of the assessment. 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.

KSDE was responsible for the development of the KAMM assessments. Items for the KAMM were selected or modified based on cognitive load. For Reading, newly developed passages and items independent of those appearing on general Reading assessment forms were selected. For Mathematics, the Mathematics Revision Committee and the Special Education Assessment Advisory Committee, in cooperation with KSDE was given the task of identifying the indicators to be included on the KAMM. In addition to defining which indicators would appear on the grade level tests, it was necessary to determine the number and kind of items to be included from the general Mathematics Assessment. Four items per grade level indicator to be assessed on KAMM (10 assessed indicators per KAMM Mathematics form) from the general Mathematics assessments were selected and modified.

Accommodations are allowed on the KAMM as IEP teams make decisions 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-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 DIF analyses. Items were reviewed by KSDE panels for bias, offensiveness and insensitivity, however.

KAMM Reading Summary Statistics

Table A.1 reports summary findings for the KAMM Reading Assessments by grade level, identifying the number of items per KAMM form, the number of students administered a particular grade level KAMM, reliability coefficients, and descriptive statistics in terms of raw total scores and percent correct total scores. All of the reliability coefficients are greater than 0.85. The percent mean correct ranged from 57.92% (SD = 17.398) at grade 7 to 76.73% (SD = 18.147) at grade 3.

Table A.1. KAMM Reading 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
3	745	30	731	0.86	23.02	5.44	76.73	18.15
4	747	37	882	0.88	26.48	6.92	71.58	18.70
5	748	45	952	0.89	31.24	8.33	69.43	18.52
6	749	47	1053	0.86	30.22	7.79	64.30	16.58
7	750	46	976	0.86	26.64	8.00	57.92	17.40
8	751	48	1024	0.86	30.02	8.56	62.54	17.84
11	752	49	730	0.90	33.55	9.04	68.48	18.45

KAMM Mathematics Summary Statistics

Table A.2 reports summary findings for the KAMM Mathematics Assessments by grade level, identifying the number of items per KAMM form, the number of students administered a particular grade level KAMM, reliability coefficients, and descriptive statistics in terms of raw total scores and percent correct total scores. The reliability coefficients for grades 3 through 6 are above 0.85 and for grades 7 and 8, above .80. The 10th grade KAMM mathematics assessment has the lowest reliability coefficient (α = 0.750). The percent mean correct ranges from 47.42% (SD = 14.722) at grade 10 to 69.65% (SD = 18.224) at grade 3.

Table A.2. KAMM Mathematics 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
3	736	40	541	0.872	27.86	7.290	69.65	18.224
4	735	40	740	0.865	26.35	7.370	65.88	18.425
5	729	40	825	0.865	26.29	7.228	65.73	18.070
6	728	40	965	0.861	24.93	7.390	62.33	18.476
7	738	40	998	0.818	23.00	6.629	57.51	16.572
8	737	40	1060	0.809	22.00	6.724	64.17	18.057
10	730	40	843	0.750	18.97	5.889	47.42	14.722

KAMM Performance Classification Reliability

As described in Section 5 of the main part of the Technical Manual, score ranges on the KAMM assessments were established through standard setting procedures to classify students into five performance level categories (*Academic Warning*, *Approaches Standard*, *Meets Standard*, *Exceeds Standard*, and *Exemplary).* Procedures to estimate classification consistency and accuracy for the KAMM assessments mirrored those used for the general assessment test forms. Tables A.3 and A.4 present summaries of the classification consistency and classification accuracy indices for the KAMM test forms across testing grade and content area. Included in the tables is information on the overall test classification reliability and related information for dichotomous decisions at three cut-points, one for *Academic Warning* versus all levels above, one for *Exemplary* versus all levels below, and one for the most important AYP reporting decision, i.e., the bottom two categories versus the upper three categories, the latter three categories defining performance judged to be acceptable.

For Reading (Table A.3), overall test classification consistency values across all categories range from .56 (grade 3) to .64 (grade 5). Classification accuracy, on the other hand, is consistently higher than classification consistency yet maintains the same pattern of variability across grades. Classification accuracy coefficients range from .69 (grade 5) to .80 (grade 10). For Mathematics (Table A.4), classification consistency values range from .45 (grade 10) to .63 (grade 4). As in reading, classification accuracy is consistently higher than classification consistency yet maintains the same pattern of variability across grades. Classification accuracy coefficients range from .56 to .73. These coefficients are slightly lower than those reported for the general assessments, likely due to the fewer number of items on the test forms.

For both Mathematics and Reading, the reliabilities of classification at a given cut point are generally high whereas probabilities of misclassifications are low. For reading, all classification accuracy coefficients for the important AYP decision (levels 12 versus 345) are equal to or over .90. For mathematics, these coefficients are all equal to or over .85. The classification consistency values are slightly lower. These values support the adequacy of the KAMM assessments for making the major decision associated with AYP reporting.

Table A.3. KAMM Classification Reliability Indices by Cut Points in Reading

Grade	Cut Point*	Classification Accuracy	Classification Consistency	False Positive	False Negative
3	Overall	0.66	0.56		
	1 / 2345†	0.97	0.96	0.01	0.02
	12 / 345	0.93	0.90	0.03	0.04
	123 / 45	0.87	0.82	0.07	0.06
	1234 / 5	0.87	0.84	0.08	0.08
4	Overall	0.71	0.62		
	1 / 2345	0.96	0.94	0.01	0.03
	12 / 345	0.92	0.89	0.03	0.05
	123 / 45	0.90	0.86	0.05	0.05
	1234 / 5	0.91	0.87	0.06	0.03
5	Overall	0.73	0.64		
	1 / 2345	0.96	0.95	0.02	0.02
	12 / 345	0.93	0.90	0.03	0.05
	123 / 45	0.91	0.87	0.04	0.05
	1234 / 5	0.92	0.89	0.05	0.02
6	Overall	0.7	0.61		
	1 / 2345	0.95	0.93	0.01	0.04
	12 / 345	0.92	0.88	0.03	0.05
	123 / 45	0.9	0.85	0.06	0.04
	1234 / 5	0.93	0.90	0.04	0.03
7	Overall	0.71	0.63		
	1 / 2345	0.93	0.91	0.02	0.04
	12 / 345	0.9	0.86	0.04	0.06
	123 / 45	0.91	0.87	0.06	0.03
	1234 / 5	0.94	0.92	0.04	0.01
8	Overall	0.71	0.62		
	1 / 2345	0.95	0.92	0.01	0.04
	12 / 345	0.92	0.88	0.03	0.05
	123 / 45	0.9	0.86	0.06	0.04
	1234 / 5	0.94	0.91	0.03	0.03
11	Overall	0.71	0.62		
	1 / 2345	0.95	0.92	0.01	0.04
	12 / 345	0.92	0.88	0.03	0.05
	123 / 45	0.9	0.86	0.06	0.04
	1234 / 5	0.94	0.91	0.03	0.03

* 1 = Academic Warning, 2 = Approaches Standard, 3 = Meets Standard, 4 = Exceeds Standard, 5 =Exemplary

Table A.4. KAMM Classification Reliability Indices by Cut Points in Mathematics

Grade	Cut Point*	Classification Accuracy	Classification Consistency	False Positive	False Negative
3	Overall	0.72	0.62		
	1 / 2345†	0.97	0.96	0.01	0.02
	12 / 345	0.91	0.88	0.04	0.05
	123 / 45	0.9	0.86	0.06	0.04
	1234 / 5	0.94	0.92	0.05	0.01
4	Overall	0.73	0.63		
	1 / 2345	0.96	0.94	0.01	0.03
	12 / 345	0.9	0.86	0.05	0.05
	123 / 45	0.91	0.88	0.05	0.03
	1234 / 5	0.96	0.94	0.04	0.00
5	Overall	0.72	0.61		
	1 / 2345	0.96	0.95	0.01	0.03
	12 / 345	0.9	0.85	0.05	0.05
	123 / 45	0.91	0.97	0.06	0.06
	1234 / 5	0.95	0.93	0.04	0.04
6	Overall	0.71	0.61		
	1 / 2345	0.94	0.92	0.02	0.04
	12 / 345	0.89	0.84	0.06	0.05
	123 / 45	0.92	0.88	0.06	0.03
	1234 / 5	0.96	0.95	0.03	0.01
7	Overall	0.69	0.57		
	1 / 2345	0.95	0.91	0.00	0.05
	12 / 345	0.85	0.79	0.09	0.06
	123 / 45	0.93	0.90	0.05	0.02
	1234 / 5	0.97	0.95	0.02	0.01
8	Overall	0.67	0.56		
	1 / 2345	0.91	0.87	0.00	0.08
	12 / 345	0.85	0.79	0.09	0.06
	123 / 45	0.93	0.91	0.05	0.02
	1234 / 5	0.96	0.95	0.03	0.01
10	Overall	0.56	0.45		
	1 / 2345	0.84	0.78	0.00	0.16
	12 / 345	0.79	0.71	0.15	0.06
	123 / 45	0.94	0.91	0.04	0.02
	1234 / 5	0.98	0.97	0.02	0.01

* 1 = Academic Warning, 2 = Approaches Standard, 3 = Meets Standard, 4 = Exceeds Standard, 5 =Exemplary

Appendix B

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

Three primary references for describing the guidelines and procedures that define the Kansas Alternate Assessments are the *Alternate Assessment 2005 Teachers' Guide, Instructions for Teachers in Preparation for Scoring the Alternate Assessment,* and the *Alternate Assessment Q&A.* All of these documents can be found by following the KS Assessment/KS Alt Assessment links at www.kansped.org on the KSDE website and must be reviewed prior to a student taking the assessment. Below is a brief summary of the requirements and procedures as well as a form for use by raters to record their scores for each student assessed. To enter the rating scores on-line, raters or district/building personnel should use the procedures outlined in this guide.
Test Coordinators, district/building administrators, and other local personnel responsible for coordinating Alternate Assessment activities must become familiar with the tasks required to implement the Kansas Alternate Assessment. Many of these tasks involve management functions that must be completed prior to professional staff involvement in data collection, scoring, and data entry activities. These management functions are identified and described in each of the component sections below.

The following identify the major activities required to implement the Kansas Alternate Assessment:

1. Students meeting the criteria for receiving the Alternate Assessment must be registered. These criteria may be found in the *Alternate Assessment 2005 Teachers' Guide* at www.kansped.org.
2. Five indicators for each content area must be selected by the local IEP team as the focus of that year's assessment. Criteria for the distribution of indicators selected across curricular standards areas are given in Section II.
3. Three pieces of evidence are to be collected during the assessment window for each indicator and submitted to a student's folio using state identified procedures and forms. This will result in 15 separate pieces of evidence in Reading and 15 in Mathematics if a student is assessed in both content areas.
4. Each piece of evidence submitted is to be <u>independently</u> rated by three local raters using the required rubric.

5. Local score ratings are to be transmitted to CETE and will serve as the data elements for scoring and reporting.
6. Schools should store and maintain the Evidence Files for each student locally. The state, through CETE, will be requesting a sample of student Evidence Files for study and evaluation. The purpose of this review is not to monitor local district implementation of the Alternate Assessment, but rather to review and evaluate the implementation process and to identify areas in need of improvement.

Detailed explanations for each of these activities are provided in the following sections..

Alternate Assessment Requirements

I. Registration of Students for the Alternate Assessment

To access the names of the Alternate Assessment students, a district or building must sign onto the CETE website using the **Alternative Assessment ID number and Password** given to either the district or to the buildings within a district. A separate listing for these students should be available in the "Alternate Assessment" file. If a list of your students is not available, contact your Test Administrator who can use the Student Editor on the CETE website (www.cete.us/disttools) to identify those students who meet the criteria for taking the Alternate Assessment.

II. Selection of Assessment Indicators for a Student

The Kansas Alternate Assessment requires that five (5) indicators be selected for a student from each content area assessed. In addition, at least one indicator must be selected for assessment from each of the Extended Curriculum Standards areas. This applies to both the Reading and the Mathematics Extended Standards. In Reading, there are three (3) Standards' areas: 1) Reading, 2) Literature, and 3) Communication for Social Interaction. At least one indicator from each of these three areas must be selected for a student's assessment in Reading. The other two indicators may both be selected from one Standards' area or spread across two areas. In Mathematics, there are four (4) Standards' areas: 1) Number and Computation, 2) Algebra, 3) Geometry, and 4) Data. At least one indicator from each of these four areas must be selected for a student's assessment in Mathematics. The remaining indicator may be selected from any of the four Standards' areas. Choose the "Instruction to Prepare for Alternate Assessment Scoring" link at www.kansped.org for more detailed information.

> **NOTE:** If a district has not already complied with the above indicator selection criteria for a student, they must do so. If the assessment is conducted using fewer indicators or on an inappropriate set of indicators, the assessment for that student will be incomplete and the student will receive scores of zero (0) for the Standard(s) in which an indicator is missing or for indicators less than five. This likely will result in the student being classified as Not Proficient (i.e., Unsatisfactory or Basic).

III. Data Collection and the Evidence Folio (File)

Three (3) pieces of evidence need to be collected for each of the five (5) indicators selected for assessment. Thus, both a Reading and a Mathematics Evidence Folio for a student should each contain 15 separate pieces of evidence from the assessment's data collection. Evidence for the Alternate Assessment indicators must be collected during the first four weeks of the assessment window.

The following are guidelines for the data collection:

- When the indicator response is scored as correct/incorrect, a minimum of 5 trials/probes is required for each piece of evidence.
- Any written transcription of student's responses must be written verbatim.
- When using worksheets as data, 3 different worksheets equal 3 different pieces of data/evidence. The worksheets could contain the same information but they may not be identical worksheets.
- When responding to 3 different individuals, tasks, or environments, each is considered to be a different piece of evidence.
- The person collecting the data and submitting the evidence must not use the Alternate Assessment skill-scoring rubric to make summary judgments about the data or about the student's performance level. Evidence should describe the assessment process and the response of the student, but should not make judgments about the level of the student's performance.

In providing support:

- The teacher should first ask the student to respond without support.
- If the teacher provides support, it should be documented on the data label.
- The support provided <u>**should not exceed**</u> the support provided during instruction.
- At the point the teacher is completing the task or leading the student to a correct response, it cannot be considered appropriate support for assessment.
- **NOTE:** Hand over hand assistance is considered appropriate for IEP goals, but is <u>not</u> 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.

The evidence file should be organized as follows:

1. A single page listing the assessed indicators for which evidence is contained in the folio. This page may be printed from the CETE website. Once signed onto the CETE website with an Alternate Assessment ID and Password for your district or building, click on "Alternate Assessment" in the menu list in the left part of the screen. Then, click on "Alt. Students." This will bring up a list of students registered for the Alternate Assessment. To select/modify or view the indicators for a student, click on the number in the column under the "Indicators" heading. Any indicators previously selected for a student will appear on the screen.

   To print this list of indicators, select the "print" command from the menu as one would do to print any document. Before printing, be sure to select landscape from the printing preferences. The whole screen will be printed, but included is the list of indicators for the selected student. This printed listing may serve as the Cover Sheet for the Evidence Folio.

2. Each of the 3 pieces of evidence for the first indicator on the list needs to have a completed Evidence Label Form as the first sheet. This form is used to describe the assessment procedure for that piece of evidence. The form may be downloaded as a Word document from the *Instructions for Teachers in Preparation for Scoring the Alternate Assessment* at www.kansped.org by following the KS Assessment/KS Alt Assessment links. The instruction document also provides guidelines for completing the form.

3. The 3 pieces of evidence for the second, third, fourth, and fifth indicators need to follow the first indicator set in the evidence file. All pieces of evidence should include a completed Evidence Label Form.

IV. 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 rater.** It is recommended that the other two raters 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 is providing training in March. Information about the KSDE training dates, locations, and materials can be found in *Instructions for Teachers in Preparation for*

Scoring the Alternate Assessment at www.kansped.org. Training and checking for inter-rater reliability on evidence examples are important components of the process that need to be completed prior to rating actual evidence samples. However, checking for inter-rater agreement **should not occur** during the actual rating of student evidence folios. **These ratings must be done independently, and individual rater ratings must not be shared with other raters.**

A Rater Recording Worksheet specific to each student being assessed is provided through the CETE website. (Both a sample generated worksheet and a blank worksheet are included on the following pages.) When filling out the worksheet, raters must briefly describe the activity used to collect data for each piece of evidence. This description is required to make sure that each rater is evaluating and recording his/her score for the same piece of evidence labeled as Evidence Sample #1, Sample #2, and Sample #3. It is critical that Evidence Sample #1 be labeled as such on the Evidence Label and that all raters treat it as the first evidence sample when recording their scores. All raters must know which evidence samples have been labeled as #1, #2, and #3 and must record their scores on the worksheet to correspond with the appropriately labeled piece of evidence.

In order to make sure that all raters are viewing the same pieces of evidence labeled as Evidence #1, #2, and #3, a designated person (either the special education teacher serving as a rater or other district/building personnel) needs to first download and print the Rater Recording Worksheet(s) from the CETE website. In the space provided on the form, he/she is then to write a brief description of the activity used to collect data for each of the pieces of evidence in the folio. This description serves to distinguish one piece of evidence from another. Before making and recording his/her ratings, this person is to make two additional copies of the recording form with the evidence sample activities described, giving a copy to each of the other two raters for use in recording their ratings.

To download or print a copy of the worksheet with the student's name and indicators already identified on the worksheet, do the following:

1. Sign onto the CETE website with an Alternate Assessment ID and Password for your district or building.

2. Click on "Alternate Assessment" from the menu list in the left part of the screen.

3. Next, click on "Alt. Students." This will bring up a list of students registered for the Alternate Assessment.

4. In the box above the student listing is the command "Print Rater Recording Worksheets." Click on this command.

5. A screen appears with options to generate all students' Reading or Mathematics worksheets or to print worksheets for individual students. Click on the appropriate option to generate the worksheets.

6. Save or print the worksheets generated. Note that these worksheets are in a "pdf" format.

SCORER WORKSHEET FOR RECORDING KANSAS ALTERNATE READING ASSESSMENT RATINGS
Student name: Dehaemers, Carmen   Rater (circle one): 1 2 3

Indicator 1: ER1.1.1 - The student has a level of alertness that is influenced by external events.

Record Evidence #1 Score:		Record Evidence #2 Score:		Record Evidence #3 Score:	
(Briefly Describe activity)		(Briefly Describe activity)		(Briefly Describe activity)	

Indicator 2: ER2.1.1 - The student identifies the main character.

Record Evidence #1 Score:		Record Evidence #2 Score:		Record Evidence #3 Score:	
(Briefly Describe activity)		(Briefly Describe activity)		(Briefly Describe activity)	

Indicator 3: ER3.1.1 - The student acknowledges a potential communication partner.

Record Evidence #1 Score:		Record Evidence #2 Score:		Record Evidence #3 Score:	
(Briefly Describe activity)		(Briefly Describe activity)		(Briefly Describe activity)	

Indicator 4: ER3.1.2 - The student greets others.

Record Evidence #1 Score:		Record Evidence #2 Score:		Record Evidence #3 Score:	
(Briefly Describe activity)		(Briefly Describe activity)		(Briefly Describe activity)	

Indicator 5: ER3.2.1 - The student listens attentively.

Record Evidence #1 Score:		Record Evidence #2 Score:		Record Evidence #3 Score:	
(Briefly Describe activity)		(Briefly Describe activity)		(Briefly Describe activity)	

RATER WORKSHEET FOR RECORDING KANSAS ALTERNATIVE ASSESSEMENT RATINGS

Student name:____________________________ ☐ Reading ☐ Mathematics Rater (circle one): 1 2 3

Indicator 1:____________

Record **Evidence #1** Score:	Record **Evidence #2** Score:	Record **Evidence #3** Score:
(Briefly Describe Activity)	(Briefly Describe Activity)	(Briefly Describe Activity)

Indicator 2:____________

Record **Evidence #1** Score:	Record **Evidence #2** Score:	Record **Evidence #3** Score:
(Briefly Describe Activity)	(Briefly Describe Activity)	(Briefly Describe Activity)

Indicator 3:____________

Record **Evidence #1** Score:	Record **Evidence #2** Score:	Record **Evidence #3** Score:
(Briefly Describe Activity)	(Briefly Describe Activity)	(Briefly Describe Activity)

Indicator 4:____________

Record **Evidence #1** Score:	Record **Evidence #2** Score:	Record **Evidence #3** Score:
(Briefly Describe Activity)	(Briefly Describe Activity)	(Briefly Describe Activity)

Indicator 5:____________

Record **Evidence #1** Score:	Record **Evidence #2** Score:	Record **Evidence #3** Score:
(Briefly Describe Activity)	(Briefly Describe Activity)	(Briefly Describe Activity)

In making their judgments about a student's skill level, raters should use the following rubric.

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:

1. Rate each piece of evidence separately using the rubric definitions and point values. Record your rating in the appropriate cell on your Rater Recording Worksheet. Make sure that the evidence sample you are evaluating and calling sample #1, #2, or #3 corresponds with the activity description written for that evidence sample on your recording form. It is critical that the scores be recorded consistently in the same cells for the same pieces of evidence across raters.

2. A rater is not to discuss his/her ratings or the contents of an evidence file with other raters. The ratings must be done independently, based on the judgment of each rater, given whatever evidence is in the Evidence Folio.

3. When a trials or probes design for data collection is used for a single piece of evidence, a minimum of 5 trials is required. If fewer than 5 trials are used, the missing trials should be counted as incorrect when 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 used in assigning a point value on the rubric. In this example, available evidence demonstrates 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.

4. 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 performed correctly to the corresponding value on the 5-point rubric scale.

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

5. The level and type of support provided when responding to the student is to be described on the Evidence Label for each piece of evidence. 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). However, the teacher is not to lead the student to an answer or response with skill related cues/prompts. If the teacher completes the task or leads the student to a

correct response, it cannot be considered appropriate support for assessment. If the evidence file indicates he/she is doing so, 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 must be assigned a rating of "1" on the skill rubric scale.

6. 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 scale for each missing evidence piece.

V. Transmission of Score Ratings to CETE

Once a rater has completed his/her ratings and recorded the ratings on a Rater's Recording Worksheet, he/she needs to enter those ratings on-line at the CETE website. Each rater should enter his/her scores independent of other raters to avoid being influenced by the scores given by other raters.

To facilitate this, districts/buildings are able to create Alternate Assessment "rater accounts" on the CETE website. When an account is created for a rater, a rater ID and Password will be issued specific to that rater. This ID and Password will be used by the rater to log onto the CETE website where he/she can enter score ratings. A rater account should be created for each individual rater assigned to rate at least one evidence file for at least one student.

To create the "rater accounts," use the following guidelines

- Each individual rater in a district/building must have a separate ID and Password.
- A rater working with multiple students need have only one rater account ID and Password. He/She will be able to enter scores for multiple students using the single ID and Password.
- Each district/building should create a minimum of three (3) "rater account" IDs and Passwords (if only one student is being assessed) to a maximum of three (3) times the number of students being assessed (creating an entirely different ID/Password set for all 3 raters for each student).

A. Generating "Rater Account" IDs and Passwords

The "rater account" IDs and Passwords may be generated at either the district or building level. To generate these accounts, complete the following steps:

1. Log onto the CETE website at www.cete.ku.edu/disttools using the Alternate Assessment ID and Password assigned to your district or building by CETE. Check with your district Test Coordinator if you need the ID and Password information specific to managing the Alternate Assessment.
2. Click on "Alt. Assessment" in the left column menu list.
3. Click on "Rater Accounts."

4. Click on “Create New Raters” in the first box. You will be asked to provide the name of the rater and then to identify that person’s job role in your district/building from the following list:

   1 - special education teacher providing the student's special education services
   2 - general education teacher
   3 - general education administrator
   4 - special education administrator/coordinator
   5 - special education teacher, but not for this student
   6 - school psychologist
   7 - speech pathologist
   8 - other special education/related service provider
   9 - other education service provider (counselor, Title I, reading, etc.)

   Once this has been completed, the rater’s name will appear on a list with an assigned ID and Password.

5. Give the assigned ID and Password to the appropriate individual rater to be used to enter scores on the CETE website.

B. Entering Rater Scores on the CETE Website

Once “rater account” IDs and Passwords have been generated and assigned to specific raters, they are to be used to enter scores on the CETE website. Ideally, each individual rater will sign onto the website and enter the scores he/she has recorded on the Rater Worksheet for a student. Because separate IDs and Passwords have been assigned to raters, this process maintains the independence of the ratings across raters. On the CETE website, raters will be able to enter and review their individual scores, but will not be able to view scores assigned by other raters.

If the district/building wants to assume responsibility for the entry of all raters’ scores for all students assessed, they may do so. However, the person responsible for entering the scores cannot be one of the raters. A rater should never have access to the scores assigned by other raters. If a district/building chooses this approach for score entry, it is **critical** that the person entering the data **log onto the CETE website using the individual rater’s ID and Password when entering the scores recorded by that rater.**

In addition to scores, raters are asked an evaluative question as they enter data on the CETE website. These questions provide feedback to the state on the perceived adequacy of the Evidence Folios in the Alternate Assessment process. If a district/building assumes responsibility for the entry of all raters’ scores for all students assessed, it must also collect and enter rater responses to the evaluative question. As the question is specific to a content area evidence folio for a student, responses to the question are requested for each set of data entered for a rater. Raters should answer the following evaluative question:

1. 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?
   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).

To enter rating scores, a rater or assigned district/building personnel should do the following:

1. Log onto the CETE website at www.cete.ku.edu/disttools using the Alternate Assessment ID and Password assigned by CETE to your district or building.

2. Click on "Alt. Assessment" in the left column menu list.

3. Click on "Score Entry."

4. A list of students assigned to the Alternate Assessment should appear. Click on "Scores" in the row for the student for whom you want to enter scores. Note that if a student was registered for the Alternate Assessment in both Reading and Mathematics, his/her name will be listed twice. Make sure you click "Scores" for the Content Area (Reading or Mathematics) for which you will be entering scores.

5. On the screen that appears, respond to the question and enter your ratings for each indicator piece of evidence by clicking on the appropriate response option. Check to make sure the correct student name is identified and that the indicators identified are the indicators for which you will be entering scores.

6. The response configuration allows you to enter 15 scores, one score for each of the 3 pieces of evidence for each of 5 indicators assessed. If ratings do not exist for one or more cells because one or more pieces of evidence was not collected or fewer than 5 indicators were assessed, click on the "Did not rate" option. This option will record a score of "0."

7. When finished entering the 15 score values, click on the "SAVE" button at the bottom of the screen. The values saved will be displayed for your review and editing.

8. The list of students will appear again. Continue entering rating scores for students. When finished, log out.

Any student data entered under a specific rater ID and Password are automatically linked to that rater. CETE will link data for a student across raters by using the student's state ID number.

VI. Locally Maintaining Student Evidence Folios

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

The state, through CETE, will be requesting that a sample of student Alternate Assessment Evidence Files be sent to the state for study and evaluation. This review is not intended to monitor local district implementation of the Alternate Assessment, but rather to review and evaluate the implementation process and to identify areas in need of improvement. One or more of your student Evidence Folios may be part of the requested sample. If so, you will receive a request indicating the materials to be sent.

Part 2: Alternate Assessment Reliability and Validity Information

Alternate Assessment Reliability Information

Reading Score Reliability

1. The correlation coefficients among all raters' total rating scores for 1783 Alternate Assessment students were:
   a. rater1 vs rater 2, .916
   b. rater1 vs rater 3, .912
   c. rater 2 vs rater 3, .923

2. The percent perfect agreement among all raters across the 15 indicators ranged from a low of 84.5 percent to a high of 87.1 percent of 1783 ratings per indicator. This percent agreement was over 98 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 125 student data folios was .504.

4. When 22 student evidence files found deficient by the external "expert" reviewer were removed from the analysis, the correlation between the average rating for the three local ratings and a fourth external "expert" reviewer of a sample of 103 student data folios was .711.

Mathematics Score Reliability

1. The correlation coefficients among all raters' total rating scores for 2100 Alternate Assessment students were:
   a. rater1 vs rater 2, .940
   b. rater1 vs rater 3, .929
   c. rater 2 vs rater 3, .930

2. The percent perfect agreement among all raters across the 15 indicators rated ranged from a low of 84.2 percent to a high of 87.0 percent of 2100 ratings per indicator. This percent agreement was over 98 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 130 student data folios was .651.

4. When 13 student evidence files found deficient by the external "expert" reviewer were removed from the analysis, the correlation between the average rating for the three local ratings and a fourth external "expert" reviewer of a sample of 117 student data folios was .738.

KAA Performance Classification Reliability

As described in Section 5 of the main part of the 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)*. Procedures to estimate classification consistency and accuracy for the KAMM assessments mirrored those used for the general assessment and KAMM 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 two sets of values are presented, one set for Reading and one set for Mathematics. Table B.1 presents summaries of these classification consistency and classification accuracy indices for the KAA system in Reading and Mathematics. Included in the table is information on the overall test classification reliability and related information for dichotomous decisions at three cut-points, one for *Academic Warning* versus all levels above, one for *Exemplary* versus all levels below, and one for the most important AYP reporting decision, i.e., the bottom two categories versus the upper three categories, the latter three categories defining performance judged to be acceptable.

In Table B.1, the data for both Reading and mathematics are near mirror images of one another. The overall test classification consistency across all categories is .79. The classification accuracy coefficient is .85. These coefficients are somewhat higher than those reported for either the general or KAMM assessments.

For both Mathematics and Reading, the reliabilities of classification at a given cut point are all high. For both content areas, the classification accuracy coefficient for the important AYP decision (levels 12 versus 345) is .97. These values support the adequacy of the KAA for making the major decision associated with AYP reporting.

Table B.1. KAA Classification Reliability Indices by Cut Points in Reading and Mathematics

Grade	Cut Point*	Classification Accuracy	Classification Consistency	False Positive	False Negative
Mathematics					
	Overall	0.85	0.79		
	1 / 2345	0.99	0.99	0.00	0.01
	12 / 345	0.97	0.96	0.01	0.02
	123 / 45	0.95	0.93	0.03	0.02
	1234 / 5	0.94	0.91	0.04	0.03
Reading					
	Overall	0.85	0.79		
	1 / 2345	0.99	0.99	0.00	0.01
	12 / 345	0.97	0.96	0.01	0.01
	123 / 45	0.95	0.94	0.02	0.02
	1234 / 5	0.93	0.91	0.04	0.03

* 1 = Academic Warning, 2 = Approaches Standard, 3 = Meets Standard, 4 = Exceeds Standard, 5 =Exemplary

Alternate Assessment Validity Information

The results from two separate data collection efforts are presented below. First, when local raters were entering their ratings on-line on CETE's website, they were requested to rate the overall clarity of the evidence samples for the student for whom they were entering data. This information provides feedback on the perception of the quality of the evidence in a student's data folio from the rater's perspective. Second, as part of the Standard Setting (cut scores) process, a sample of student data folios were sent for review to 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.

Local Rater Evidence Clarity Judgments – Reading

When entering their ratings, 1850 local raters provided judgments on the overall clarity of the evidence samples for the reading indicators in allowing them to make their ratings with confidence. The first table below summarizes the job position of the raters responding. Of those responding, 86.6 percent indicated they "had no problems in making judgments for any of the evidence samples." An additional 10.6 percent indicated they "had difficulty in making judgments for a limited number of the evidence samples (1 – 2), but the rest were sufficiently clear." This brings the total to 97.2 percent indicating the data were sufficiently clear to pose no or very limited problems in allowing them to make their ratings with confidence.

# of raters	Rater job responsibility:
1141	special education teacher providing the student's special education services
40	general education teacher
21	general education administrator
94	special education administrator/coordinator
142	special education teacher, but not for this student
191	school psychologist
99	speech pathologist
89	other special education/related service provider
33	other education service provider (counselor, Title I, reading, etc.)
1850	

# of raters	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?
1384 (86.6%)	I had no problems in making judgments for any of the evidence samples.
169 (10.6%)	I had difficulty in making judgments for a limited number of the evidence samples (1 – 2), but the rest were sufficiently clear.
37 (2.3%)	I had difficulty in making judgments for several of the evidence samples (3 – 7), but over half were sufficiently clear.
7 (.4%)	I had difficulty in making judgments for more than half of the evidence samples (8 – 13) with only a few being sufficiently clear.
1 (.1%)	I had difficulty in making judgments for almost all of the evidence samples (14 – 15).
1598	

Local Rater Evidence Clarity Judgments – Mathematics

When entering their ratings, 2191 local raters provided judgments on the overall clarity of the evidence samples for the mathematics indicators in allowing them to make their ratings with confidence. The first table below summarizes the job position of the raters responding. Of those responding, 87.8 percent indicated they "had no problems in making judgments for any of the evidence samples." An additional 9.6 percent indicated they "had difficulty in making judgments for a limited number of the evidence samples (1 – 2), but the rest were sufficiently clear." This brings the total to 97.4 percent indicating the data were sufficiently clear to pose no or very limited problems in allowing them to make their ratings with confidence.

# of raters	Rater job responsibility:
1379	special education teacher providing the student's special education services
62	general education teacher
43	general education administrator
112	special education administrator/coordinator
175	special education teacher, but not for this student
180	school psychologist
100	speech pathologist
96	other special education/related service provider
44	other education service provider (counselor, Title I, reading, etc.)
2191	

# of raters	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?
1666 (87.8%)	I had no problems in making judgments for any of the evidence samples.
183 (9.6%)	I had difficulty in making judgments for a limited number of the evidence samples (1 – 2), but the rest were sufficiently clear.
42 (2.2%)	I had difficulty in making judgments for several of the evidence samples (3 – 7), but over half were sufficiently clear.
5 (.3%)	I had difficulty in making judgments for more than half of the evidence samples (8 – 13) with only a few being sufficiently clear.
1 (.1%)	I had difficulty in making judgments for almost all of the evidence samples (14 – 15).
1897	

Independent Expert Panel Judgments

At the conclusion of the spring 2006 testing, a random sample of Alternate Assessment students from across the state was identified. Schools were contacted and were requested to send in the data folios for these students. From this pool of randomly sampled student data folios, samples representing students across grades and content were drawn to be sent for review by an independent panel (n=18) of trained expert judges. As part of their review, these expert judges were asked to respond to the following questions addressing the quality (validity) of the student data folios.

2. 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).

3. 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.

4. 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.

In addition, the following statement provided the opportunity for the expert judge to comment on the quality of the evidence file for a student.

5. 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.

Reading Judgments

In reading, 127 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 reading data folios, 74.1 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, 76.4 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, 69.3 percent of the folios were judged to be perfectly or highly compliant with an additional 16.5 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?
67 (52.8%)	I had no problems in making judgments for any of the evidence samples.
27 (21.3%)	I had difficulty in making judgments for a limited number of the evidence samples (1 – 2), but the rest were sufficiently clear.
18 (14.2%)	I had difficulty in making judgments for several of the evidence samples (3 – 7), but over half were sufficiently clear.
6 (4.7%)	I had difficulty in making judgments for more than half of the evidence samples (8 – 13) with only a few being sufficiently clear.
9 (7.1%)	I had difficulty in making judgments for almost all of the evidence samples (14 – 15).
127	

# 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?
73 (57.5%)	All assessment procedures were very appropriate.
24 (18.9%)	The vast majority were appropriate with only 1 or 2 procedures being questionable across the 3 evidence pieces for the 5 indicators.
11 (8.7%)	The majority of procedures were appropriate, but there were several (3 – 7) that were questionable across the 3 evidence pieces for the 5 indicators.
11 (8.7%)	More than half of the procedures (8 – 12) were questionable with only a few being sufficiently appropriate.
8 (6.3%)	Almost all of the procedures (14 – 15) were questionable.
127	

# 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?
59 (46.5%)	Perfectly compliant.
29 (22.8%)	Highly compliant with few irregularities.
21 (16.5%)	Generally compliant, but with some irregularities.
9 (7.1%)	Much of the information had irregularities.
9 (7.1%)	Almost all of the information had irregularities.
127	

Mathematics Judgments

In mathematics, 132 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 reading data folios, 79.5 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, 78.1 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, 71.2 percent of the folios were judged to be perfectly or highly compliant with an additional 22.7 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?**
66 (50.0%)	I had no problems in making judgments for any of the evidence samples.
39 (29.5%)	I had difficulty in making judgments for a limited number of the evidence samples (1 – 2), but the rest were sufficiently clear.
16 (12.1%)	I had difficulty in making judgments for several of the evidence samples (3 – 7), but over half were sufficiently clear.
8 (6.1%)	I had difficulty in making judgments for more than half of the evidence samples (8 – 13) with only a few being sufficiently clear.
3 (2.3%)	I had difficulty in making judgments for almost all of the evidence samples (14 – 15).
132	

# 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?**
69 (52.3%)	All assessment procedures were very appropriate.
34 (25.8%)	The vast majority were appropriate with only 1 or 2 procedures being questionable across the 3 evidence pieces for the 5 indicators.
17 (12.9%)	The majority of procedures were appropriate, but there were several (3 – 7) that were questionable across the 3 evidence pieces for the 5 indicators.
9 (6.8%)	More than half of the procedures (8 – 12) were questionable with only a few being sufficiently appropriate.
3 (2.3%)	Almost all of the procedures (14 – 15) were questionable.
132	

# 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?**
57 (43.2%)	Perfectly compliant.
37 (28.0%)	Highly compliant with few irregularities.
30 (22.7%)	Generally compliant, but with some irregularities.
5 (3.8%)	Much of the information had irregularities.
3 (2.3%)	Almost all of the information had irregularities.
132	

Changes in KAA for 2007

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 2007 implementation to provide more direction and standardization of the assessment process.

In addition, procedures are in place to collect data using the Learning Characteristics Inventory developed by the University of Kentucky. This inventory has been recommended as a means to provide validity information on the description and appropriateness of the student population participating in the Alternate Assessment.