

# TECHNICAL REPORT

## Y2000 Kansas Assessments in Mathematics, Reading, and Writing

Prepared by:

Douglas R. Glasnapp, John P. Poggio and Md Hafidz Omar,

Center for Educational Testing and Evaluation

University of Kansas

### Table of Contents

Introduction and Background .....	2000-2
Development of the Assessments and Content-Related Validity Evidence.....	2000-7
Differential Item Functioning .....	2000-18
Test Equating .....	2000-33
Classification Accuracy .....	2000-47
References.....	2000-68
Attachment A: Content Validation Review Form .....	2000-69

## **Introduction and Orientation to the Kansas Assessments**

This technical manual provides information on the psychometric properties of the year 2000 *Kansas Assessments in Reading Writing, and Mathematics*. The purposes of Kansas assessments are to:

- (1) provide aggregate state accountability and progress information toward meeting the Kansas Curriculum Standards in the tested areas;
- (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, writing and mathematics were planned, created beginning in May 2000 and then administered in the Spring 2000. Grade 5, 8 and 11 students participated in the reading and writing assessments. Grade 4, 7 and 10 students participated in the mathematics 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 2000 administration of the Kansas assessments serves as the baseline for a new cycle of state assessments. The assessments administered were all newly developed to measure the new targeted indicators (outcomes) in the most recent editions of the state curricular Standards for the content areas. These documents must be referenced when examining and evaluating any of the information resulting from the state assessment programs. The 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](http://www.ksde.org).

As the baseline year of the new round of assessments, the Spring 2000 administration of the Kansas assessments incorporated important changes from prior Kansas assessments. Curriculum standards were changed and targets for the assessments restricted, performance assessments were formally abandoned by the Board of Education, and test specifications revised (e.g., reading text types expanded, math skills narrowed, and writing at various grades more narrowly specified and controlled). In effect, no comparison to past student, building, district or state performance can be made. To achieve a long term assessment and accountability system projected to be in place for a minimum of five academic years, there were four different parallel forms of both the reading and mathematics tests created then administered at each grade level. The tests were distributed and administered such that equivalent groups of students within classrooms, buildings, districts and across the state took each form of a content area test. In subsequent years according to a specified plan, different intact forms will be cycled through the assessment to afford comparisons for growth over time at the school, district and state levels. 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 (Y2000). 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.

Because writing still offers flexibility in the way in which the assessments are implemented across districts, standardization of the process is expected (and required) to occur locally. Using 2000 writing results as a baseline for trend is appropriate only when the local implementation process does not change in subsequent years.

In each reading assessment test form, four authentic, extended, authentic reading selections representing different text types were included (a narrative, an expository, a technical, and a persuasive selection). The questions asked over these selections were in

the form of multiple-correct yes/no items. That is, a question was posed to the student and several (typically 4 or 5) choices were presented as alternative responses to the question. The student responded “yes” or “no” to each answer choice to indicate whether that option was a correct (yes) or incorrect (no) response to the question. More than one of the answer choices could be correct to any one question. Given this format proscribed by KSDE, the reading test took on the format of a multiple yes/no (or true/false) test for the student. Each yes/no response was scored as correct/incorrect and contributed equally to a student’s score. A student’s score over a specific text was determined by the number of correct decisions made.

As in prior assessment years, the writing assessment continued to be based on a Six-Trait scoring model. However, the type of writing required from 8<sup>th</sup> and 11<sup>th</sup> grade students changed from prior years. The writing prompts at grade 8 were intended to elicit explanation/expository writing. At grade 11 the prompt were intended to elicit convincing/persuasive writing. The grade 5 assessment continued to require that students write to narrative prompts as in prior years. When a student’s paper was determined to be “off topic” the paper was scored as zero. The six traits remain constant at all grade levels with ideas/content, organization, voice, word choice, sentence fluency and conventions being measured. While the trait identifiers are the same across grade levels, the scoring rubrics that define each trait differ across grade levels and are specific to the type of writing required at that grade level. Faculty and staff reviewing writing assessment results refer to each grade’s rubric to evaluate specific strengths and weaknesses while scoring the paper.

Additionally, the writing assessment was flexible in its implementation locally. While the recommended time for the assessment was four separate days, some districts used fewer days for administering the assessment. Some districts allowed students to use a word processor; some did not. The other notable difference in implementation of the writing assessment was the number of ratings of student responses conducted locally. Some districts provided two ratings locally and thus had only 10 percent of their students’ papers also scored by state raters, while other districts chose to provide only one local rating and have the state raters score all their students’ papers. These choice points

exercised by local districts (which are to remain constant over years) created a basis for specific norms construction.

The mathematics assessments followed a multiple choice, selected response testing format. Items on these tests were multiple-choice with only one correct answer to be selected from the response options provided to a question. Each question on a test form contributed equally to a score value. The numbers of items measuring each indicator in the Kansas Curricular Standards for Mathematics were not equal on a test form, however. Thus, indicators were differentially represented by items based on specifications set down by the state. All test forms at a grade level did follow the same specifications for weighting the indicators.

This technical report is organized first presenting information on the item/test development procedures aimed at maximizing the content validity of the assessments as measures of the targeted indicators in the state's Curricular Standards. Then results from psychometric analyses are presented in the sequence in which they were conducted for decision-making. The first psychometric results provide information on the results 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 are presented. The equating results are followed by a discussion of the analyses and procedures to determine the cut-scores for classifying students into one of five performance levels defined by the state. After the presentation of these initial, but necessary analyses and results, information is provided on the technical psychometric characteristics of the resulting assessments in each of the content areas.

Providing credible information as to the standing and progress of education to schools and others in Kansas with reference to the curriculum indicators targeted for assessment is the central mission of the Kansas assessments. Decisions about the assessment program's structure and how it is offered is formed demonstrably by this specific and proscribed intention. The methods and procedures put in place to monitor and evaluate the assessments are responsive to this purpose. For example, when presenting reliability information in the latter sections, the presentation distinguishes

between high and low stakes decisions. High stakes decisions have serious and often universal consequences for either the student, the school building or the school district, e.g., using the test information as part of the school accreditation process. Low stakes decisions do not automatically impact students or buildings, e.g., requiring the use of results to plan instruction. For use in high stakes decisions, test scores should have reliabilities at or above .85 (Nunnally, 1978; Herman, Aschbacher, & Winters, 1992). For use in low stakes decisions, a minimum reliability of .70 is used as the criterion for test scores (Nunnally, 1978; Herman, Aschbacher, & Winters, 1992; Feldt & Brennan, 1989; Reckase, 1997). Whether used for high or low stakes decisions, a test score alone should not to be used without other corroborating information or evidence (NCR, 1998).

The next section presents information related to the construction of the assessments.

## **DEVELOPMENT OF THE ASSESSMENTS AND CONTENT RELATED VALIDITY EVIDENCE**

### **Program Overview**

The Kansas assessments are planned and created to point to, reflect and otherwise operationalize the direction for needed curriculum and instruction changes in Kansas K-12 schools. The assessments grow out of, in part, the premise that what is tested is what gets taught in schools. In recent 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 are used to help monitor annual school progress to support Title I monitoring and evaluation requirements. As related to accountability, students are classified into one of five performance categories (Advanced, Proficient, Satisfactory, Basic and Unsatisfactory). Based on performance, a school may be identified as having achieved the state's "Standard of Excellence". Student classification and school decisions points have been arrived at using typical standard setting approaches, although final cutpoints are established by the Kansas Department of Education based on a review and consideration of the actual score distributions. Important to underscore is that results from the state assessments are not used at any level to make a predetermined decision (continuing accreditation, financial rewards, advancement, promotion, etc.).

The assessments set out academic challenges and high standards for Kansas students and educators. The assessments present challenging tasks drawn from fairly well defined content standards. At this time the state assessments are constructed to provide input and assist with local educator's understanding a student's achievement with reference to the Kansas subject area Curricular Standards, and to inform officials as to the performance of schools toward achieving these Standards. Any other use, action or inference based on performance on the Kansas assessments was not considered during the development of the assessments. In this section the general procedures followed to construct the Kansas assessments In Reading, Writing, Mathematics, Science and Social Studies are detailed. The presentation offers a description as to what methods are characteristically followed to

ready the assessments for all populations tested. Later sections of this manual are more exact in describing specific procedures that may only be touched upon in this presentation (e.g., empirical DIF, setting cutscores, item analysis results, etc.). As the development methodology follows a “content validation” approach, we begin with a limited description of the populations for whom assessments are prepared and information regarding the content of the assessments themselves.

### **Who is Tested, When, and Over What**

All Kansas students at the designated grades including special education and English Language Learners (ELL) students are tested. Students in both public and private schools (accredited and non accredited) are tested. The only children excluded for the state’s assessments are ELL students whose English proficiency is extremely low (e.g., LAS scores of 1). While the vast majority of all tested students sit for the general education assessment in a content area and grade level, the program also provides for an Alternate Assessment (that 1 percent of the school population whose learning needs cannot be met by regular school curriculum), Plain English assessments (specially design and prepared for ELL students to reduce the language load of the general education examinations while preserving the actual skills being tested), and a series of Modified assessments (approximately 3 percent of the students whose IEP specifically demonstrate that the general education assessments content standards are not suitable for the individual student). All assessments are developed based on or derived from the curriculum/content standards associated in each academic subject (Reading, Writing, Mathematics, Science, and Social Studies). In addition to reports that summarize the performance of all students tested, separate individual, building and district summary reports are also prepared for regular education and special education, LEP, and migrant students. Examinations are administered on a somewhat variable calendar: writing, reading and mathematics from mid-February through mid-March; science and social studies from early-March to early April.

As mentioned, the assessments grow from the Kansas Curriculum Standards in the five subject fields. Curriculum standards are periodically reviewed and changes are made as determined by state curricular advising committees. During the latest cycle of review (1999 and 2000), two strategic decisions toward shaping the curriculum standards in each content area were imposed: first, the standards have been rewritten during revision to maximize the specificity of the “indicator” statements; and second, while many indicators are specified at each grade, we have moved to “targeting” the most important skills and indicators for the assessment. For example, grade 4 mathematics posits more than 70 indicators as learning expectations, yet only 25 have been targeted for the assessment, that is, will actual be used to form the assessment at that grade. With the exception of writing, typically 25 to 30 content area indicators are targeted and fixed for each of the state assessments at each grade. These targeted indicators are shared widely with the schools. The choice of the targeted indicators is left to the discretion of the assorted advisory committees, but the criteria embraced to reduce the list of indicators to be tested annually are to center on three considerations: (1) those indicators viewed as the most important and relevant to what needs to be mastered at the tested grade, (2) those indicator viewed as important to have mastered for the student to be successful at the next grade in the subject area, or (3) the indicator points to a skill that merits attention in the curriculum (that is, there is a sense that the skill has not been adequately attended to in the past, thus the test drives curriculum reform).

Curriculum standards are targeted at higher order outcomes including critical thinking, diverse communication skills, problem solving, reasoning, and decision making as well as those key and essential knowledges and understandings important for student to have in order to assure their core knowledge. Given the grade and content area from 35 to 65 percent of the coverage of an examination will focus on advanced cognitive processing skills. Using the targeted Kansas Curriculum Standards as the beacon, annually the state assessments are crafted largely by Kansas educators identified by their districts and state professional association leaders. The assessments are a product of Kansas educators whose development is coordinated by the Center for Educational Testing and Evaluation

(CETE) at the University of Kansas and the Kansas Department of Education (KSDE). The next section discusses the approaches to assessment development.

## **METHODOLOGY AND PROCEDURES FOR ASSESSMENT DEVELOPMENT**

The approach to development of an assessment relies almost entirely on Kansas educators and resources (the Braille test booklets are produced outside of Kansas and are prepared by the testing staff at the American Printing House for the Blind). The model of test construction is a content validation development approach (judgmental) that over time is then supported by empirical validation for alteration and change. A series of steps are generally followed leading to the creation of an assessment. Before actual development of the items are begun, a critical first task is CETE's work with the appropriate state advising committee to define the general structure and format for each assessment (by grade as well). Once there is agreement with KSDE on the specifications for an examination (number of items, amount of time for testing, format/layout of items, distribution of items/specifications, structure and coverage of the questions, responsibility of differing item types for content coverage, etc.), the steps to create the items ordinarily proceed as follows.

1. Four to 6 experienced (minimum of three years teaching at the grade level in the content area), highly regarded Kansas teachers at the grade for the content area are selected based on nominations received from local districts. Persons selected are then trained in item writing techniques and rules and begin the creation of the assessment questions using the applicable Curriculum Standards as their sole guide. Brought to a common location for initial training, teachers following training (about two days) will work independently crafting their first draft items. During this initial stage of item development CETE staff and content experts are available to work with the item writers as questions are conceived, drafted and finalized. While working on items participants are continuously encouraged that items must be appropriate for the specific indicator and the students at the tested grade.

We have tried out two distinct approaches to this “first stage drafting of items approach”: (1) work with and train teachers for 2 to 3 days, then send them “home” to complete their assigned task (typically each person is expected to prepare 25 to 40 items); and (2) convene and keep together a working team in a content area and grade for a few weeks (3 to 4) during which time they draft and work together to ready the needed collection of items. The former approach tends to produce lots of questions that yield adequate stimulus material for the eventual questions, whereas the latter produces fewer questions, but ones of better quality. We have observed this in all content areas. Though we have considerable experience at both approaches, we have arrived at no conclusion as to a “preferred” or “superior” approach. Frankly the latter approach is very time consuming and requires constant management. The former as noted yields the quota of items, but the quality is not as good. Depending on the test constraints and time, we will follow either approach (e.g., for 2000 reading and math we convened and sequestered teams for 8 weeks, but for 2001 science and social studies item writers were trained, then worked for a few days).

2. As the second step of development, CETE staff receives work products, reviews, items, does some editing of items and prepares items for the next stage. Much work happens at this point in development. CETE’s development group reviews all items, screens for redundancy, and edits, and modifies items. The CETE group is made up of testing experts and content area experts. The goal at this stage is to include as many of the emerging items as possible, cleaning up, clarifying and rewriting items to move on to the next stage so items are received as credible, useable items. At this stage, content experts consider the fit of items to the particular indicators being measured and make adjustments as needed.
3. The work products (items) are next reviewed, revised, modified and contributed to by a second round of developers comprised of 4 to 7 Kansas curriculum specialists specific to the content area and grade level of an item pool. This is a vitally important step. It is at this stage that the expertise of the content disciplines is

necessarily brought to the task. Also, we begin to focus on breadth and depth of coverage in consideration of the indicators to be assessed. The process at this third step by participants not only edits, reviews, and modifies items based on expertise and experiences in the classroom, but new items are written as necessary to broaden and support coverage. As with the step 1 development task, the focus of this work is inclusive; dropping items to refine the pool is not a consideration or goal; working to improve and salvage as many of the drafted items as possible is the goal. This includes continuing to write items whenever judged necessary. Much of the effort at this stage focuses on improving the quality of the distractors in the items. The goal is to create a set of high “quality” questions before the field test stage. The approach taken by CETE is in contrast to one which only goes through steps 1 and 2 above and a larger number of items, but depends mainly on field-testing to separate the good from the bad. In addition to the quality and accuracy of the item, participants review and revise items to assure the representativeness of the items to the specific standard as well as their appropriateness for students.

Again, we have tried this stage of development and review in two ways: (1) convening groups of experts (4 to 7 in a subject area) who as a team work through the items for a content area and grade level (about two weeks), and (2) large (10 to 14) numbers of field-based educators who receive material via the mail, including instructions and guidance as to what they are expected to do (edit, revise, make specific suggestions, insert, add questions, etc.), and return their comments to CETE after a period of time. Our experience has been that both procedures produce high quality items with the first being more time and staff efficient.

4. Next CETE staff (testing experts and content specialists) go over all input received from Step 3 with a single purpose: to carefully, diligently and closely review each item given the revisions and comments received and attempt to finalize each item to as sound and defensible a form as possible. This step is done by a team of measurement and content area people (usually 3 to 5 persons) going over each item. Depending upon schedules and availability, KSDE curriculum staff participate. At

a minimum, KSDE provides written comments as part of this review stage process. At the conclusion of this step, there exists a well-found and strong pool of candidate items for the final forms of the tests. The goal at the end of this stage is to have developed approximately twice the number of items needed to produce the final forms of each of the content area grade level tests.

5. The test items as they come through Step 4 are readied and grouped by indicator (i.e., content standard) for review by other independent groups of field-based reviewers. As noted, on the order of twice the actual number of items thought to be needed go into this stage of review. This step is the "item to tested skill alignment" phase. From 10 to 14 persons for each grade level/content area examination are brought together to review the entire pool of questions. The group at each grade level by content area are comprised of approximately one third state content area advising committee members, and two-thirds classroom teachers who are teaching at the level for which the test is intended. Teachers are selected based on experience and evidence of their ongoing involvement and commitment at local and state levels to the content area. The task at this stage can be broadly defined as critical review and analysis of items for accuracy, readability, appropriateness for the students, as well as fit to and representativeness for the indicators. This is a vitally important step as it summates the external review of the items by content experts and field-based personnel. Attached is one of the reviewing forms used at this stage to direct the process. Participants are convened in a central location, the task discussed and detailed, then individuals work alone completing their critical review and appraisal as called for and described by the instructions form provided in Attachment A. Once the review of all items is completed (ordinarily a two day process of participant working individually), time is spent with the group at a grade level and content area reviewing and discussing their findings. This "discussion and debriefing" is used to identify if there are weaknesses in the pool, coverage, or structure of the configured items. The goal at this point is to receive feedback that guides CETE toward final edits and some trimming of the pool of questions. At

this stage and based on the review and feedback, it is sometimes the case that a few (very few) items are newly written.

6. CETE staff review results from step 5 and finalize items in the pool. Items judged to be misfits are dropped; item editing suggestions are accepted when there is evidence that the input is justified. If gaps in coverage are identified or shortages result, persons from steps 1 and 3 are asked to write specific needed items. Generally very few items are abandoned as poorly fitting (on the order of two to three per item pool), and only one or two items are necessary to add into the pool. The pool is framed and defined and is not likely to be added to or significantly altered coming out of this stage.
7. Though there has been up to this point considerable input and direction from very diverse vested parties to craft appropriate and suitable assessment items, this next step is vitally important and useful. All items coming to this stage and which could likely appear on an assessment are reviewed for bias, insensitivity and offensiveness by a committee composed of impacted class members (note: when actual testing is completed and before the results are officially sent to schools, empirical DIF procedures are used to evaluate for evidence of bias). Persons involved are chosen to represent the largest ethnic, cultural and racial populations of students in the state. In addition gender and disability advocates become involved.

CETE has conducted the logical review for sensitivity, bias and offensiveness following two formats. Early on independent panels of 4 to 6 persons representing an impacted class were convened to review items. Each panel would then set about reviewing all items. More recently we have come to use members of the Kansas Equity Council. Some years ago, federal resources were used to create such a council to serve as advocates for minority and historically impacted groups. They received training and have often been called upon to advise state agencies on issues and concerns that would arise related to fairness and equity. The Kansas Equity Council is used by CETE to carry out the logical review for bias, offensiveness and

insensitivity of all Kansas assessments. Additional members to the review panel are added to represent an impacted group if at the time of a review meeting specific members of the Equity Council are unable to attend. Use of the Council has been outstanding in the results obtained. Their review offers significant benefits to the entire development process. Their training makes them especially able to identify many problems and issues in the items and stimulus material readied for the assessments. It is no understatement to conclude that the Equity Council review strengthens the quality of the assessments and goes a long way to assuring fairness in the assessments. It is not at all unusual for ten to fifteen items in each pool to be identified as problematic and meriting alteration.

8. Coming out of the Equity Council review and advisory step, CETE prepares items for large scale pilot testing. Pilot testing is carried out in mid to later September (mandated testing occurs in February and March) using volunteer schools with students at the grade level above that scheduled for testing. Pilot test data are secured from 200 to 400 students per item. CETE spirals items onto forms based on the content standard indicator and then distributes multiple pilot forms in a way such that no school district obtains more than one-third of the item pool and preferably not more than a quarter. Pilot forms are prepared in a manner to limit actual pilot testing time to about 35 to 40 minutes under power test expectations. The pilot test booklets sent to a school are randomly distributed to students in each participating class. Administration of the pilot tests are carried out by local administrators or school counselors (and not classroom teachers). At least 20 percent of the schools in the state become involved in pilot testing (a much greater number at high school grades). Following the pilot test period in the schools, there is also an evaluation using student interviews to obtain general and specific reactions of students to the exposed questions (issues of readability, clarity, familiarity, etc.). Test administrators gage the time needed to administered the assessments and specifically monitor the readability of the assessments. Both classical item analysis and limited IRT methods are used to evaluate questions once data are returned. Based on data, CETE makes simple edits to an item, and

abandon items whose statistics indicate a problem. Statistics are used by CETE and KSDE to identify items that go onto final forms and balance the forms when more than one form is being created. Finally, items for the final booklets are chosen to assure maximum fit to and coverage of the specific indicators in correct proportion to the test specifications set down by KSDE and its advisors for the specific exam.

9. The pilot testing phase results in the “official” Kansas tests. Administration manuals and scoring guides are finalized, booklets and materials are proofread by two to four separate readers (editing of items continues at this stage), and finally printed and distributed to all Kansas districts. If any problem is identified with a specific item from the field during the test administration or from subsequent post-testing item analysis or empirical DIF analyses, the item is dropped or scoring is modified before student, school and district reports are returned.

### **Modifications to the process for select instruments**

Reading: As the reading assessment relies on authentic selections, that process begins by CETE working with about a dozen Kansas librarians who are commissioned to find reading selections fitting the content standards that are appropriate for the grade of students being tested in consideration of the standards. The librarians, nominated by their schools, are selected to represent the grade ranges for which an assessment is intended. The librarians meet once to review their selections and ideas and then go off to finalize their list of nominated pieces and selections. The state advising committee is next asked to review the nominated selections and sign off on their appropriateness. Before item development on passages begins, associated with any selection, the Equity panel screens the selections for their acceptability and offers suggestions and direction. Item development on the surviving passages then follow the processes detailed above, including a re-screen of the items by the Equity Panel.

Writing: Readyng the writing assessment scoring rubrics and prompts is a somewhat-truncated process from that detailed above. A grade level panel of 4 to 8 Kansas writing instructors were convened on separate occasions and worked on preparing the scoring rubrics and prompts in line with the state’s writing curricular standards and the KSDE specifications for the writing model (the six trait model, etc.). There is discussion and interaction among participants at and across grades. Prompts for grade assessments are drafted during these meetings. Participants go off with a selection of their grade level prompts and do limited local pilot testing of the prompts with classes at their school/district (a grade above the intended grade). Following their pilot, which includes scoring of paper, this group then reconvenes to finalize prompts given the students work in response to the emerging prompts. Steps 7 and 8 are then followed once the prompts have been finalized by the panel.

Special Education and ELL assessments: As some of these tests are drawn from the general education assessments, just prior to pilot testing special educators review assessments and make content changes as determined needed and appropriate given the SPED Modified Curriculum Standards. The same additional step is taken for the Plain English forms (except two reviews, by testing experts at CETE and UCLA, and separately by Kansas ELL instructors), work reviewing and editing items to reduce the language load of all questions. When an examination is prepared as a unique assessment, that is, it is not directly formed out of the general education assessment, a process as that described above is followed by appropriate Kansas school based personnel.

The next section begins the technical psychometric documentation associated with the Kansas assessments.

## **Differential Item Functioning (DIF) Analyses**

When tests are constructed, it is important that test items be examined to minimize any bias, insensitivity or offensiveness toward any gender or ethnic group. Evidence to address these issues is typically collected using two different approaches: 1) a logical judgmental review of test items by panels of persons representing impacted gender and ethnic groups, and 2) an empirical analysis of item responses. All reading passages, reading test items, mathematics test items and writing prompts were subjected to a logical review prior to the production of test booklets and forms. The majority of individuals forming the panels for the logical review of items and prompts for bias, insensitivity and offensiveness were class members of the state trained Equity Council (see discussion in the prior section). Additions to the reviewing panels were made if a sufficient number of individuals representing a specific ethnic group was not available on the review dates. Three individuals representing a specific ethnic group was set as the minimum representation. In addition to females as a historically affected gender group, participants were included to represent African-Americans, Asian-Americans, Hispanics and Native Americans, and handicapping conditions as historically impacted groups. All panels were convened in a central location. The procedure had the participants conduct individual reviews of each item in mathematics and reading, and the writing prompts after which there was entire group discussion and comment. Deletion of specific reading selections, prompts and items and revisions of items occurred following the recommendations of the review panel.

When test items are actually taken by students, they can function differently for ethnic and gender groups. One of the reasons for differential functioning is bias where an item can favor a specific group not because of their academic abilities but because of their cultural or gender experiences. There are also other rival explanations as to why an item can differentially function such as district curriculum differences and unintended multidimensional rather than a single dimensional construct measured by a test.

Thus in addition to the logical reviews of test items during the test development process, empirical studies of ethnic and gender DIF are important to help minimize the possibility that one group of test takers is being disadvantaged or advantaged due to characteristics of test items other than the intended content knowledge or skills. DIF analyses identify items that have meaningful performance differences between specified subgroups after matching on ability (or

total test scores). Theoretically, if students in two comparison groups are matched across the ability continuum, but differences still show up in an item's correct response rates, then there may be factors, extraneous to the construct measured, that caused performance differences. Hence, the item may be potentially biased.

With regard to studies of test item bias or differential item functioning (DIF), there is currently no single industry standard for conducting these studies in terms of either methodology or criteria used for making decisions. The literature contains several proposed procedures, each different in the way they statistically handle item data and the indices they produce as criteria for identifying DIF items.

Unfortunately, while there is some consistency across procedures in identifying DIF items, this consistency is far from resulting in 100 percent agreement. As with any statistical procedure, one can expect a few extreme indices to result due to sampling fluctuations for the sample data used. Thus, a few items in any analysis might be identified as exhibiting DIF in one sample whereas in another sample of data, they would not. Additionally, our experience has found that 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 then would typically be compared to a random sample of white test takers across the entire state.

## **PROCEDURES**

**Samples.** Taking the above into account, the DIF analysis procedures and criteria put in place emphasized curriculum matching as a basis for making decisions and recommendations. Analyses were conducted using racial/ethnic (Asian Americans, African Americans, Hispanic Americans, and Native Americans) and gender groups and samples of whites from only schools that had minority groups.

For the spring assessments, adopting the national federal mandate, students could choose to identify themselves as belonging to one or more than one ethnic/racial group. Nevertheless, in Kansas, students who identified themselves as belonging to only one ethnic group defined most ethnic groups in the DIF analyses. For Native Americans, because in most cases the number of students belonging to only this group is relatively low (sometimes in the fifties), two DIF samples were used. One sample consisted of examinee records when students also considered

themselves as whites. This group is referred to as Native American Doubletons throughout this report. Another sample with smaller sizes consisted of those examinees who indicated that they belong to the Native American group only. This group is referred to as Native American Singletons in this paper. This second group, although small, was compared against whites to supplement the findings for the first group. For the ethnic comparisons, the base or reference group was whites and the focal group was one of the five minority groups mentioned above. In addition to these analyses, differential item performance analyses were made between females and males using examinee responses by each test form from the entire state. For the gender comparison, the reference group was the male group while the focal group was female. Only student responses with at least 90% of the test attempted were used in these analyses. The numbers of students in each gender and ethnic group used in the DIF analyses are given in Tables 2 through 4 and 6 through 8.

**Items.** The mathematics items and reading stems from alternate test forms at all grade levels were analyzed for DIF. In mathematics, each form at both grades 4 and 7 had 52 items whereas the grade 10 forms had 47 items. All the items were in the multiple choice format and thus were scored dichotomously. Prior to conducting the DIF analyses, item analyses were completed and decisions as to the adequacy of items made. In the end, one item in form 20 at grade 7 and, at grade 10, 2 items in form 76 and 1 item in form 99 were dropped as poorly functioning items.

In reading, each form had four text types with differing numbers of multiple yes/no items that were related to between 8 to 14 stems at each text type. Prior to the DIF analyses, and as was done for the mathematics forms, a few items were dropped as poorly functioning items. Each multiple yes/no item was scored dichotomously and these item scores were added together for each stem. Then, analyses for reading were carried out with stems identified as single items. Analyses at the stem level were done to avoid the dependency issue between related questions that belong to the same stem in the multiple yes/no format.

**Statistical Methods.** The main procedure used was the Mantel-Haenszel (MH) technique. In mathematics, the dichotomous MH procedure was employed to analyze the single correct multiple-choice items. The analyses were carried out at the total test level. The criteria used in these analyses were chi-squared extreme area probabilities ( $p$ ) less than 0.001 and absolute ETS delta values greater than 1.5. Items with negative delta values were seen to

disadvantage the focal group while positive values advantaged this group in comparison to the reference group.

Because the analyses for Reading were done at the stem rather than the multiple yes/no level, polychotomous scores for each comparison group were analyzed using the ordinal version of the generalized Mantel-Haenszel procedure. The hypothesis tested by this procedure was whether the ordinal distribution of item scores at each ability level was the same for both groups. The statistic used in this procedure was the chi-square with degrees of freedom 1. Because the chi-squared is a non-directional statistic, the Standardized Mean Difference (SMD) statistics which looks at the standardized differences in item means at each ability level was used to corroborate the evidences given by the chi-square value. Negative values of the SMD suggest that items disadvantaged a focal group while positive values suggest the opposite. Severe DIF is indicated if SMD values exceed 0.2 in absolute values while DIF detection with the chi-square is indicated when the extreme area probability is smaller than the alpha level of 0.001. The criteria used for the analyses summarized below relied on detection with chi-square extreme area probability ( $p < .001$ ) and corroborated by the SMD absolute values exceeding 0.2.

## **RESULTS**

A sample output for a mathematics DIF analysis at grade 10, form 30 is given in Table 1. From the DELTA\_E column of Table 1, items 3, 4, 11, 30, and 36 appeared to show DIF for the Asian versus White comparison. However, with a chi-squared extreme area probability less than 0.001, only item 3 appeared to show significant DIF. Item 3 seemed to disadvantage Asians.

Tables 2 through 4 gives a summary of items flagged by the Mantel-Haenszel procedure for each DIF comparison by form at each of the three grade levels. Table 2 shows that 3 items were flagged at grade 4, all with negative DIF. None of the items were flagged for the gender DIF comparisons but were flagged for the Asian, Blacks, and Hispanics comparisons. Table 3 showed 8 flagged items at grade 7, two of which were positive DIF and six were negative DIF. Most of these items were flagged for ethnic comparisons. Table 4 showed 8 flagged items at grade 10. One item was flagged showing positive DIF and the remaining exhibiting negative DIF. All these items were flagged for ethnic comparisons.

For DIF analyses in Reading, Table 5 gives a sample output of the ordinal generalized Mantel-Haenszel procedure. The output is for a gender comparison and is organized by text type for form 19 at grade 5. First, look at the column “p <” for items with chi-square extreme area

probabilities less than alpha level of 0.001. Then, look under the SMD column for absolute values exceeding 0.2. According to the criteria discussed earlier, no gender DIF stems existed for any text type of this grade 5 form.

Tables 6 through 8 summarize these DIF results for Reading at grade levels 5, 8, and 11. The tables are organized by comparison groups, text types, and form. In addition, a count of the number of items flagged is also given. Table 6 indicates that three stems in reading at grade 5 appeared to show some differential performance among examinee groups. Two exhibited negative DIF. The numbers are higher at higher grade levels as evidenced by Tables 7 and 8. Table 7 for grade 8 reading shows 13 items total flagged by the generalized Mantel-Haenszel procedure. Approximately half of these stems exhibited positive DIF. Table 8 for grade 11 reading shows 17 items flagged for the same reasons, 7 of which demonstrated positive DIF. It is uncertain at this point whether items flagged at the stem level exhibited differential performance due to potential cultural cues at the stem or at the yes/no item level.

#### **DIF SUMMARY**

Fewer items were found to exhibit DIF for mathematics than for reading. If items were to be dropped, this may have consequences on which test specification cells the items reside in and may have implications on the equating work that follows. In Reading, stems flagged may have been highlighted because of differential performance at the yes/no item level rather than potential cultural cues at the stem level. Potentially, this begged the question of whether to delete a whole stem that consisted of 4 to 5 yes/no questions or to abandon single yes/no questions related to the stem in question. A logical follow-up of these DIF analyses was a review by impacted groups who can answer some lingering questions of practical importance of these highlighted items.

Table 1. Example DIF output for Mathematics

MANTEL-HAENSZEL DIP ANALYSIS FOR 2000 mathematics grade 10 Kansas asian dif form 30 (Whites=6,Asian=2)

NUMBER OF ITEMS = 47 & CHK = .000

ITEM	P-6	PB-6	P-2	PB-2	P-6+2	PB-6+2	CHI-I	APLHA-I	DELTA-I	CHI-E	APLHA-E	DELTA-E
1	.71	.42	.64	.44	.70	.42	3.68	1.59	-1.09	4.85	1.69	-1.23
2	.40	.39	.43	.64	.41	.43	.02	.94	.13	.00	1.01	-.02
3	.71	.41	.57	.45	.69	.41	16.38	2.61	-2.25	14.42	2.39	-2.05
4	.46	.60	.39	.73	.45	.62	8.62	2.47	-2.13	5.29	1.97	-1.59
5	.34	.49	.45	.60	.36	.51	3.70	.61	1.15	3.34	.64	1.07
6	.34	.34	.39	.18	.34	.32	.03	.94	.14	.03	.94	.14
7	.37	.43	.44	.42	.38	.43	.67	.81	.51	.47	.84	.42
8	.32	.28	.31	.44	.32	.31	.57	1.23	-.48	.66	1.23	-.48
9	.42	.36	.46	.48	.42	.38	.16	.89	.26	.47	.84	.40
10	.42	.40	.42	.50	.42	.41	.05	1.08	-.17	.04	1.07	-.16
11	.56	.54	.67	.64	.58	.55	6.24	.49	1.67	5.46	.53	1.50
12	.40	.43	.46	.49	.41	.44	.64	.81	.48	.62	.82	.47
13	.60	.36	.58	.39	.60	.36	.09	1.09	-.20	.17	1.12	-.26
14	.61	.54	.60	.58	.61	.54	.24	1.17	-.36	.33	1.18	-.38
15	.63	.35	.65	.32	.64	.35	.01	.96	.11	.21	.89	.27
16	.29	.49	.33	.50	.30	.49	.00	.98	.05	.11	.90	.25
17	.57	.46	.67	.35	.58	.44	3.93	.63	1.08	4.97	.60	1.22
18	.48	.49	.48	.66	.48	.52	.52	1.23	-.48	.38	1.19	-.41
19	.36	.27	.33	.42	.35	.29	1.32	1.33	-.66	.85	1.25	-.52
20	.42	.52	.47	.58	.43	.53	.34	.85	.39	.36	.84	.40
21	.31	.53	.37	.58	.32	.54	.01	.94	.15	.00	.98	.06
22	.38	.54	.42	.59	.39	.55	.00	.97	.06	.00	.96	.09
23	.21	.27	.25	.21	.22	.26	.00	1.03	-.07	.01	1.06	-.13
24	.20	.28	.16	.45	.19	.30	3.41	1.91	-1.52	2.95	1.75	-1.32
25	.82	.36	.77	.44	.81	.37	2.96	1.63	-1.15	2.75	1.60	-1.11
26	.49	.52	.51	.47	.49	.51	.01	.99	.02	.03	.94	.15
27	.82	.24	.79	.25	.81	.24	.17	1.15	-.32	.70	1.27	-.56
28	.62	.50	.63	.40	.62	.48	.03	.94	.15	.08	.92	.19
29	.44	.39	.41	.44	.43	.39	1.27	1.32	-.65	.92	1.26	-.54
30	.56	.50	.46	.63	.55	.52	10.49	2.19	-1.84	9.53	2.18	-1.83
31	.23	.33	.23	.35	.23	.33	.02	1.06	-.15	.06	1.10	-.22
32	.31	.26	.36	.31	.31	.27	.55	.83	.43	.52	.84	.41
33	.39	.36	.41	.50	.39	.39	.02	.94	.13	.12	.90	.24
34	.19	.19	.17	.31	.19	.21	.56	1.29	-.60	.44	1.25	-.52
35	.79	.27	.75	.36	.78	.29	.56	1.24	-.51	.24	1.16	-.34
36	.56	.47	.69	.60	.58	.49	8.63	.46	1.85	9.37	.45	1.86
37	.68	.42	.68	.54	.68	.44	.01	1.00	-.00	.00	1.02	-.04
38	.28	.41	.36	.45	.29	.42	1.81	.71	.79	1.81	.72	.78
39	.50	.33	.52	.34	.50	.33	.04	.94	.15	.01	.96	.11
40	.40	.31	.39	.47	.40	.34	.03	1.06	-.15	.01	1.04	-.10
41	.51	.32	.56	.42	.52	.34	.76	.81	.49	.64	.83	.45
42	.65	.51	.67	.48	.65	.51	.01	.99	.02	.02	.95	.13
43	.49	.53	.46	.64	.49	.55	1.57	1.42	-.83	1.85	1.44	-.86
44	.45	.38	.55	.48	.46	.40	3.73	.63	1.07	3.62	.65	1.00
45	.40	.42	.48	.41	.41	.42	1.10	.78	.58	.68	.82	.46
46	.47	.46	.55	.51	.49	.47	.99	.78	.59	2.05	.70	.82
47	.63	.45	.73	.48	.64	.46	5.71	.54	1.44	5.30	.55	1.40

\*\* SUMMARY DATA \*\*

	MEAN	SD	CASES	KR-20
GROUP-6	22.157	9.113	721.	.892
GROUP-2	23.083	10.489	132.	.923
TOTAL	22.300	9.345	853.	.898

Table 2. Summary of differentially functioning Items for Grade 4 Math Forms

Form	Reference	DIF Group		DIF items	n
		n	Focal		
10	Male	3992	Female	4113	
	White	902	Asians	111	P1Q13-
			Blacks	493	
			Hispanics	344	
			Natives Dn	292	
			Natives Sn	81	
					<u>0 +; 1 -</u>
33	Male	4016	Female	4118	
	White	893	Asians	85	P3Q3-
			Blacks	530	
			Hispanics	342	
			Natives Dn	281	
			Natives Sn	86	
					<u>0 +; 1 -</u>
56	Male	3966	Female	4083	
	White	874	Asians	108	P4Q6-
			Blacks	466	
			Hispanics	358	
			Natives Dn	275	
			Natives Sn	91	
					<u>0 +; 1 -</u>
79	Male	3818	Female	3971	
	White	822	Asians	99	
			Blacks	460	
			Hispanics	352	
			Natives Dn	272	
			Natives Sn	86	
					<u>0 +; 0 -</u>
					<u>0 +; 3 -</u>

Table 3. Summary of differentially functioning Items for Grade 7 Math Forms

Form	Reference	DIF Group		DIF items	n
		n	Focal n		
20	Male	3999	Female 4155		
	White	961	Asians 122 Blacks 441 Hispanics 399 Natives Dn 385 Natives Sn 56		
					<u>0 +; 0 -</u>
43	Male	4041	Female 4113	P3Q5-	
	White	895	Asians 121 Blacks 436 Hispanics 389 Natives Dn 387 Natives Sn 78	P3Q1- P3Q1-	
					<u>0 +; 2 -</u>
66	Male	3976	Female 4126		
	White	982	Asians 142 Blacks 454 Hispanics 363 Natives Dn 347 Natives Sn 63	P1Q8-, P2Q10+, P3Q11+	
					<u>2 +; 1 -</u>
89	Male	3888	Female 4067		
	White	998	Asians 144 Blacks 426 Hispanics 360 Natives Dn 368 Natives Sn 67	P3Q11- P3Q1- P3Q1- P2Q12-	
					<u>0 +; 3 -</u>
					<u>2 +; 6 -</u>

Table 4. Summary of differentially functioning Items for Grade 10 Math Forms

Form	Reference	DIF Group		DIF items	n
		n	Focal n		
30	Male	3918	Female 3842		
	White	721	Asians 132 Blacks 366 Hispanics 293 Natives Dn 199 Natives Sn 54	P1Q3- P1Q1-, P1Q3-	
					<u>0 +; 2 -</u>
53	Male	3684	Female 3936		
	White	728	Asians 124 Blacks 349 Hispanics 285 Natives Dn 206 Natives Sn 54	P2Q6+, P3Q3-, P3Q4-, P3Q5-	
					<u>1 +; 3 -</u>
76	Male	3717	Female 3919		
	White	780	Asians 158 Blacks 367 Hispanics 305 Natives Dn 194 Natives Sn 56	P1Q1-	
					<u>0 +; 1 -</u>
99	Male	3744	Female 3773		
	White	711	Asians 130 Blacks 344 Hispanics 289 Natives Dn 204 Natives Sn 49	P1Q9-	
					<u>0 +; 1 -</u> <u>1 +; 7 -</u>

Table 5. Example DIF output for Reading

GENERALIZED MANTEL-HAENSZEL -- FORM 19 CONTENT Expository - Whites VS. Asians

Item	ORD MH	p <	SMD
1	1.917	0.166	0.083
2	1.522	0.217	0.086
3	0.000	0.984	0.006
4	0.052	0.820	0.017
5	8.586	0.003	-0.211
6	0.187	0.666	-0.031
7	0.091	0.763	0.020
8	1.143	0.285	-0.073
9	8.226	0.004	0.219
10	0.068	0.795	0.012
11	3.468	0.063	-0.127

FOCAL GROUP N = 120  
 REFERENCE GROUP N = 840  
 GENERALIZED MANTEL-HAENSZEL -- FORM 19 only CONTENT narrative - Whites VS. Asians

Item	ORD MH	p <	SMD
1	0.009	0.925	0.006
2	0.357	0.550	0.035
3	0.230	0.632	0.023
4	0.338	0.561	0.035
5	0.177	0.674	-0.022
6	0.207	0.649	0.027
7	0.944	0.331	0.046
8	0.197	0.657	-0.022
9	0.004	0.948	-0.002
10	2.206	0.138	-0.125

FOCAL GROUP N = 120  
 REFERENCE GROUP N = 840  
 GENERALIZED MANTEL-HAENSZEL -- FORM 19 CONTENT Persuasive - Whites VS. Asians

Item	ORD MH	p <	SMD
1	0.901	0.342	-0.059
2	8.659	0.003	-0.181
3	1.178	0.278	-0.050
4	0.595	0.441	-0.058
5	1.872	0.171	0.107
6	0.893	0.345	0.068
7	0.016	0.900	0.003
8	4.637	0.031	0.189
9	0.133	0.715	-0.019

FOCAL GROUP N = 120  
 REFERENCE GROUP N = 840  
 GENERALIZED MANTEL-HAENSZEL -- FORM 19 CONTENT Technical - Whites VS. Asians

Item	ORD MH	p <	SMD
1	0.730	0.393	0.031
2	0.113	0.737	-0.025
3	1.389	0.239	0.071
4	0.261	0.609	0.033
5	0.350	0.554	0.043
6	1.752	0.186	-0.087
7	0.757	0.384	-0.066

FOCAL GROUP N = 120  
 REFERENCE GROUP N = 840

Table 6. Summary of differentially functioning Items for Grade 5 Reading Forms

Form	DIF Group				Text Types				
	Reference	n	Focal	n	Expository	Narrative	Persuasive	Technical	All
19 Male	3730	Female	3803						
White	840	Asians	120						
		Blacks	474						
		Hispanics	321				S8+		
		Natives Dn	279						
		Natives Sn	65						
					0	0	1	0	1 +; 0 -
36 Male	3538	Female	3790						
White	834	Asians	107						
		Blacks	474				S4-		
		Hispanics	292						
		Natives Dn	305						
		Natives Sn	81						
					0	0	0	1	0 +; 1 -
53 Male	3753	Female	3717						
White	816	Asians	114						
		Blacks	459						
		Hispanics	291						
		Natives Dn	279						
		Natives Sn	63						
					0	0	0	0	0 +; 0 -
70 Male	3604	Female	3666						
White	785	Asians	118						
		Blacks	423				S6-		
		Hispanics	294						
		Natives Dn	290						
		Natives Sn	80						
					0	0	1	0	0 +; 1 -
					0	0	2	1	1 +; 2 -

Table 7. Summary of differentially functioning Items for Grade 8 Reading Forms

Form	Reference	DIF Group		Text Types					
		n	Focal	n	Expository	Narrative	Persuasive	Technical	All
28	Male	3809	Female	3935			S13-		
	White	872	Asians	136			S13-		
			Blacks	425					
			Hispanics	338		S12+	S12+		
			Natives Dn	268					
			Natives Sn	83					
					0	1	2	0	2 +; 1 -
45	Male	3759	Female	3923					
	White	854	Asians	122					
			Blacks	438	S13-			S5+	
			Hispanics	337					
			Natives Dn	275					
			Natives Sn	78					
					1	0	0	1	1 +; 1 -
62	Male	3754	Female	3842		S8-			
	White	848	Asians	156		S1-		S4-,S10-	
			Blacks	434			S10-		
			Hispanics	325					
			Natives Dn	237					
			Natives Sn	63					
					0	2	1	2	0 +; 5 -
80	Male	3692	Female	3942					
	White	823	Asians	125	S6+				
			Blacks	411	S10+				
			Hispanics	332	S10+	S9+			
			Natives Dn	250					
			Natives Sn	70					
					2	1	0	0	3 +; 0 -
					3	4	3	3	6 +; 7 -

Table 8. Summary of differentially functioning Items for Grade 11 Reading Forms

Form	Reference	DIF Group		Text Types					
		n	Focal	n	Expository	Narrative	Persuasive	Technical	All
37	Male	3237	Female	3531	S12-			S11-	
	White	603	Asians	155				S11-	
			Blacks	291	S12-			S11-	
			Hispanics	241				S7+	
			Natives Dn	152					
			Natives Sn	39					
					1	0	0	2	1 +; 2 -
54	Male	3219	Female	3496					
	White	630	Asians	126				S11-	
			Blacks	271		S9+			
			Hispanics	252		S9+,S11-			
			Natives Dn	150				S2+	
			Natives Sn	53					
					0	2	0	2	2 +; 2 -
71	Male	3221	Female	3411				S4-,S9+	
	White	557	Asians	118					
			Blacks	258	S10-	S4+	S13+		
			Hispanics	226		S8-		S9+	
			Natives Dn	132					
			Natives Sn	39					
					1	2	1	2	3 +; 3 -
98	Male	3175	Female	3317	S6+		S9-		
	White	588	Asians	125	S6+,S11-				
			Blacks	272	S11-			S11-	
			Hispanics	196					
			Natives Dn	168					
			Natives Sn	42					
					2	0	1	1	1 +; 3 -
					4	4	2	7	7 +; 10 -

## **IMPACTED GROUP RECOMMENDATION**

Dropping an item based solely on statistical findings is ill advised and is a practice that should be avoided. Statistical reasons for dropping items may not necessarily mean these items were biased against the impacted groups. There could exist reasons totally unrelated to item bias. Therefore, items that exhibited negative DIF were reviewed by Equity panel members of the specific impacted group. These groups explored why these items exhibited negative DIF, whether or not these items appeared to be actually biased, and whether or not any of the items should be dropped or revised. Table 9 summarizes the recommendations given by the impacted groups review panel with respect to the negatively functioning items in mathematics and reading. One prompt was dropped from the writing assessment.

Based on these recommendations, some of the negative DIF items were dropped from future administration in mathematics and reading. Revised items of comparable specifications, difficulty, and response time would be used to replace these items so that equating relationships between forms derived under certain test specifications and length considerations may still apply.

## **CONCLUSION**

Statistical analyses of the items across gender and racial groups identified several items that were differentially functioning. Because statistical reasons alone were not sufficient for definitive identification of biased items, a group of special reviewers from each of the impacted groups were invited to review, critique, and make recommendations regarding the negatively functioning items. The group recommended dropping a subset of the items identified as negative DIF items. Negatively functioning items were dropped as recommended and similar items would replace them in future administration of reading and mathematics assessments.

Table 9. Recommendations made by impacted group on negatively functioning items

Subject	Grade	Form	Part	Item	Impacted Group	Recommendations
Mathematics	4	10	1	13	Asians	retain item
		33	3	3	Blacks	retain item
		56	4	6	Hispanics	retain item
	7	43	3	1	Blacks & Hispanics	drop item
			3	5	Female	drop item
		66	1	8	Asians	retain item
		89	2	12	Native American	retain item
		3	1	Blacks & Hispanics	drop item	
	10	30	1	1	Asians	retain item
			1	3	Asians & Blacks	drop item
		53	3	3	Asians	retain item
			3	4	Asians	retain item
			3	5	Asians	retain item
		76	1	1	Hispanics	retain item
		99	1	9	Blacks	retain item
Reading	5	19	Persuasive	stem 2	Asians	retain stem
		36	Technical	stem 4	Blacks	retain stem
		53	Expository	stem 2	Asians	retain stem
	8	28	Persuasive	stem 13	Asians, Blacks, & Hispanics	drop stem
		28	Technical	stem 4	Hispanic	retain stem
		45	Persuasive	stem 11	Asians	retain stem
		45	Technical	stem 10	Asians	retain stem
		62	Persuasive	stem 8	Blacks	retain stem
		62	Persuasive	stem 10	Blacks	retain stem
		62	Technical	stem 10	Asians	retain stem
	11	37	Expository	stem 12	Female & Blacks	drop stem
		37	Technical	stem 11	Female, Asians, Blacks, & Hispanics	drop stem
		54	Narrative	stem 11	Asians & Hispanics	retain stem
		71	Technical	stem 4	Female	retain stem
		98	Persuasive	stem 9	Female	retain stem
		98	Expository	stem 11	Asians & Blacks	retain stem

## Test Equating

An important property of test equating is equity (Kolen and Brennan, 1995; Lord, 1980). Simply put, this property requires that it should be a matter of indifference to examinees at every ability level whether they have to respond to form X or form Y of the test. Two other important properties are symmetry and same test specifications (Kolen and Brennan, 1995). A prior section described procedures used to establish equivalent tests based on specifications. Without these three properties or assumptions, a test form cannot be said to be satisfactorily equated, in the sense of the term, even if sophisticated methods were used.

With newly developed Kansas assessments in Mathematics and Reading, scores from parallel test forms administered to different groups needed to be equated to ensure equitability of scores to every examinee. This section summarizes the description of test forms, the design and the methods used, the decision taken, and discusses issues in equating multiple forms of the Kansas Assessments. The main purpose of equating was to ensure equity to examinees at every ability level in the state of Kansas. Another purpose, and an important byproduct, of this equating was a common metric for expressing equitable examinee scores.

For adequacy of an equating design, sufficient groundwork on test development was needed to ensure that test forms were classically parallel. Good and defensible test development practices that ensure the same number of items on each form with the same test specifications had to be followed every step of the test development process. Data gathering procedures, that inform test item review and accurately describe item properties for proper assembly of test forms, are also crucial.

### PROCEDURES

**Test Forms Preparation.** As a result of pilot testing in the Fall semester 1999, four parallel forms in mathematics were developed at three grade levels; fourth, seventh, and tenth. The pilot testing design sampled from volunteering Kansas schools exposing not more than 25% of the item pool at any site. On the order of 200 to 350 student responses per item were captured for the pilot analyses. Using both IRT modified three-parameter model (with  $c=0.1$ ) and classic indices of difficulty and discrimination, poorly performing items were, for the most part, abandoned. Any apparent poorly functioning item retained was done so based on a judgment that the item was an appropriate (valid) measure of important content, but students were performing poorly on the item due to lack of instructional opportunity to learn the content.

In Mathematics four independent forms (no common items) were developed with equal number of items between forms representing each standard, benchmark, and indicator combination as dictated by the percentages agreed upon by KSDE. Items were randomly assigned to forms, then forms adjusted to assure adequate/proportionate indicator sampling. For example, the percentage of items measuring knowledge and application indicators were 50% at grades 4 and 7. The number of items at both the fourth and seventh grade was 52 while the number at tenth grade was 47. All items in mathematics were multiple choice in format.

In Reading, four forms each containing four types of reading passages were developed at three grade levels; fifth, eighth, and eleventh. The four types of passages were Expository, Persuasive, Technical, and Narrative. For each form, passages that were easy were teamed with those that were difficult to balance total difficulty as much as possible between forms. In addition, the lengths of passages and test questions were also controlled to minimize the structural differences between forms. All items in reading were multiple yes/no in format. This is a format where students are required to choose yes or no (mark correct or incorrect) to a series of choice options that relate to a stem question. In reading, there were between a hundred fifty to two hundred multiple yes/no items (choices) and between 40 to 56 stems per form at each grade level.

**Equating Design.** The data collection method for the equating was the Random Groups design. The design was implemented by spiraling four different forms at each grade level in both reading and mathematics during test administration in Kansas classrooms. With approximately 30,000 – 32,000 regular education students taking the test at each grade level, about 7500 - 8000 students took each of the four test forms. This represents a number that is more than adequate for a random groups equating.

Table 10 shows the percentages of students taking alternate forms of mathematics and reading assessments across schools in Kansas. For the values in Table 10, percentages of students taking each form were obtained for each school and these percentages were summarized into means and standard deviations. In addition, the table provides percentages of students taking each form by gender, race, and educational classifications. At every grade level, while mean percentages associated with the last form were slightly lower than those for the first listed form, percentages based on demographic information supports the equivalence of groups obtained

through this data collection design. In other word, data in Table 10 strongly suggest the equivalence of the groups responding to each form, at all grades, for each tested content area.

**Statistical Procedures.** Both classical and IRT test equating were examined. For the classical equating, linear and equi-percentile methods were investigated. For the IRT test equating, the IRT observed score equating was studied. Each form was separately calibrated under the 3-parameter model in mathematics and the generalized partial credit model in reading and observed scores frequency distributions were obtained by summing the compound binomial (or multinomial) distribution across all values of theta. Then, equi-percentile method of equating was used on the obtained observed scores frequency distributions.

In mathematics, examinee scores were equated at the process skills (knowledge and application) and the total score level and expressed in the percent correct metric. In Reading, equating was conducted at each of the four text-type levels and at the total percent correct score level.

**Equating Criteria.** Because both classical and IRT test equating were examined, comparison between several competing methods were possible. These equating methods had to be reviewed and deliberate decisions had to be made as to which produced the most reasonable conversion for students in the state of Kansas. To assist in selecting the best equating conversion, the following criteria in the order listed were used.

1. Fidelity to the equated data

An equating conversion that provided the closest approximation to the base form distributional moments gave the best score transformation. When there were no difference in form difficulty, the distributional moments of the equated scores would 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 were similar to each other, the simpler conversion was used. The standard error for the equi-percentile equating at each score level was used to judge the degree of similarity between equating conversions.

### 4. Smoothed distributional properties

An equating conversion that provided fewer gaps at the top or bottom of the percent correct scale was chosen.

These criteria were used simultaneously, favoring methods meeting all or most criteria.

Table 10. Percentages of Students in Kansas Schools Taking Alternate Forms

Subject	Grade	N of schools	Form	Statistics		Gender		Race		Education	
				M	SD	Female	Male	White	Minority	Regular	Sped
Math	4	843	10	25.6	4.5	50.7	49.3	80.0	20.0	89.4	10.6
			33	25.3	3.9	50.5	49.5	80.2	19.8	90.4	9.6
			56	25.0	4.3	50.6	49.4	80.1	19.9	90.9	9.1
			79	24.1	4.4	50.8	49.2	80.1	19.9	90.4	9.6
			All			50.6	49.4	80.1	19.9	90.3	9.7
	7	508	20	25.1	4.7	50.9	49.1	82.8	17.2	90.4	9.6
			43	25.6	6.3	50.4	49.6	81.9	18.1	91.0	9.0
			66	24.7	4.2	50.8	49.2	82.1	17.9	91.2	8.8
			89	24.6	6.2	51.0	49.0	81.9	18.1	90.9	9.1
			All			50.8	49.2	82.2	17.8	90.9	9.1
	10	387	30	25.3	4.3	50.4	49.6	83.5	16.5	92.7	7.3
			53	25.0	5.0	51.6	48.4	83.7	16.3	92.1	7.9
			76	25.1	5.3	51.2	48.8	83.0	17.0	92.6	7.4
			99	24.6	5.4	50.3	49.7	83.1	16.9	92.5	7.5
			All			50.9	49.1	83.3	16.7	92.5	7.5
Reading	5	819	19	26.1	6.4	50.4	49.6	83.7	16.3	90.4	9.6
			36	24.7	4.8	51.5	48.5	83.1	16.9	90.0	10.0
			53	24.8	5.3	49.6	50.4	84.2	15.8	90.8	9.2
			70	24.4	4.9	50.1	49.9	83.9	16.1	90.6	9.4
			All			50.4	49.6	83.7	16.3	90.5	9.5
	8	470	28	26.0	6.6	50.7	49.3	82.3	17.7	90.8	9.2
			45	25.1	4.4	51.0	49.0	81.9	18.1	91.1	8.9
			62	24.5	4.2	50.3	49.7	81.8	18.2	90.7	9.3
			80	24.4	7.1	51.8	48.2	82.5	17.5	91.6	8.4
			All			51.0	49.0	82.1	17.9	91.1	8.9
	11	353	37	25.7	4.8	51.8	48.2	80.3	19.7	93.8	6.2
			54	25.5	5.6	51.7	48.3	80.8	19.2	94.1	5.9
			71	24.8	5.5	51.3	48.7	81.3	18.7	94.0	6.0
			98	24.1	4.4	51.2	48.8	80.9	19.1	93.6	6.4
			All			51.5	48.5	80.8	19.2	93.9	6.1

## RESULTS

Table 11 shows a descriptive summary of the equating samples obtained in mathematics. The bolded numbers in the table describe the characteristics of the base form at each grade level. These base forms were primarily chosen for their complete test specifications as intended by KSDE. That is, these forms had no items dropped after administrations due to printing errors, item analyses, or differential item functioning analyses. When more than one test form had complete items as intended such as at grade 4, the form that had medium mean total scores among the forms was chosen as the base. Table 11 also shows that all the scales had sufficient reliability for equating purposes. In addition, these reliability measures were not different from one form to the other.

Table 11. Descriptive Statistics for Equating Samples for Mathematics Assessments by Test Form and Grade Level

Grade	Form	N	Knowledge				Application				Total			
			n	M	SD	Rxx'	n	M	SD	Rxx'	n	M	SD	Rxx'
4	10	7564	26	15.36	4.63	0.79	26	13.37	4.31	0.72	52	28.73	8.37	0.86
	33	7562	26	14.77	4.53	0.76	26	12.95	4.50	0.75	52	27.72	8.41	0.86
	56	7379	26	15.67	4.79	0.80	26	12.50	4.60	0.75	52	28.17	8.82	0.87
	<b>79</b>	<b>7315</b>	<b>26</b>	<b>15.77</b>	<b>4.61</b>	<b>0.77</b>	<b>26</b>	<b>12.70</b>	<b>4.72</b>	<b>0.78</b>	<b>52</b>	<b>28.28</b>	<b>8.79</b>	<b>0.87</b>
7	20	7409	26	13.38	5.11	0.81	24	12.89	4.72	0.79	50	26.28	9.28	0.89
	43	7594	25	12.90	5.36	0.84	25	12.81	4.26	0.74	50	25.71	9.04	0.88
	<b>66</b>	<b>7433</b>	<b>26</b>	<b>13.36</b>	<b>5.35</b>	<b>0.83</b>	<b>26</b>	<b>13.20</b>	<b>5.09</b>	<b>0.81</b>	<b>52</b>	<b>26.56</b>	<b>9.93</b>	<b>0.90</b>
	89	7343	25	11.92	4.91	0.80	26	14.46	4.85	0.79	51	26.38	9.20	0.89
10	30	7394	17	7.29	3.56	0.74	29	13.16	5.53	0.82	46	20.45	8.50	0.88
	<b>53</b>	<b>7299</b>	<b>17</b>	<b>8.10</b>	<b>3.79</b>	<b>0.77</b>	<b>30</b>	<b>13.92</b>	<b>5.20</b>	<b>0.78</b>	<b>47</b>	<b>22.02</b>	<b>8.37</b>	<b>0.87</b>
	76	7096	17	7.33	3.41	0.71	27	13.09	5.03	0.78	44	20.41	7.82	0.85
	99	7108	17	7.17	3.94	0.84	29	14.30	5.74	0.78	46	21.47	9.09	0.90

Table 12 shows a descriptive summary of the equating samples in reading. In addition to the previous statistics, Table 12 shows the maximum score points that can be obtained for each text type and form. Again, bolded numbers in the table describe characteristics of the base forms. Because reading was not based on strict percentages of items within a set of specification cells, selection of base forms were made solely on the basis of medium percent correct on the Total scale. While the Total scale had reliabilities in the .90s, the text types in some cases had reliability measures as low as the .60s. However, because most text types had reliability indices in the high .70s and low .80s, equating at the text type level did not appear objectionable.



Table 12. Descriptive Statistics for Equating Samples of Reading Assessments by Test Form, Text Types, and Grade Level

Grade	Form	N	Expository					Persuasive					Technical					Narrative					Total				
			n	max	M	SD	Rxx'	n	max	M	SD	Rxx'	n	max	M	SD	Rxx'	n	max	M	SD	Rxx'	n	max	M	SD	Rxx'
5	19	<b>7838</b>	<b>11</b>	<b>44</b>	<b>35.46</b>	<b>6.54</b>	<b>0.83</b>	<b>9</b>	<b>37</b>	<b>29.17</b>	<b>3.96</b>	<b>0.62</b>	<b>7</b>	<b>30</b>	<b>25.03</b>	<b>3.54</b>	<b>0.67</b>	<b>10</b>	<b>39</b>	<b>33.53</b>	<b>4.81</b>	<b>0.79</b>	<b>37</b>	<b>150</b>	<b>82.21</b>	<b>10.18</b>	<b>0.91</b>
	36	7869	10	40	31.66	4.89	0.73	10	40	33.90	4.38	0.76	8	32	25.48	4.84	0.77	10	39	33.62	4.92	0.80	38	151	82.43	10.33	0.92
	53	7561	10	40	33.11	4.54	0.70	10	40	33.88	4.47	0.74	9	37	30.64	4.18	0.70	10	38	31.24	4.41	0.73	39	155	83.12	9.13	0.90
	70	6929	10	40	32.02	5.16	0.74	10	40	33.54	4.27	0.70	9	37	30.36	5.01	0.73	10	42	34.43	5.52	0.80	39	159	81.98	10.47	0.91
8	28	7491	11	45	36.95	5.63	0.78	12	48	38.13	6.00	0.79	12	50	42.05	5.74	0.79	14	56	46.09	7.09	0.83	49	199	81.99	10.27	0.93
	45	7629	12	49	40.53	5.37	0.77	14	58	45.50	7.35	0.82	11	46	39.09	4.63	0.72	12	49	39.60	6.31	0.81	49	202	81.74	9.66	0.93
	<b>62</b>	<b>7461</b>	<b>13</b>	<b>54</b>	<b>43.04</b>	<b>6.14</b>	<b>0.78</b>	<b>11</b>	<b>42</b>	<b>34.39</b>	<b>4.56</b>	<b>0.67</b>	<b>11</b>	<b>49</b>	<b>41.54</b>	<b>5.74</b>	<b>0.79</b>	<b>10</b>	<b>43</b>	<b>35.88</b>	<b>4.80</b>	<b>0.71</b>	<b>45</b>	<b>188</b>	<b>82.45</b>	<b>9.05</b>	<b>0.91</b>
	80	7810	13	54	43.62	6.58	0.80	11	42	33.45	4.96	0.71	12	50	43.25	6.20	0.82	10	40	34.21	5.01	0.79	46	186	83.11	10.04	0.93
11	<b>37</b>	<b>6478</b>	<b>12</b>	<b>49</b>	<b>39.11</b>	<b>6.24</b>	<b>0.78</b>	<b>11</b>	<b>46</b>	<b>33.72</b>	<b>6.29</b>	<b>0.77</b>	<b>12</b>	<b>51</b>	<b>43.20</b>	<b>5.87</b>	<b>0.80</b>	<b>13</b>	<b>54</b>	<b>45.79</b>	<b>6.37</b>	<b>0.82</b>	<b>48</b>	<b>200</b>	<b>80.66</b>	<b>10.51</b>	<b>0.93</b>
	54	5934	12	52	44.48	6.41	0.83	11	51	41.06	5.89	0.78	13	55	46.66	6.52	0.79	14	59	48.87	6.81	0.81	50	217	83.43	10.23	0.93
	71	6304	10	47	35.35	6.04	0.76	13	55	43.72	6.92	0.82	13	56	48.83	5.87	0.80	11	48	38.21	6.66	0.74	47	206	80.38	10.36	0.92
	98	6201	13	55	41.93	7.08	0.78	12	50	39.47	6.36	0.78	10	42	34.16	4.81	0.71	12	55	44.55	6.87	0.82	47	202	79.37	10.35	0.92

max refers to the maximum possible points

An example of selected parts of an equating output is given in Table 13. The base form for this equating is form 79, the form to be equated is form 56, and the scale to be equated is the Mathematics knowledge sub-scale with 26 items. In exercising the criteria listed in the previous paragraphs, the four moments of the equated scores from several competing methods were compared to that of the base form. Although moments from IRT observed scores are not typically computed, the moments for this method was also calculated to facilitate comparison. Table 13 shows that the linear and equi-percentile equating appeared to give equated scores with the closest approximation to the first two moments of the base form distribution. The IRT observed scores equating provided moments most dissimilar to that of the base distribution.

Table 13. Moments of the Equated Form 56  
by Equating Method

Test Form/Method	Mean	SD	Skewness	Kurtosis
Old Form: 4K79.DAT n=7315; New Form: 4K56.DAT n=7379;				
Raw Scores				
4K79.DAT	16.0135	4.4448	-0.2189	2.5051
4K56.DAT	16.0995	4.6200	-0.2322	2.4167
4K56.DAT equated to 4K79.DAT				
unsmoo	16.0134	4.4327	-0.2138	2.4782
s=0.01	16.0144	4.4402	-0.2141	2.4817
s=0.05	16.0145	4.4406	-0.2138	2.4795
s=0.10	16.0147	4.4430	-0.2136	2.4724
s=0.20	16.0160	4.4496	-0.2139	2.4471
s=0.30	16.0176	4.4537	-0.2167	2.4338
s=0.50	16.0216	4.4589	-0.2255	2.4261
s=0.75	16.0258	4.4615	-0.2313	2.4259
s=1.00	16.0258	4.4615	-0.2313	2.4259
linear	16.0135	4.4448	-0.2322	2.4167
irt obs	15.9711	4.4347	-0.2176	2.4638

Because the mean differences between base and form 56 was .08, Figure 1 suggests that most equating methods considered provided reasonable conversions at all score levels where the majority of the data congregated (between raw scores of 7 and 24). Figure 1 also shows that the linear equating offered similar conversions to that of the equi-percentile equating at these data score congregations.

Figure 1. knowledge Grade 4 Form 56 equated to base form 79

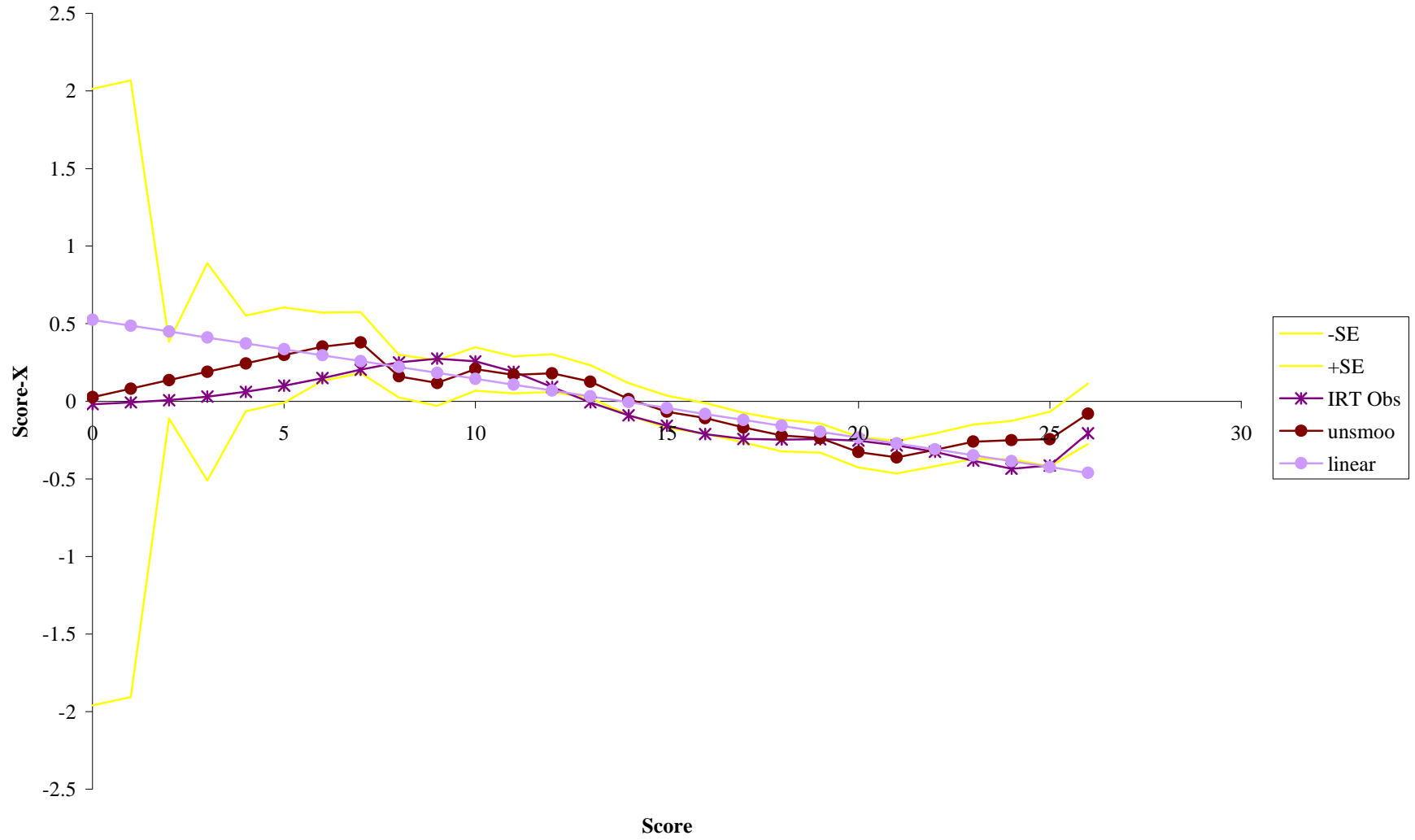


Table 14 shows related conversion tables for all competing methods. It appeared that most conversions showed reasonable progression of equated scores at the top of the raw score scale. Table 15 shows the same conversion table when transformed onto the percent correct metric. In this percent correct metric, the difference in distributional smoothness, at the top of the scale, between methods was also minimal. Given the multiple criteria considered, it appeared that the linear method gave the most parsimonious yet reasonable conversion.

Table 14. Conversion Table for Competing Methods

Raw	IRT	Obs	unsmoo	s=0.01	s=0.05	s=0.10	s=0.20	s=0.30	s=0.50	s=0.75	s=1.00	linear	frequency
0	-0.018	0.027	0.025	0.029	0.028	0.031	0.033	0.032	0.030	0.030	0.030	0.525	
1	0.992	1.081	1.074	1.086	1.084	1.093	1.098	1.095	1.089	1.089	1.089	1.487	
2	2.007	2.136	2.124	2.143	2.141	2.155	2.163	2.158	2.148	2.148	2.148	2.449	
3	3.030	3.190	3.173	3.201	3.197	3.217	3.228	3.221	3.208	3.208	3.208	3.411	8
4	4.060	4.244	4.222	4.258	4.253	4.279	4.293	4.284	4.267	4.267	4.267	4.373	20
5	5.101	5.298	5.278	5.312	5.301	5.329	5.344	5.331	5.314	5.314	5.314	5.335	48
6	6.150	6.351	6.331	6.295	6.278	6.299	6.308	6.294	6.279	6.279	6.279	6.297	80
7	7.204	7.379	7.299	7.257	7.252	7.269	7.272	7.257	7.244	7.244	7.244	7.259	121
8	8.251	8.161	8.185	8.205	8.225	8.238	8.235	8.219	8.209	8.209	8.209	8.221	144
9	9.275	9.118	9.135	9.173	9.202	9.206	9.198	9.182	9.174	9.174	9.174	9.183	241
10	10.257	10.208	10.177	10.173	10.183	10.172	10.160	10.145	10.139	10.139	10.139	10.145	298
11	11.191	11.170	11.182	11.174	11.160	11.136	11.121	11.107	11.104	11.104	11.104	11.107	340
12	12.093	12.181	12.171	12.156	12.128	12.095	12.079	12.069	12.069	12.069	12.069	12.070	440
13	12.993	13.126	13.119	13.105	13.081	13.049	13.036	13.031	13.034	13.034	13.034	13.032	435
14	13.910	14.014	14.019	14.027	14.022	13.999	13.991	13.993	13.999	13.999	13.999	13.994	498
15	14.843	14.932	14.936	14.949	14.957	14.947	14.946	14.955	14.964	14.964	14.964	14.956	518
16	15.789	15.891	15.887	15.886	15.893	15.895	15.902	15.918	15.929	15.929	15.929	15.918	589
17	16.758	16.831	16.830	16.832	16.834	16.847	16.860	16.881	16.894	16.894	16.894	16.880	555
18	17.753	17.779	17.789	17.785	17.782	17.804	17.821	17.846	17.859	17.859	17.859	17.842	594
19	18.756	18.763	18.752	18.738	18.738	18.768	18.788	18.812	18.824	18.824	18.824	18.804	529
20	19.746	19.672	19.683	19.691	19.707	19.740	19.759	19.779	19.789	19.789	19.789	19.766	499
21	20.716	20.639	20.654	20.673	20.695	20.723	20.736	20.748	20.754	20.754	20.754	20.728	455
22	21.674	21.687	21.696	21.695	21.702	21.715	21.719	21.719	21.719	21.719	21.719	21.690	396
23	22.617	22.740	22.747	22.734	22.724	22.714	22.705	22.690	22.685	22.685	22.685	22.652	283
24	23.565	23.750	23.766	23.765	23.751	23.717	23.694	23.662	23.650	23.650	23.650	23.614	173
25	24.586	24.756	24.787	24.793	24.780	24.722	24.684	24.635	24.615	24.615	24.615	24.577	91
26	25.793	25.919	25.927	25.928	25.923	25.901	25.886	25.864	25.851	25.851	25.851	25.539	24

## EQUATING DECISIONS

Each of the twenty-seven equating analysis in mathematics and the forty-four equatings in reading was performed and subjected to the criteria previously listed. The selected equating for each scale in reading and mathematics is summarized in this section of the report. For all

scales, a zero on the raw score scale converts to zero regardless of equating method used. In addition, equated scores that were negative in value were set to the minimum score of zero. For student reports, the top equated scores were set to the top scores on the base form. For building reports, because mean scores rarely consists of perfect scores from all students, the top equated scores were used as is, without setting them to the top scores on the base form. Furthermore, setting top scores at the building level may change the moments of the equated distributions.

Table 15. Conversion Table for Competing Methods Expressed in Percent Correct Metric

Raw	IRT Obs	unsmoo	s=0.01	s=0.05	s=0.10	s=0.20	s=0.30	s=0.50	s=0.75	s=1.00	Linear
0	-0.07	0.10	0.10	0.11	0.11	0.12	0.13	0.12	0.12	0.12	2.02
1	3.82	4.16	4.13	4.18	4.17	4.20	4.22	4.21	4.19	4.19	5.72
2	7.72	8.22	8.17	8.24	8.23	8.29	8.32	8.30	8.26	8.26	9.42
3	11.65	12.27	12.20	12.31	12.30	12.37	12.42	12.39	12.34	12.34	13.12
4	15.62	16.32	16.24	16.38	16.36	16.46	16.51	16.48	16.41	16.41	16.82
5	19.62	20.38	20.30	20.43	20.39	20.50	20.55	20.50	20.44	20.44	20.52
6	23.65	24.43	24.35	24.21	24.15	24.23	24.26	24.21	24.15	24.15	24.22
7	27.71	28.38	28.07	27.91	27.89	27.96	27.97	27.91	27.86	27.86	27.92
8	31.74	31.39	31.48	31.56	31.63	31.68	31.67	31.61	31.57	31.57	31.62
9	35.67	35.07	35.13	35.28	35.39	35.41	35.38	35.32	35.28	35.28	35.32
10	39.45	39.26	39.14	39.13	39.17	39.12	39.08	39.02	39.00	39.00	39.02
11	43.04	42.96	43.01	42.98	42.92	42.83	42.77	42.72	42.71	42.71	42.72
12	46.51	46.85	46.81	46.75	46.65	46.52	46.46	46.42	46.42	46.42	46.42
13	49.97	50.48	50.46	50.40	50.31	50.19	50.14	50.12	50.13	50.13	50.12
14	53.50	53.90	53.92	53.95	53.93	53.84	53.81	53.82	53.84	53.84	53.82
15	57.09	57.43	57.45	57.50	57.53	57.49	57.48	57.52	57.55	57.55	57.52
16	60.73	61.12	61.10	61.10	61.13	61.13	61.16	61.22	61.27	61.27	61.22
17	64.46	64.73	64.73	64.74	64.75	64.80	64.85	64.93	64.98	64.98	64.92
18	68.28	68.38	68.42	68.40	68.39	68.48	68.54	68.64	68.69	68.69	68.62
19	72.14	72.17	72.12	72.07	72.07	72.18	72.26	72.35	72.40	72.40	72.32
20	75.95	75.66	75.70	75.73	75.80	75.92	76.00	76.07	76.11	76.11	76.02
21	79.68	79.38	79.44	79.51	79.60	79.70	79.75	79.80	79.82	79.82	79.72
22	83.36	83.41	83.45	83.44	83.47	83.52	83.53	83.53	83.53	83.53	83.42
23	86.99	87.46	87.49	87.44	87.40	87.36	87.33	87.27	87.25	87.25	87.12
24	90.64	91.35	91.41	91.40	91.35	91.22	91.13	91.01	90.96	90.96	90.82
25	94.56	95.22	95.33	95.36	95.31	95.08	94.94	94.75	94.67	94.67	94.53
26	99.20	99.69	99.72	99.72	99.70	99.62	99.56	99.48	99.43	99.43	98.23

Table 16 shows the type of equating decisions taken for mathematics. Except for the linear method, all methods required the use of conversion tables. These conversion tables are not provided here to conserve space. When linear equating was chosen, Table 16 also provides the equating constants for the conversion function. After comparing several methods, most equating

method chosen were the unsmoothed equi-percentile equating. Conversion tables related to this method were saved and used for the equating subroutines implemented for the 2000 assessment year. Linear equating and smoothed equi-percentile equating methods were chosen five times each. IRT observed scores equating were chosen for two knowledge subscales at grade 7.

Table 16. Summary of Equating Decisions for Mathematics

Grade	Form		Scale <sup>a</sup>	Decision	Equating	
	Base	Equated			Slope	Intercept
4	79	10	Knowledge	Unsmoothed Equipercentile	-	-
			Application	Unsmoothed Equipercentile	-	-
			Total	Unsmoothed Equipercentile	-	-
	79	33	Knowledge	Unsmoothed Equipercentile	-	-
			Application	Unsmoothed Equipercentile	-	-
			Total	Unsmoothed Equipercentile	-	-
	79	56	Knowledge	Linear Method	0.962076923	0.525
			Application	Unsmoothed Equipercentile	-	-
			Total	Linear Method	0.997711538	0.201
7	66	20	Knowledge	IRT Observed Scores	-	-
			Application	Unsmoothed Equipercentile	-	-
			Total	Linear Method	1.0738	-1.723
	66	43	Knowledge	IRT Observed Scores	-	-
			Application	Unsmoothed Equipercentile	-	-
			Total	Linear Method	1.095326531	-1.581
	66	89	Knowledge	Smoothed Equipercentile (s=1.00)	-	-
			Application	Unsmoothed Equipercentile	-	-
			Total	Smoothed Equipercentile (s=0.75)	-	-
10	53	30	Knowledge	Unsmoothed Equipercentile	-	-
			Application	Unsmoothed Equipercentile	-	-
			Total	Unsmoothed Equipercentile	-	-
	53	76	Knowledge	Unsmoothed Equipercentile	-	-
			Application	Smoothed Equipercentile (s=0.01)	-	-
			Total	Linear Method	1.071113636	0.183
	53	99	Knowledge	Unsmoothed Equipercentile	-	-
			Application	Smoothed Equipercentile (s=0.01)	-	-
			Total	Smoothed Equipercentile (s=0.10)	-	-

<sup>a</sup> All scales were equated at the raw score level and then expressed in the percent correct metric.

Table 17 shows the types of equating chosen for Reading. Because the narrative text for forms 19 and 36 at grade 5 was the same type, no equating was conducted for narrative form 36. The methods chosen for Reading were unsmoothed equi-percentile, unsmoothed equi-percentile with line segments, smoothed equi-percentile, and linear equating. The unsmoothed equi-percentile equating with line segments were used when no method produced smooth progression of equated scores at the top of the score scale and trends at the top of the equi-percentile could use some smoothing. The line segments in these cases were derived by fitting a least squares regression line for the unsmoothed equi-percentile equating at the top of the score scale.

## **SUMMARY AND DISCUSSION**

The reliabilities for Mathematics process skills (Knowledge and Application) and Total score appeared acceptable for equating purposes. Although two text types at grade 5 in Reading had reliability values below the .70 cut-off, all other text types were adequate for equating. Therefore, equating of reading at the text type level did not appear objectionable. In addition, data collected from the spring 2000 administration show that the groups formed for the equating work appeared to be random.

Numerous equating methods were considered including IRT and classical methods. The 3-parameter model for Mathematics and the generalized partial credit model for Reading were used to produce score distributions for IRT observed scores equating. Because several competing methods were considered, a few criteria were used to select the method that would provide for the most equitable scores for Kansas examinees. Methods that best fit the data through the criteria listed were chosen.

In this equating, the test forms equated were designed to be equal in regards to the test specifications laid out by KSDE. However, several items were lost after spring 2000 administration of the tests. A few items were dropped after item analyses, a couple of items were lost due to printing errors, and some items were dropped due to DIF. This item-loss created empty or under-measured test specification cells in at least one of the equated forms. Because the equating work completed was based on certain fixed length response time and test specification assumption, these items would have to be replaced with like items for future administration so that the equating relationship derived here would still apply.

Table 17. Summary of Equating Decisions for Reading

Grade	Form <sup>a</sup>	Scale <sup>b</sup>	Equating		
			Decision	Slope	Intercept
5	36	Expository	Unsmoothed Equipercentile	-	-
		Persuasive	Smoothed Equipercentile (s=0.01)	-	-
		Technical	Smoothed Equipercentile (s=0.01)	-	-
		Narrative	-	-	-
		Total <sup>c</sup>	Smoothed Equipercentile (s=0.05)	-	-
	53	Expository	Unsmoothed Equipercentile	-	-
		Persuasive	Smoothed Equipercentile (s=0.01)	-	-
		Technical	Unsmoothed Equipercentile	-	-
		Narrative	Unsmoothed Equipercentile	-	-
		Total <sup>c</sup>	Smoothed Equipercentile (s=0.10)	-	-
	70	Expository	Smoothed Equipercentile (s=0.01)	-	-
		Persuasive	Smoothed Equipercentile (s=0.01)	-	-
		Technical	Unsmoothed Equipercentile	-	-
		Narrative	Unsmoothed Equipercentile	-	-
		Total <sup>c</sup>	Smoothed Equipercentile (s=0.20)	-	-
8	28	Expository	Smoothed Equipercentile (s=0.30)	-	-
		Persuasive	Smoothed Equipercentile (s=0.01)	-	-
		Technical	Smoothed Equipercentile (s=0.01)	-	-
		Narrative	Unsmoothed Equipercentile	-	-
		Total <sup>c</sup>	Smoothed Equipercentile (s=0.01)	-	-
	45	Expository	Smoothed Equipercentile (s=0.30)	-	-
		Persuasive	Combination of 1) Unsmoothed Equipercentile below 32 2) line segment of Unsmoothed EQE thereafter	0.608	6.839
		Technical	Smoothed Equipercentile (s=0.01)	-	-
		Narrative	Smoothed Equipercentile (s=0.01)	-	-
		Total <sup>c</sup>	Smoothed Equipercentile (s=0.01)	-	-
	80	Expository	Smoothed Equipercentile (s=0.50)	-	-
		Persuasive	Smoothed Equipercentile (s=0.01)	-	-
		Technical	Smoothed Equipercentile (s=1.00)	-	-
		Narrative	Unsmoothed Equipercentile	-	-
		Total <sup>c</sup>	Smoothed Equipercentile (s=0.05)	-	-
11	54	Expository	Smoothed Equipercentile (s=0.10)	-	-
		Persuasive	Smoothed Equipercentile (s=0.05)	-	-
		Technical	Smoothed Equipercentile (s=0.05)	-	-
		Narrative	Combination of 1) Unsmoothed Equipercentile below 50 2) line segment of Unsmoothed EQE thereafter	0.877	3.152
		Total <sup>c</sup>	Smoothed Equipercentile (s=0.10)	-	-
	71	Expository	Unsmoothed Equipercentile	-	-
		Persuasive	Combination of 1) Unsmoothed Equipercentile below 49 2) line segment of Unsmoothed EQE thereafter	1.343350655	-27.6401056
		Technical	Smoothed Equipercentile (s=0.10)	-	-
		Narrative	Smoothed Equipercentile (s=0.05)	-	-
		Total <sup>c</sup>	Smoothed Equipercentile (s=0.01)	-	-
	98	Expository	Linear method	0.883309091	2.074
		Persuasive	Smoothed Equipercentile (s=0.01)	-	-
		Technical	Unsmoothed Equipercentile	-	-
		Narrative	Unsmoothed Equipercentile	-	-
		Total <sup>c</sup>	Smoothed Equipercentile (s=0.01)	-	-

<sup>a</sup> Base form at grade 5 is form 19, at grade 8 is form 62, and at grade 11 is form 37.

<sup>b</sup> These scales except Total were equated at the raw score level and expressed in percent correct metric.

<sup>c</sup> Reading Total Percent correct scores were transformed to a scale of with minimum -20 and maximum 80 prior to equating. Only the Score of 0 and above for this new scale were equated. After equating, 20 points were added back to all equated scores.

Another area of relative concern to equating is DIF. The DIF analyses reported in the previous section were done sometimes with small ethnic groups. In particular, Native American groups were typically small. It may be conceivable to repeat DIF analyses for some of these small ethnic groups in future administration years. However, dropping items after the equating work is completed will open up a new need to re-equate test forms.

Finally, a constant trade-off between test development and psychometrics typically occur in the area of improvement of test items. Changes, if any, made to improve items after equating is completed are often done to clarify items. Although these changes are done with good intentions, they may enhance the items and would result in the items becoming easier than when they were originally used to establish equating relationship. If such practices were not refrained from, drastic change in item difficulties may result. Since equating is concerned with minimizing form to form difficulty, such changes may unintentionally nullify the equating work.

## **Student Performance Level Classification Cut Scores**

As part of legislative and State Board of Education mandates, performance category definitions and criteria were to be determined for each of the assessments given by the state. At this writing the categories have been set, but definitions and the criteria (cut-scores or decision rules) for classifying students into a performance category level have been determined only for students taking the general assessments in reading and mathematics. For students taking the general reading, writing and mathematics assessments, five category levels of student performance have been defined: Advanced, Proficient, Satisfactory, Basic, and Unsatisfactory.

To match performance on the state's assessments to these categories, test cut-scores were identified to define the rules for classification of students. Cut-scores for each test were determined based on information gathered from teacher ratings of student classroom performance, student performance on the state's assessment tests, and the expert judgments of teachers, curriculum directors and principals. Teacher data on the judged performance levels of random samples of approximately 8000 students at a grade level and content area were obtained independent of the teacher's knowledge of student performance on the state assessments. These teacher ratings were matched with student test scores to provide data for an empirical contrasting groups approach to setting cut scores. These data were shared with KSDE personnel and KSDE content area advisory groups who subsequently determined the cut scores to be used for each content area test. As groups reviewed the contrasting groups data, an iterative process was used to arrive at consensus with the last step being a reality check review of actual student test score distribution given the identified cut scores.

Table 1 gives the proficiency level cut-scores for each grade level test in reading or mathematics based on the total percent correct score on a test. The "unsatisfactory" category indicates the point below which scores in that category would fall. The other categories include all scores falling at or above the cut-score indicated.

Table 1. Performance Level Category Cut-Scores Based on Total Percent Correct Scores

<b>Content Area and Grade Level</b>	<b>Advanced</b>	<b>Proficient</b>	<b>Satisfactory</b>	<b>Basic</b>	<b>Unsatisfactory</b>
Reading					
Grade 5	93	87	80	68	<68
Grade 8	93	87	80	68	<68
Grade 11	93	87	80	68	<68
Mathematics					
Grade 4	75	60	48	35	<35
Grade 7	75	60	48	35	<35
Grade 10	70	60	48	35	<35

As examples, students who had total scores of 93 percent correct or above on the 5<sup>th</sup> grade reading test would be classified as “Advanced”; students who had total scores of 87 percent correct or above, but below 93 percent correct would be classified as “Proficient”; students who had total scores of 80 percent correct or above, but below 87 percent correct would be classified as “Satisfactory”; students who had total scores of 68 percent correct or above, but below 80 percent correct would be classified as “Basic”; and students with percent correct total scores below 68 would be classified as “Unsatisfactory.” Note that the cut-scores for tests across grade levels within a content area are the same, except for the cut-scores needed to be in the “Advanced” category in mathematics. The “Advanced” category cut-score at grades 4 and 7 is 75 percent correct, but at grade 10 it is 70 percent correct.

### **Mathematics Assessment Test Characteristics**

As identified previously, the Mathematics Assessment at the 4<sup>th</sup>, 7<sup>th</sup>, and 10<sup>th</sup> grades consisted of four administered parts which contained objective items in a single-correct multiple-choice format. Each administered part was designed for a testing period of 30 minutes. In addition, at each grade level four parallel forms of the assessments were administered in a spiraled design in classrooms in Kansas.

Two process skill scores and a total score were reported for the Kansas Mathematics Assessment. The reported score for each student was the percentage of total points attained based on points available for each score.

In the year 2000, Mathematics Assessment items were administered in Kansas schools for the first time. Because of the psychometric analyses reported earlier in this manual, items with undesirable properties were dropped and results from different test forms were put on a common scale. In total, five items at grade 7 and five items at grade 10 were dropped across all forms in the 2000 Mathematics Assessment. At the seventh grade, two items in form 20 were not scored after item analyses. Two items in form 43 and one item in form 79 were not scored after a DIF analysis and item review identified these items as being possibly biased. At 10th grade, one item in form 30 was not scored after a DIF analysis and item review identified this item as potentially biased. Three items in form 76 and one item in form 99 were dropped due to printing errors or confusion in wording detected through the item analyses.

The measures and scores, which are determined and reported to students and schools, include the following. For a complete definition of each score, the Kansas Curricular Standards for Mathematics should be referenced.

#### **A. Mathematics Process Skill Scores**

- 1) Knowledge - (percent correct score reported)

Knowledge requires a student to know and/or to be able to do a set of mathematical concepts, facts, and/or procedures. Knowledge includes basic computation facts within all of the various number systems, capability in performing mathematical procedures as well as identification and explanation of mathematical concepts.

- 2) Application - (percent correct score reported)

Application requires students to describe how mathematical knowledge should be used or applied in the real world. Application includes:

- (a) translating various graphical, mental representations, and equation forms of mathematical concepts or procedures;
- (b) making inferences by deductive, inductive, proportional, or spatial reasoning;
- (c) generating new mathematical problems from a given set of known facts; and
- (d) applying mathematical concept, facts, or procedures in order to solve a novel problem.

Application involves higher order level processing of information.

## B. Mathematics Total Scores

### 3) Total - (percent correct score reported)

Total percent correct score is arrived at by adding the scores on all knowledge and application items and expressing this score as a percentage of the total number of items on knowledge and application combined. There was an approximate equal distribution of knowledge and application items on each grade 4 and grade 7 test form while on the grade 10 test forms, the distribution of knowledge to application items was approximately 35 percent to 65 percent. The exact number of items of each type on each grade level form is given in Table 19.

## C. Mathematics Results

Tables 18 and 19 report summary psychometric findings for the Mathematics Assessment. Table 18 identifies the mean student percent correct across the state at each grade level for the two process skill scores (Knowledge and Application) and the Mathematics total score.

Table 18  
Math Assessment Descriptive Statistics of Percent Correct<sup>a</sup>  
Mean (Standard Deviation) for Spring 2000 Scores Of Regular Education Students

Grade	Number Tested <sup>b</sup>	Process Scores		Total
		Knowledge	Application	
4	32113	61.04 (17.23)	49.69 (18.02)	55.39 (16.60)
7	32671	52.83 (19.95)	52.14 (19.08)	52.49 (18.54)
10	30782	48.93 (22.02)	47.42 (17.05)	47.99 (17.51)

<sup>a</sup> Values are mean equated percent of points available

<sup>b</sup> Number of students at each grade level on which means are based (includes all regular education students and gifted students from both public and private schools).

Because different test specifications were used at different grade levels and no vertical scaling was performed, it is impossible to compare performance across grade levels in Table 18. However, Application items appeared more difficult than Knowledge items for students at all grade levels.

Table 19 provides student and building level reliability coefficients for the 2000 Mathematics Assessment percent correct scores for each test form at each of the three grade levels. Based on the values in Table 19, all score reliabilities achieved acceptable levels. Student level Mathematics Total scores show evidence of a high level of reliability for the intended purposes of the testing program with coefficients ranging from a low of .85 to a high of .90 across all forms and grade levels. The Knowledge and Application process scores show

satisfactory reliability as well with coefficients ranging from .71 to .84 for the Knowledge subscales and from .72 to .82 for the Application subscales.

Table 19  
Math Assessment  
Student<sup>a</sup> and Building<sup>b</sup> Level Reliabilities for Spring 2000 Scores

Grade	Form	Mean class size	Scores								
			Knowledge			Application			Total		
			N of Items	Reliabilities		N of Items	Reliabilities		N of Items	Reliabilities	
		Student	Building <sup>c</sup>		Student	Building <sup>c</sup>		Student	Building <sup>c</sup>		
4	10	9	26	0.79	0.76	26	0.72	0.73	52	0.86	0.76
	33	9	26	0.76	0.77	26	0.75	0.75	52	0.86	0.78
	56	9	26	0.80	0.77	26	0.75	0.74	52	0.87	0.77
	79	9	26	0.77	0.77	26	0.78	0.75	52	0.87	0.77
	all	36		-	0.90		-	0.88		-	0.91
7	20	17	26	0.81	0.85	24	0.79	0.82	50	0.89	0.84
	43	17	25	0.84	0.86	25	0.74	0.82	50	0.88	0.85
	66	16	26	0.83	0.88	26	0.81	0.87	52	0.90	0.88
	89	16	25	0.80	0.87	26	0.79	0.85	51	0.89	0.87
	all	65		-	0.95		-	0.93		-	0.95
10	30	21	17	0.74	0.81	29	0.82	0.85	46	0.88	0.85
	53	20	17	0.77	0.83	30	0.78	0.80	47	0.87	0.82
	76	20	17	0.71	0.80	27	0.78	0.80	44	0.85	0.81
	99	20	17	0.84	0.82	29	0.78	0.83	46	0.90	0.83
	all	81		-	0.92		-	0.92		-	0.93

<sup>a</sup> Student level reliabilities are estimated using coefficient alpha.

<sup>b</sup> Building level reliabilities are for when the same set of items are repeated over two years with different cohorts of average sized schools. Buildings with smaller cohorts would have lower score reliabilities and buildings with larger cohorts would have higher score reliabilities.

<sup>c</sup> Building level reliabilities are estimated based on Feldt and Brennan(1989).

The reliability coefficients reported for buildings are for means scores. The last line reported in Table 19 for each grade level set of information is the reliability of building means for a building with an average class size for the state (36 for grade 4, 65 for grade 7 and 81 for grade 10). These reliability values are all quite high with the lowest coefficient being .88 for grade 4 building Application mean scores. The remainder of the coefficients are all .90 or greater. As the size of the building mean score reliabilities are dependent on class size, it should be noted that building means based on fewer students will have lower score reliabilities and those means for buildings with a greater number of students will have higher score reliabilities.

To provide some indication of the score reliability for small school building means, building score reliabilities also are reported for each form assuming that each form would have been administered to one-fourth of the students in the average size building. Thus, building score reliabilities are reported for building mean scores for small class sizes of 9 at grade 4, 16 – 17 at grade 7 and 20 – 21 at grade 10. These coefficients may be used to provide guidance to smaller school buildings in judging the reliability of their building mean scores.

In addition to reporting on percent correct scores, students also receive performance level classification scores identifying one of the five performance levels in which their percent correct score would classify them. To estimate the reliabilities of the performance level classifications for students, a procedure described by Livingston and Lewis (1995) was used which provides for an estimation for both classification accuracy and consistency for making student classification decisions based on scores from only one form of a test. As defined by Livingston and Lewis, accuracy “...refers to the extent to which the actual classifications of test takers (on the basis of their single-form scores) agree with those that would be made on the basis of their true scores, if their true scores could somehow be known (p. 180).” Consistency “refers to the agreement between the classifications based on two nonoverlapping, equally difficult forms of the test (p. 180).” The accuracy and consistency of classifications can be estimated overall and for each of the five distinct levels and also for dichotomous classification decisions about whether a student is in one of the categories above a specific cut score or in one of the categories below the cut score, e.g., being classified in the proficient or advanced level versus being classified in one of the combined satisfactory, basic or unsatisfactory levels.

Table 20 provides accuracy and consistency reliability estimates for the overall classification of students and for the classification of students into each of the five distinct levels. As indicated by the coefficients, the classification reliability is greater for the two end categories, advanced and unsatisfactory. This is not atypical as there is more opportunity for misclassification in the middle categories that have levels both below and above them.

Table 21 provides accuracy and consistency reliability estimates for specific dichotomous combined levels of classification. All of these coefficients are sufficiently high to provide confidence in the dichotomous decision being made about a student’s classification. Where students are concerned, the decision with the major instructional consequence is the

classification in the Unsatisfactory category. With the exception of the consistency index at grade 10 (.79), all coefficients for this classification decision are in the .90s and mid .80s, thus providing confidence that appropriate decisions are being made.

Table 20. Classification Reliability Accuracy and Consistency Estimates for each Performance Level Category and Overall Based on Total Mathematics Scores

Performance Level Classification	Grade 4 Tests		Grade 7 Tests		Grade 10 Tests	
	Accuracy	Consist.	Accuracy	Consist.	Accuracy	Consist.
Advanced	.77	.62	.85	.71	.82	.70
Proficient	.69	.58	.59	.48	.50	.37
Satisfactory	.57	.47	.59	.47	.48	.36
Basic	.66	.54	.51	.42	.48	.38
Unsatisfactory	.77	.66	.85	.74	.79	.68
Overall	.67	.56	.66	.57	.61	.51

Table 21. Classification Reliability Accuracy and Consistency Estimates for Specific Dichotomous Combined Classification Levels Based on Total Mathematics Scores

Dichotomous Classification	Grade 4 Tests		Grade 7 Tests		Grade 10 Tests	
	Accuracy	Consist.	Accuracy	Consist.	Accuracy	Consist.
Advanced vs Others	.94	.92	.95	.92	.95	.92
Adv,Prof vs Others	.89	.85	.91	.88	.91	.88
Adv,Prof,Sat vs Other	.89	.85	.90	.85	.88	.83
Others vs Unsatisfact	.94	.91	.90	.86	.85	.79

## **Reading Assessment Test Characteristics**

The Reading Assessments at the 5th, 8th, and 11th grades consisted of 4 comprehension passages representing 4 distinct text types: expository, persuasive, technical, and narrative. Expository texts are informational and can have causal and comparative organizational structures. Persuasive texts are designed to convince the reader to adopt a particular opinion or to perform a certain action. Technical texts inform the reader how to complete a certain technical task. Narrative texts, which include adventure stories and mysteries, frequently consist of a problem, conflict, and resolution. A series of multiple yes/no questions that related to a stem question accompanied every passage. Depending on the number of multiple yes/no items, a stem could have between 4 and 6 points total.

Each text and its questions were designed for a testing period of 30 minutes. In addition, at all grades four parallel forms of the assessments were administered in a spiraled design in Kansas's classrooms. Four text scores and a total score were reported for the Kansas Reading Assessments. Because of the psychometric analyses reported earlier in this manual, items with undesirable properties were dropped and results from different test forms were put on a common scale. At all grades, the first multiple yes/no Expository item was not scored as it was considered a practice item. At grade 5, four multiple yes/no items related to the same stem in form 19 for the Technical passage and, in form 53, one Persuasive and one Narrative yes/no question were not scored after item analyses. At grade 8, the list of items dropped across all four forms was longer. On form 28, four same-stem multiple yes/no Expository items, four same-stem Persuasive items and one Technical multiple yes/no item were not scored. Similarly, one Expository multiple yes/no item in form 45, two Expository items, two Persuasive items, and four same-stem Technical items in form 62, and one Expository and three Persuasive items in form 80 were also dropped after these initial analyses. At grade 10, one Persuasive and two Technical questions in form 37, one Expository item in form 54, one Technical item in form 71, and one Expository item and 5 same-stem Technical items were not scored based on the initial analysis results.

In Reading, items were also dropped due to bias. At grade 8, 4 same-stem Persuasive items in form 28 and, at grade 11 in form 37, 4 same-stem Expository items and 4 same-stem Technical items were not scored after DIF analysis and review identified them as possibly bias.

Table 7 continued

## Reading Results

Tables 22 and 23 report summary psychometric findings for the Reading Assessments. Table 22 identifies the mean student performance (percent correct scores) across the state at each grade level for the text type scores (Expository, Persuasive, Technical, and Narrative) and the Reading total scores. The most difficult text, at about 73 percent correct, was the Persuasive text at grade 10. However, because different forms were used without vertical scaling, performance comparisons across grade levels should not be made.

It should be noted when interpreting the percent correct averages that score values are expected to be above 50 percent correct given the type of item (multiple yes/no, i.e., true-false) on the test and the scoring procedure employed (number of correct decisions). If students were guessing at random, the expected average percent correct for any grouping of the items should have a lower bound of 50 percent correct. Thus, the effective score range for the values reported is from a lower bound of 50 to the upper bound of 100 on the percent correct scale. Given that very few, if any, students should be totally guessing, it is logical that a lower bound for the actual mean level of performance should be in the vicinity of 55 to 60 percent correct.

Table 22  
Reading Assessment Descriptive Statistics of Percent Correct<sup>a</sup>  
Mean (Standard Deviation) for Spring 2000 Scores of Only Regular Education Students

Grade	Number Tested <sup>b</sup>	Text Type Scores				Reading Total <sup>c</sup>
		Expository	Persuasive	Technical	Narrative	
5	31604	81.20 (14.46)	79.41 (10.35)	83.93 (11.39)	86.23 (12.26)	82.77 (10.16)
8	32710	80.21 (11.08)	82.53 (10.52)	85.26 (11.36)	83.62 (10.98)	82.97 (9.14)
11	28480	79.83 (12.72)	73.27 (13.74)	84.71 (11.51)	84.51 (11.87)	80.77 (10.58)

<sup>a</sup> Values are mean equated percent of points available

<sup>b</sup> Number of students at each grade level on which means are based (includes all regular education students and gifted students from both public and private schools).

<sup>c</sup> Reading Total is an equally weighted average of the Text Types percent correct scores

Table 23 provides student and building level reliability estimates for the 2000 Reading Assessments for each text type and form at all grade-levels. Based on these reliability estimates, total scores are highly consistent and achieve acceptable levels. Student total scores show evidence of satisfactory reliability for the purposes of the testing program. Part (or process) scores show satisfactory reliability as well with exception of a few text types at fifth grade where

Table 7 continued

the reliability index dipped below .70. Reliability of building averages for the individual forms, with exception of several text types at grade 5, are adequate. The same reliabilities when calculated for all forms combined are especially high. Using criteria discussed in the introduction, several text types at grade 5 such as form 19 Persuasive and Technical appear inadequate for making even low stakes decision. In contrast, other text types at the fifth grade appeared quite reasonable. At grades 8 and 11, all scores have reliability values that are adequate for use in low stakes decisions. Where the reliability is more than .85, scores can also be used for making high stakes decisions provided that corroborating evidence is used.

As for mathematics, performance level classifications are also made for each student based on their total reading percent correct score. Estimates of the classification reliability were made applying the same procedure as used for the mathematics assessments. Table 24 provides accuracy and consistency reliability estimates for the overall classification of students and for the classification of students into each of the five distinct levels. As indicated by the coefficients, the classification reliability is greater for the two end categories, advanced and unsatisfactory, and between these two categories, is higher for classifying students as unsatisfactory.

Table 25 provides accuracy and consistency reliability estimates for specific dichotomous combined levels of classification. All of these coefficients are sufficiently high to provide confidence in the dichotomous decision being made about a student's classification. Where students are concerned, the decision with the major instructional consequence is the classification in the Unsatisfactory category. In contrast with math where classifying students as advanced versus proficient or below has the greatest accuracy and consistency indices, the decision to classify students as unsatisfactory or not in reading has the greatest accuracy and consistency indices. All coefficients for this classification decision are in the mid .90s, thus providing a high degree of confidence that appropriate decisions are being made when students are classified as unsatisfactory on the basis of their reading test total scores.

Table 23  
 Student<sup>a</sup> and Building<sup>b</sup> Level Reliabilities for Spring 2000 Reading Assessment Scores

Grade	Form	Mean Bldg size	Text Type																			
			Expository				Persuasive				Technical				Narrative				Total			
			n		max		n		max		n		max		n		max		n		max	
			Student	Bldg <sup>c</sup>	Student	Bldg <sup>c</sup>	Student	Bldg <sup>c</sup>	Student	Bldg <sup>c</sup>	Student	Bldg <sup>c</sup>	Student	Bldg <sup>c</sup>	Student	Bldg <sup>c</sup>	Student	Bldg <sup>c</sup>	Student	Bldg <sup>c</sup>		
5	19	9	11	44	0.83	0.68	9	37	0.62	0.70	7	30	0.67	0.69	10	39	0.79	0.64	37	150	0.91	0.71
	36	9	10	40	0.73	0.76	10	40	0.76	0.71	8	32	0.77	0.75	10	39	0.80	0.69	38	151	0.92	0.75
	53	9	10	40	0.70	0.68	10	40	0.74	0.68	9	37	0.70	0.69	10	38	0.73	0.67	39	155	0.90	0.72
	70	9	10	40	0.74	0.68	10	40	0.70	0.68	9	37	0.73	0.67	10	42	0.80	0.72	39	159	0.91	0.73
	all	36			-	0.85			-	0.85			-	0.86			-	0.84			-	0.88
8	28	17	11	45	0.78	0.78	12	48	0.79	0.83	12	50	0.79	0.82	14	56	0.83	0.83	49	199	0.93	0.81
	45	17	12	49	0.77	0.85	14	58	0.82	0.85	11	46	0.72	0.81	12	49	0.81	0.85	49	202	0.93	0.87
	62	16	13	54	0.78	0.79	11	42	0.67	0.90	11	49	0.79	0.79	10	43	0.71	0.81	45	188	0.91	0.81
	80	16	13	54	0.80	0.83	11	42	0.71	0.82	12	50	0.82	0.83	10	40	0.79	0.83	46	186	0.93	0.85
	all	65			-	0.91			-	0.94			-	0.93			-	0.93			-	0.94
11	37	19	12	49	0.78	0.89	11	46	0.77	0.83	12	51	0.80	0.85	13	54	0.82	0.83	48	200	0.93	0.87
	54	19	12	52	0.83	0.79	11	51	0.78	0.82	13	55	0.79	0.89	14	59	0.81	0.83	50	217	0.93	0.85
	71	19	10	47	0.76	0.84	13	55	0.82	0.81	13	56	0.80	0.84	11	48	0.74	0.83	47	206	0.92	0.85
	98	19	13	55	0.78	0.82	12	50	0.78	0.83	10	42	0.71	0.83	12	55	0.82	0.83	47	202	0.92	0.86
	all	74			-	0.92			-	0.92			-	0.95			-	0.92			-	0.94

<sup>a</sup> Student level reliabilities are estimated using coefficient alpha.

<sup>b</sup> Building level reliabilities are for when the same set of items are repeated over two years with different cohorts of average sized schools. Buildings with smaller cohorts would have lower score reliabilities and buildings with larger cohorts would have higher score reliabilities.

<sup>c</sup> Building level reliabilities are estimated based on Feldt and Brennan(1989).

Table 24. Classification Reliability Accuracy and Consistency Estimates for each Performance Level Category and Overall Based on Total Reading Scores

Performance Level Classification	Grade 5 Tests		Grade 8 Tests		Grade 11 Tests	
	Accuracy	Consist.	Accuracy	Consist.	Accuracy	Consist.
Advanced	.81	.62	.	.48	.72	.49
Proficient	.57	.50	.53	.51	.66	.59
Satisfactory	.58	.48	.64	.55	.63	.54
Basic	.71	.63	.68	.62	.74	.67
Unsatisfactory	.82	.77	.84	.78	.84	.81
Overall	.67	.58	.63	.57	.70	.62

Table 25. Classification Reliability Accuracy and Consistency Estimates for Specific Dichotomous Combined Classification Levels Based on Total Reading Scores

Dichotomous Classification	Grade 5 Tests		Grade 8 Tests		Grade 11 Tests	
	Accuracy	Consist.	Accuracy	Consist.	Accuracy	Consist.
Advanced vs Others	.89	.86	.86	.85	.92	.90
Adv,Prof vs Others	.90	.87	.88	.85	.90	.88
Adv,Prof,Sat vs Other	.92	.89	.92	.89	.92	.90
Others vs Unsatisfact	.96	.94	.96	.95	.96	.94

## Writing Assessment Characteristics

Identical to reading, local districts conducted the 2000 Kansas Writing Assessment at grades 5, 8, and 11. The implementation of the writing assessment locally was not standardized statewide with local districts having some flexibility in its administration to align with local instructional practice. For all districts, a list of 2-3 prompt options at each grade level was provided. Students chose one writing prompt from this option list on which to write.

Student papers were scored by either two local raters or rated at the local level by one rater and then submitted for evaluation by a state rater. When districts provided two local reads, ten percent or 20 of their papers (whichever was greater) were returned to the state for scoring and evaluation. Each student's final writing sample was judged using the Six-Trait Analytical Scoring Model to evaluate the following: Ideas and Content, Organization, Voice, Word Choice, Sentence Fluency, and Conventions. The scores on the six traits were weighted and averaged to produce a composite (total) writing score. Each trait was rated on a five-point scale (1=low, 5=high). Half-point ratings were possible (i.e., 2.5), but raters were encouraged to use them infrequently. The definitions of the scale score values were as follows.

- 1) Beginning - searching, exploring, struggling -- looking for a sense of purpose or way to begin.
- 2) Emerging - Moments that trigger reader's/writer's questions -- Stories/ideas buried within the text.
- 3) Developing - Writer begins to take control, begins to shape ideas -- writing has definite direction, coherence, momentum, and sense of purpose.
- 4) Maturing - More control -- Writer has confidence to experiment; is about a draft away.
- 5) Strong - writer in control -- skillfully shaping and directing the writing, evidence of fine-tuning.

The reported score for each student was a mathematical average of the local rater scores and when available the state rater score. The composite score was a weighted average of the six trait scores. For this average, ideas/content and organization were given a weight of three, voice and word choice were given a weight of two, and sentence fluency and conventions a weight of one.

Because the writing program was unstandardized in its implementation, results are reported based on district clusters that reflect administration practices in the local districts in regards to raters, administration time, and student use of word processors. Tables 26, 27, 28 and 29 show summary psychometric findings for the Writing Assessment.

Table 26  
Writing Assessment Descriptive Statistics for Spring 2000 Scores by Grade Level and Program Cluster

Grade	cluster <sup>a</sup>	Number Tested <sup>b</sup>	Trait						
			Ideas /Content	Organization	Voice	Word Choice	Sentence Fluency	Convention	Composite
5	1	671	3.20 (0.86)	3.06 (0.91)	3.27 (0.86)	3.11 (0.81)	2.97 (0.87)	2.95 (0.93)	3.12 (0.81)
	2	1457	3.14 (0.82)	3.01 (0.82)	3.18 (0.80)	3.01 (0.73)	2.93 (0.82)	2.89 (0.88)	3.05 (0.74)
	3	7728	3.40 (0.84)	3.29 (0.87)	3.45 (0.84)	3.30 (0.79)	3.22 (0.86)	3.27 (0.92)	3.34 (0.78)
	4	13469	3.34 (0.81)	3.22 (0.85)	3.38 (0.83)	3.20 (0.76)	3.15 (0.83)	3.12 (0.89)	3.26 (0.76)
	5	81	2.95 (0.97)	2.84 (0.97)	3.14 (0.90)	2.86 (0.89)	2.74 (0.95)	2.77 (1.01)	2.91 (0.89)
	6	415	3.02 (0.82)	2.80 (0.85)	3.10 (0.81)	2.86 (0.73)	2.73 (0.85)	2.65 (0.90)	2.90 (0.75)
	7	831	3.41 (0.79)	3.30 (0.85)	3.46 (0.80)	3.29 (0.72)	3.24 (0.85)	3.29 (0.89)	3.35 (0.75)
	8	2791	3.25 (0.82)	3.08 (0.86)	3.30 (0.84)	3.10 (0.77)	3.01 (0.85)	2.98 (0.92)	3.15 (0.77)
	9	2040	3.31 (0.85)	3.19 (0.88)	3.35 (0.86)	3.18 (0.80)	3.10 (0.87)	3.10 (0.92)	3.23 (0.79)
8	1	1647	3.36 (0.82)	3.16 (0.86)	3.42 (0.77)	3.17 (0.71)	3.08 (0.83)	3.05 (0.87)	3.24 (0.73)
	2	2112	3.33 (0.78)	3.16 (0.82)	3.40 (0.72)	3.17 (0.68)	3.11 (0.76)	3.04 (0.82)	3.23 (0.70)
	3	9460	3.62 (0.80)	3.50 (0.82)	3.66 (0.74)	3.46 (0.71)	3.42 (0.81)	3.42 (0.85)	3.54 (0.71)
	4	10096	3.54 (0.78)	3.43 (0.83)	3.56 (0.72)	3.36 (0.69)	3.36 (0.78)	3.31 (0.84)	3.45 (0.70)
	5	170	3.38 (0.77)	3.18 (0.80)	3.35 (0.70)	3.15 (0.67)	3.08 (0.75)	3.05 (0.84)	3.23 (0.68)
	6	709	3.23 (0.79)	2.97 (0.87)	3.26 (0.77)	3.03 (0.70)	2.93 (0.80)	2.84 (0.85)	3.08 (0.72)
	7	1172	3.68 (0.75)	3.57 (0.79)	3.70 (0.72)	3.50 (0.68)	3.46 (0.77)	3.46 (0.84)	3.59 (0.68)
	8	2530	3.49 (0.77)	3.31 (0.83)	3.45 (0.72)	3.28 (0.67)	3.25 (0.77)	3.23 (0.85)	3.36 (0.69)
	9	2578	3.52 (0.86)	3.40 (0.92)	3.54 (0.85)	3.34 (0.80)	3.30 (0.86)	3.26 (0.91)	3.42 (0.81)
11	1	3007	3.36 (0.84)	3.29 (0.88)	3.53 (0.77)	3.23 (0.78)	3.22 (0.87)	3.17 (0.91)	3.32 (0.77)
	2	3267	3.28 (0.83)	3.23 (0.84)	3.46 (0.75)	3.18 (0.73)	3.16 (0.79)	3.09 (0.84)	3.25 (0.73)
	3	8134	3.58 (0.80)	3.53 (0.83)	3.69 (0.72)	3.42 (0.74)	3.44 (0.82)	3.42 (0.88)	3.53 (0.72)
	4	5841	3.54 (0.76)	3.50 (0.78)	3.66 (0.67)	3.41 (0.67)	3.43 (0.74)	3.37 (0.80)	3.50 (0.67)
	5	619	3.25 (0.75)	3.20 (0.78)	3.44 (0.67)	3.17 (0.66)	3.14 (0.75)	3.13 (0.81)	3.24 (0.67)
	6	1695	3.20 (0.71)	3.12 (0.76)	3.34 (0.62)	3.10 (0.64)	3.10 (0.72)	3.07 (0.79)	3.17 (0.64)
	7	1332	3.51 (0.73)	3.46 (0.78)	3.59 (0.67)	3.36 (0.68)	3.44 (0.75)	3.47 (0.80)	3.48 (0.66)
	8	2008	3.51 (0.73)	3.43 (0.80)	3.59 (0.66)	3.35 (0.69)	3.39 (0.75)	3.38 (0.81)	3.46 (0.67)
	9	3695	3.31 (0.82)	3.26 (0.85)	3.48 (0.74)	3.20 (0.75)	3.19 (0.82)	3.14 (0.89)	3.28 (0.74)

<sup>a</sup> Writing program clusters are identified by the different practices followed by classrooms in buildings across the state of Kansas. Cluster 1 is a program which allows less than 4 editing/writing days and use of word processor and papers are rated by 2 local raters. Cluster 2 allows less than 4 editing/writing days and no use of word processor and papers are rated by 2 local raters. Cluster 3 allows 4 editing/writing days and use of word processor and papers are rated by 2 local raters. Cluster 4 allows 4 editing/writing days and no use of word processor and papers are rated by 2 local raters. Cluster 5 allows less than 4 editing/writing days and use of word processor and papers are rated by one local and one state raters. Cluster 6 allows less than 4 editing/writing days and, no use of word processor and papers are rated by one local and one state raters. Cluster 7 allows 4 editing/writing days and use of word processor and papers are rated by one local and one state raters. Cluster 8 allows 4 editing/writing days and no use of word processor and papers are rated by one local and one state raters. Cluster 9 is an unidentified writing practice.

<sup>b</sup> Number of students at each grade level on which means are based (includes all regular education students and gifted students from both public and private schools).

Table 26 displays the mean writing scores for papers rated by state raters by writing program cluster at the three grade levels. In addition, the number of student papers in each cluster is provided. Most papers at all grade levels are rated by two local raters (clusters numbered at or below 4) and more students are given writing and editing periods of four days to complete their work (clusters 3, 4, 7, and 8). The even numbered clusters across grade levels show that more students did not use word processing. The number of students allowed use of word-processors was higher at the 11th grade than those in the other two grade levels. In one instance at this grade level, the number of student papers word-processed outnumbered those that were not. In general, the mean Writing trait and composite scores were higher when scored by two local raters than when scored by one local and one state rater.

Table 27 reports information addressing the consistency or ratings between local and state raters when scoring student writing papers. The local score in this table for most papers was the single local rater but for some when available the average of two local raters. The state raters scored student papers lower than the local raters. However, the difference was relatively modest from 0.07 to 0.48. The correlations ranged from a low of 0.36 to a high of 0.78 indicating moderate consistency.

Table 28 shows percent of agreement between local and state raters. The percent of exact agreements were relatively low with a high of 38.3%, but local and state discrepancies of less than or equal to 1.5 were all, except one (88.2%), in the 90% range.

The rater correlations in Table 27 were used to estimate student level rater reliabilities in Table 29. Table 29 provides student and building level reliability coefficients for the 2000 Writing Assessment. Student information is based on the correlations between scores from state and local raters. These rater reliabilities are the maximum that the reliability could be for the student level writing scores. Because the writing scores are based on one item, score reliabilities cannot be estimated. The rater reliabilities are mostly in a range from 0.52 to 0.88. Rater reliabilities are slightly higher for the Conventions and Composite scores. Measurement error for this assessment at the student level is related to the use of only one item.

Building level reliabilities were obtained by treating means for each prompt chosen by students as different sets of building scores. These prompt means were then supplied into the coefficient alpha formula to obtain building level reliabilities. In general, average building sizes

for writing programs (clusters 1, 2, 5, and 6) that allow less than 4 editing/writing days for students to complete their work were lower than other clusters reported in Table 29. This is more pronounced at grade 5. Cluster 2 at grade 5, for instance, had prompt means based on an average of 5 students. Because building level reliabilities for this cluster at four trait levels and the composite score levels were close to zero, mean scores coming from this cluster were not consistent. Cluster 5 at grade 11 also suffered from inconsistencies between mean scores. Building reliabilities for other clusters were generally in the mid .60s for all trait levels with the exception of only cluster 8 at all grade levels and cluster 4 at only grade 8, which appeared to show adequate reliability values.

It should be noted that neither the student nor building reliability index reflects measurement errors due to the interaction between raters and prompts. In addition, the index also does not take into account the interaction between student and prompt which is confounded given that students write on only one prompt of choice. At the student level, the writing scores should not be used alone for decision-making. The use needs to be limited to supporting instructional and diagnostic planning. At the building level, writing scores for only cluster 8 at all grade levels and cluster 4 at only grade 8 meet the minimum standard acceptable for use in low stake decisions. These writing scores are generally not consistent enough for use in tracking growth in Writing for Kansas's schools.

Table 27  
Writing Rater Consistency Descriptive Statistics by Grade Level and Program Clusters

Grade	cluster	Statistics	N	Trait						
				Ideas /Content	Organization	Voice	Word Choice	Sentence Fluency	Convention	Composite
5	1	Local Mean	671	3.24	3.10	3.31	3.14	2.99	2.96	3.16
		State Mean	289	3.08	2.94	3.20	3.03	2.93	2.93	3.03
		Correlations	289	0.66	0.65	0.62	0.64	0.65	0.70	0.74
	2	Local Mean	1457	3.17	3.05	3.20	3.02	2.95	2.89	3.08
		State Mean	741	3.06	2.90	3.14	2.99	2.89	2.92	3.00
		Correlations	741	0.62	0.62	0.54	0.61	0.62	0.67	0.70
	3	Local Mean	7726	3.43	3.34	3.47	3.31	3.24	3.28	3.37
		State Mean	3420	3.31	3.18	3.44	3.27	3.18	3.23	3.28
		Correlations	3420	0.60	0.60	0.58	0.62	0.64	0.67	0.69
	4	Local Mean	13466	3.38	3.28	3.39	3.22	3.19	3.15	3.30
		State Mean	6702	3.25	3.11	3.37	3.17	3.10	3.10	3.20
		Correlations	6701	0.61	0.61	0.59	0.62	0.64	0.68	0.70
	5	Local Mean	81	3.01	2.91	3.19	2.88	2.77	2.75	2.95
		State Mean	81	2.88	2.78	3.11	2.86	2.72	2.80	2.87
		Correlations	81	0.65	0.70	0.60	0.74	0.78	0.78	0.78
	6	Local Mean	415	3.18	2.94	3.17	2.90	2.80	2.68	3.00
		State Mean	415	2.88	2.68	3.04	2.84	2.68	2.63	2.81
		Correlations	415	0.54	0.60	0.51	0.53	0.55	0.59	0.65
	7	Local Mean	831	3.56	3.47	3.57	3.38	3.34	3.38	3.47
		State Mean	831	3.27	3.15	3.36	3.21	3.15	3.21	3.23
		Correlations	831	0.59	0.60	0.55	0.58	0.64	0.64	0.68
	8	Local Mean	2790	3.39	3.22	3.36	3.17	3.12	3.05	3.26
		State Mean	2788	3.12	2.96	3.25	3.05	2.93	2.92	3.06
		Correlations	2787	0.58	0.60	0.56	0.59	0.61	0.67	0.68
9	Local Mean	2014	3.36	3.25	3.39	3.19	3.13	3.11	3.27	
	State Mean	1152	3.15	3.00	3.22	3.09	3.00	3.00	3.09	
	Correlations	1142	0.59	0.57	0.60	0.60	0.63	0.65	0.69	
8	1	Local Mean	1647	3.39	3.20	3.45	3.17	3.09	3.06	3.26
		State Mean	580	3.24	3.07	3.35	3.16	3.07	3.05	3.17
		Correlations	580	0.60	0.62	0.53	0.59	0.58	0.64	0.70
	2	Local Mean	2112	3.36	3.19	3.43	3.18	3.12	3.04	3.25
		State Mean	713	3.27	3.07	3.35	3.17	3.15	3.10	3.20
		Correlations	713	0.57	0.64	0.53	0.56	0.63	0.68	0.69
	3	Local Mean	9458	3.66	3.53	3.69	3.47	3.43	3.43	3.56
		State Mean	3285	3.52	3.39	3.54	3.41	3.41	3.41	3.46
		Correlations	3284	0.58	0.60	0.52	0.57	0.63	0.64	0.68
	4	Local Mean	10093	3.56	3.46	3.58	3.37	3.37	3.31	3.47
		State Mean	3255	3.46	3.31	3.49	3.34	3.35	3.32	3.39
		Correlations	3255	0.59	0.61	0.50	0.57	0.60	0.64	0.68
	5	Local Mean	170	3.59	3.36	3.60	3.27	3.11	3.10	3.39
		State Mean	170	3.20	3.04	3.13	3.06	3.07	3.01	3.09
		Correlations	170	0.61	0.56	0.36	0.45	0.59	0.59	0.60
	6	Local Mean	709	3.41	3.12	3.36	3.08	2.99	2.91	3.20
		State Mean	708	3.06	2.83	3.17	2.98	2.87	2.77	2.97
		Correlations	708	0.56	0.60	0.53	0.53	0.61	0.63	0.68

Table 6 continued

Grade cluster	Statistics	N	Trait							
			Ideas /Content	Organization	Voice	Word Choice	Sentence Fluency	Convention	Composite	
7	Local Mean	1172	3.89	3.79	3.91	3.63	3.58	3.58	3.77	
	State Mean	1170	3.48	3.36	3.50	3.38	3.36	3.36	3.42	
	Correlations	1170	0.54	0.57	0.51	0.53	0.58	0.62	0.65	
8	Local Mean	2527	3.63	3.45	3.56	3.36	3.30	3.29	3.47	
	State Mean	2524	3.36	3.18	3.36	3.22	3.20	3.18	3.27	
	Correlations	2521	0.53	0.57	0.51	0.55	0.57	0.64	0.65	
9	Local Mean	2495	3.58	3.46	3.60	3.38	3.34	3.29	3.47	
	State Mean	1122	3.42	3.27	3.43	3.30	3.27	3.25	3.34	
	Correlations	1089	0.58	0.57	0.54	0.59	0.64	0.67	0.67	
11	1	Local Mean	3005	3.38	3.32	3.56	3.26	3.24	3.18	3.35
		State Mean	979	3.13	3.03	3.32	3.03	3.04	3.01	3.10
		Correlations	979	0.64	0.66	0.55	0.64	0.66	0.68	0.73
	2	Local Mean	3266	3.30	3.25	3.48	3.20	3.16	3.10	3.27
		State Mean	927	3.20	3.10	3.42	3.09	3.12	3.06	3.17
		Correlations	927	0.56	0.61	0.53	0.60	0.64	0.66	0.68
	3	Local Mean	8134	3.62	3.58	3.72	3.45	3.47	3.45	3.57
		State Mean	3017	3.42	3.35	3.54	3.30	3.33	3.32	3.39
		Correlations	3017	0.58	0.59	0.50	0.59	0.62	0.64	0.67
	4	Local Mean	5841	3.57	3.54	3.70	3.44	3.45	3.39	3.54
		State Mean	1885	3.40	3.34	3.49	3.27	3.34	3.30	3.36
		Correlations	1885	0.60	0.62	0.52	0.61	0.63	0.65	0.70
	5	Local Mean	618	3.34	3.32	3.51	3.24	3.23	3.19	3.32
		State Mean	618	3.18	3.10	3.37	3.10	3.06	3.07	3.16
		Correlations	617	0.57	0.60	0.49	0.54	0.57	0.59	0.67
	6	Local Mean	1694	3.29	3.22	3.42	3.17	3.18	3.14	3.25
		State Mean	1694	3.12	3.02	3.27	3.03	3.02	3.00	3.08
		Correlations	1693	0.58	0.59	0.50	0.57	0.59	0.64	0.67
	7	Local Mean	1329	3.63	3.59	3.71	3.45	3.54	3.58	3.59
		State Mean	1331	3.39	3.34	3.48	3.29	3.35	3.37	3.37
		Correlations	1328	0.53	0.56	0.47	0.51	0.54	0.57	0.62
	8	Local Mean	2007	3.65	3.57	3.69	3.45	3.48	3.49	3.57
		State Mean	2005	3.39	3.30	3.50	3.26	3.31	3.28	3.35
		Correlations	2004	0.58	0.60	0.51	0.58	0.60	0.61	0.68
9	Local Mean	3529	3.37	3.32	3.52	3.24	3.22	3.15	3.32	
	State Mean	1595	3.17	3.10	3.36	3.09	3.10	3.10	3.16	
	Correlations	1577	0.62	0.65	0.52	0.64	0.64	0.68	0.72	

Table 28

## Writing Rater Percent Agreements Descriptive Statistics by Grade Level and Writing Program Clusters

Grade	cluster	Percent	Trait						
		Agreement per Discrepancy	Ideas /Content	Organization	Voice	Word Choice	Sentence Fluency	Convention	Composite
5	1	Exact(0)	22.1	14.2	17.6	19.8	19.8	20.8	2.1
		≤ 1.5	97.2	97.2	97.6	98.6	97.9	97.9	96.8
		≤ 2	99.7	99.7	99	100	100	99	100
	2	Exact(0)	20.6	16.9	16.2	23.6	18.6	19.5	4
		≤ 1.5	97.6	96.9	96.4	99.1	97	97.2	97.4
		≤ 2	99.5	98.9	99.1	99.9	99.5	99.3	99.4
	3	Exact(0)	20.4	18.1	18.9	22.2	19.6	19.7	3.6
		≤ 1.5	95.7	95.3	95.5	97.5	96.8	96.6	96.1
		≤ 2	99.1	98.7	99.2	99.7	99.2	99.3	99.3
	4	Exact(0)	20.6	18	19.8	22.3	19	19.6	3.4
		≤ 1.5	96.4	95.7	96.1	97.7	96.8	97	96.7
		≤ 2	99.1	99	99.2	99.7	99.4	99.5	99.4
	5	Exact(0)	28.4	30.9	21	35.8	38.3	32.1	6.3
		≤ 1.5	93.8	96.3	93.8	100	98.8	97.5	96.2
		≤ 2	98.8	98.8	98.8		100	98.8	98.7
	6	Exact(0)	27	25.2	23.4	26.9	26.5	26.1	1.7
		≤ 1.5	92	93	93.7	95.9	94	93.7	96.3
		≤ 2	98.6	97.8	98.6	99.3	98.1	97.8	98.3
	7	Exact(0)	30.6	27	29.3	31.3	30.2	30	4.6
		≤ 1.5	95.2	94	93.8	97.1	95.8	95.4	95.8
		≤ 2	98.7	98.1	98.7	100	99	99.3	99.4
	8	Exact(0)	27.6	26.7	26.8	31.4	28.8	28.6	3.1
		≤ 1.5	93.1	94.1	93.3	95.7	95.1	95.3	95.5
		≤ 2	98.5	98.5	98.6	99.6	98.5	99	99
9	Exact(0)	23.1	19.5	20.4	26.4	24.7	21.9	2.2	
	≤ 1.5	96.1	95.4	95.8	97.1	96.8	96	96.9	
	≤ 2	99.5	98.2	99.6	99.7	99.2	99.2	99.4	
8	1	Exact(0)	20	20.2	19.1	21.6	16.8	17.9	4.1
		≤ 1.5	97.6	97.1	96.6	98.8	97.9	96.9	98.1
		≤ 2	99	99	99.1	99.7	98.8	99.7	99.8
	2	Exact(0)	17.1	19.8	18.4	22.7	20.6	20.1	3.5
		≤ 1.5	95.4	97.1	96.5	97.6	97.2	98	96.9
		≤ 2	98.3	98.6	99.2	99.7	99.2	99.4	99.3
	3	Exact(0)	21.1	18.7	19	24.3	20.8	20.1	3.8
		≤ 1.5	96.5	96.3	96.6	98	97.2	97.3	97
		≤ 2	98.8	98.8	99.2	99.5	99.6	99.4	99.3
	4	Exact(0)	21	18.1	17.9	23.7	19.9	19.3	3.3
		≤ 1.5	97	96.7	96.7	97.9	97.4	96.8	97.4
		≤ 2	99.1	99.3	99.2	99.8	99.4	99.5	99.3
	5	Exact(0)	25.3	33.7	16.6	29.6	28.4	27.6	4.2
		≤ 1.5	92.9	91.1	88.2	92.9	95.9	94.1	92.2
		≤ 2	98.8	97	97.6	98.8	98.8	98.8	98.8
	6	Exact(0)	24.9	23.6	25	31.4	28.8	28.3	2.2
		≤ 1.5	93.6	94.6	93.8	96.6	96.7	96	96.5
		≤ 2	98.3	98.7	98.7	99.3	99.3	99.3	99.6

Table 7 continued

Grade	cluster	Percent	Trait						
		Agreement per Discrepancy	Ideas /Content	Organization	Voice	Word Choice	Sentence Fluency	Convention	Composite
11	7	Exact(0)	28	25	26.5	30.4	26.8	27.6	2.7
		≤ 1.5	93.8	93.8	94.1	96.6	96.7	96.1	96
		≤ 2	98.4	98.5	98.5	99.3	99.2	98.9	99.1
	8	Exact(0)	27.3	25.5	25.5	32.2	27.1	27.6	2.8
		≤ 1.5	94.4	94	95.9	97.6	96.1	96.7	96.5
		≤ 2	97.9	98.4	99.1	99.6	99.1	99.3	99.1
	9	Exact(0)	22.7	19.4	21.4	22	22	22.5	4
		≤ 1.5	95.5	93	96.3	97.2	97.8	97	95.7
		≤ 2	99	98.3	99.1	99.9	99.4	99.4	99
11	1	Exact(0)	19.5	17.8	20.3	19.4	18.5	18.4	2.7
		≤ 1.5	97.5	96.1	96.6	97.7	96.5	97.2	96.9
		≤ 2	99.3	98.7	99.3	99.4	99.2	99.3	99.9
	2	Exact(0)	20.3	17.8	19.1	21.8	18.5	21.4	3.3
		≤ 1.5	95.1	96.1	96.6	96.9	96.7	97.1	95.9
		≤ 2	98.3	99.2	99.6	99.6	99.5	99.2	99.1
	3	Exact(0)	19	17.1	18.8	21.1	18.5	16.9	2.8
		≤ 1.5	95.8	95.4	96.3	97.2	96.9	96.3	96.7
		≤ 2	99	98.8	99.3	99.5	99.2	99.4	99.1
	4	Exact(0)	18.8	18.7	17.9	20.9	21.1	18.8	2.7
		≤ 1.5	97.3	96.5	97.8	98.7	98.3	97.3	97.6
		≤ 2	99.4	99.4	99.7	99.8	99.7	99.5	99.7
	5	Exact(0)	28	26.5	29.1	30.5	26.9	27.8	3
		≤ 1.5	96.6	96.4	96.9	97.7	97.1	96.6	97.4
		≤ 2	98.7	99.2	99.4	99.5	98.9	98.9	99.5
	6	Exact(0)	31	28.1	31.8	33.5	30.6	29.7	3.6
		≤ 1.5	96.5	96.3	97.4	98.3	97.5	97.3	98.1
		≤ 2	99.6	99.3	99.5	99.8	99.7	99.6	99.8
	7	Exact(0)	25.5	26.4	27.5	29.3	25.2	24.8	3.4
		≤ 1.5	95.8	95	95.9	96.7	96.8	95.5	97
		≤ 2	99.2	98.9	99.3	99.6	99.2	98.6	99.5
	8	Exact(0)	30.4	27.4	28.7	31.6	28.1	27.7	4.1
		≤ 1.5	96	95.4	96.9	97.6	97	95.7	97
		≤ 2	99.4	99	99.3	99.6	99.6	99	99.6
	9	Exact(0)	22.8	21	22.2	27.3	22.4	21.4	3
		≤ 1.5	96.8	96.3	96.4	98	97.2	97.1	97.7
		≤ 2	99.2	99	99.4	99.6	99.4	99.2	99.7

Table 29  
Students<sup>a</sup> and Building<sup>b</sup> Writing Score Reliabilities by Grade Level and Writing Program Clusters

Grade	cluster	mean		Rater Reliability for Students							Score Reliability for Building Averages						
		bldg size	bldg size	Ideas /Content	Organization	Voice	Word Choice	Sentence Fluency	Convention	Composite	Ideas /Content	Organization	Voice	Word Choice	Sentence Fluency	Convention	Composite
5	1	59	5	0.80	0.79	0.77	0.78	0.78	0.82	0.85	0.52	0.48	0.43	0.55	0.53	0.51	0.53
	2	110	6	0.77	0.77	0.70	0.76	0.77	0.81	0.82	0.04	-0.05	-0.03	0.06	0.22	0.34	0.04
	3	213	33	0.75	0.75	0.74	0.77	0.78	0.80	0.82	0.62	0.67	0.65	0.66	0.63	0.70	0.68
	4	389	31	0.76	0.76	0.74	0.77	0.78	0.81	0.82	0.64	0.66	0.65	0.64	0.60	0.64	0.66
	5	7	5	0.79	0.83	0.75	0.85	0.87	0.88	0.88	0.32	0.40	0.46	0.64	0.57	0.65	0.51
	6	33	6	0.70	0.75	0.67	0.69	0.71	0.74	0.78	0.69	0.51	0.53	0.56	0.69	0.53	0.63
	7	34	18	0.74	0.75	0.71	0.73	0.78	0.78	0.81	0.44	0.54	0.44	0.62	0.47	0.56	0.56
	8	92	23	0.73	0.75	0.72	0.74	0.76	0.80	0.81	0.76	0.74	0.73	0.70	0.72	0.75	0.76
	9	84	6	0.75	0.73	0.75	0.75	0.78	0.79	0.81	0.40	0.28	0.47	0.43	0.44	0.61	0.41
8	1	67	11	0.75	0.77	0.69	0.74	0.73	0.78	0.82	0.43	0.46	0.51	0.45	0.53	0.49	0.47
	2	71	14	0.72	0.78	0.69	0.72	0.77	0.81	0.82	0.72	0.60	0.71	0.61	0.66	0.68	0.69
	3	161	49	0.73	0.75	0.68	0.73	0.77	0.78	0.81	0.69	0.70	0.66	0.67	0.69	0.77	0.71
	4	142	57	0.74	0.76	0.67	0.72	0.75	0.78	0.81	0.78	0.77	0.77	0.78	0.75	0.78	0.78
	5	9	5	0.75	0.71	0.52	0.62	0.74	0.74	0.75	0.55	0.79	0.79	0.52	0.34	0.38	0.71
	6	24	15	0.72	0.75	0.69	0.69	0.76	0.77	0.81	0.68	0.67	0.70	0.51	0.47	0.25	0.64
	7	31	23	0.70	0.72	0.68	0.70	0.74	0.77	0.78	0.70	0.66	0.73	0.57	0.61	0.34	0.68
	8	44	38	0.69	0.72	0.68	0.71	0.72	0.78	0.78	0.74	0.79	0.81	0.72	0.75	0.74	0.80
	9	57	11	0.73	0.73	0.70	0.74	0.78	0.80	0.80	0.68	0.69	0.71	0.73	0.70	0.71	0.71
11	1	80	17	0.78	0.79	0.71	0.78	0.79	0.81	0.85	0.61	0.61	0.64	0.55	0.50	0.59	0.62
	2	51	31	0.72	0.76	0.70	0.75	0.78	0.80	0.81	0.61	0.69	0.65	0.60	0.61	0.56	0.66
	3	148	43	0.74	0.74	0.67	0.74	0.76	0.78	0.80	0.63	0.66	0.69	0.64	0.58	0.53	0.65
	4	85	57	0.75	0.77	0.69	0.76	0.78	0.79	0.82	0.54	0.62	0.57	0.63	0.54	0.68	0.60
	5	13	13	0.72	0.75	0.66	0.70	0.73	0.74	0.80	0.04	-0.07	-0.93	-0.15	-0.10	0.68	-0.16
	6	28	37	0.73	0.74	0.66	0.72	0.75	0.78	0.80	0.68	0.58	0.65	0.47	0.65	0.50	0.66
	7	25	29	0.69	0.71	0.64	0.68	0.70	0.73	0.77	0.73	0.65	0.67	0.59	0.51	0.35	0.67
	8	26	42	0.73	0.75	0.67	0.73	0.75	0.76	0.81	0.80	0.78	0.79	0.70	0.71	0.58	0.78
	9	70	14	0.77	0.78	0.68	0.78	0.78	0.81	0.84	0.57	0.52	0.48	0.42	0.49	0.48	0.53

<sup>a</sup> For Students these indices are rater reliabilities and do not reflect error due to the use of only one prompt per student

<sup>b</sup> Building reliabilities were based on correlations between building level trait scores for different prompts. The correlations obtained were adjusted to reflect building reliability if 2 prompts were used at grade 5 and 3 prompts were used at grades 8 and 11. Buildings which used fewer prompts and have fewer students will have lower score reliabilities.

## References

- Feldt, L.S., & Brennan, R.L. (1989). Reliability. In R. Linn (Ed.), Educational Measurement. Phoenix, AZ: Oryx Press.
- Herman, J.L., Aschbacher, P.R., & Winters, L. (1989). A Practical Guide to Alternative Assessment. Alexandria, VA: Association for Supervision and Curriculum Development.
- Kolen, M.J. & Brennan, R.L. (1995). Test Equating: Methods and Practices. New York, NY: Springer-Verlag.
- Livinston, S.A & Lewis, C. (1995). Estimating the Consistency and Accuracy of Classifications Based on Test Scores, Journal of Educational Measurement, 32, 179-197.
- Lord, F.M. (1980). Applications of Item Response Theory to Practical Testing Problems. Hillsdale, NJ: Erlbaum.
- Nunnally, J.C. (1978). Psychometric Theory. New York, NY: McGraw-Hill.
- Reckase, M.D. (1997, March). Statistical Test Specification for Performance Assessment: Is this an Oxymoron? Paper presented at annual convention of American Educational Research Association in Chicago, Illinois.

## **Attachment A:**

### **Illustrative Content Validation and Curricular Alignment Evaluation Form**

The example that follows is taken from the mathematics assessment review activity. Other content area relied on forms comparable to this document.

## **Mathematics Item Review Guidelines**

**Center for Educational Testing and Evaluation  
University of Kansas  
August 1999**

We are in the process of developing items for new assessments to be administered in the spring of 2000 to measure the revised Kansas Mathematics Curriculum Standards at grades 4, 7, and 10. Based on the Standards alone, we have worked with numerous Kansas mathematics educators this summer preparing a pool of items to measure those indicators identified for state assessment. The number of items prepared needs to represent a sufficient pool of questions to produce multiple forms of each grade test.

At this stage in the item development process, we need a formal content validity check of the individual items and items sets measuring each indicator. That is your task during these two days:

- to make judgements about the appropriateness of the items,
- edit and revise items to improve their suitability as measures of the indicators in the new Kansas Mathematics Curriculum Standards, and
- to review the items in terms of their quality and provide input to test design plans.

In the Kansas Standards, the indicators for which items have been developed are categorized under more generalized outcome statements identified as Standards and Benchmarks. In your packet of materials, the items have been arranged in sets by Indicator and labeled by Standard, Benchmark and Indicator. In addition, the indicators are of two types, Knowledge indicators and Application indicators. Our labeling system is to list the Standard first (S), then the Benchmark (B) and then the Indicator with a designation as knowledge level (K) or application (A). For example, the label S2B3K4 refers to Standard #2, Benchmark #3 measuring Knowledge indicator 4.

Be advised: NOT ALL INDICATORS ARE BEING MEASURED, only those designated by triangles in the Curriculum Standards are eligible for the state assessment. A copy of the complete standards is available in your packet. In addition, you will NOT be reviewing all items. You will have an opportunity to review the emerging items for approximately 50% of the Standards (more exactly two of the four Standards) at one grade.

Our plan is for you to complete your review of the items by noon Wednesday. Then in the afternoon, each group will convene to discuss any major problems that are being encountered. At that time, we will also pose a few general questions to get your suggestions on a limited set of test design issues.

## Criteria to follow for the Review:

The items have been arranged in stapled sets by Indicator. Please, do not pull apart the sets. As you start a set, **read** the Standard, Benchmark, Indicator and Content description to be measured by the set of items. Your individual, professional understanding and interpretation of the indicators to be measured are important in this process review.

Once you have familiarized yourself with the indicator, carefully and thoughtfully review each individual item. **For each item**, use the following **criteria** to guide your review.

### **Regarding the content of the item:**

- |   |     |                     |                |
|---|-----|---------------------|----------------|
| 1 Does this item measure <u>content</u> called for by the indicator?                  | Yes | <b>No...*</b>       |                |
| 2 Is the <u>content tested</u> by this item trivial, isolated, or non-generalizeable? | No  | <b>Not Sure...*</b> | <b>Yes...*</b> |

### **Regarding properties of the item:**

- |   |               |  |                                    |
|---|---------------|--|------------------------------------|
| 3 Does the item measure the <u>skill</u> called for by this indicator?  | Yes           | <b>No...*</b>                              |                                    |
| 4 Judge the <u>Reading level</u> of this item for students at this grade:   | Not a problem | <b>Too difficult for many students...*</b> |                                    |
| 5 Evaluate the <u>appropriateness of the difficulty</u> of this item for students at this grade:  | Too easy      | Okay/<br>Acceptable                        | <b>Unnecessarily difficult...*</b> |
| 6 Is the complete item as presented (stem, supporting charts, graphs, illustrations, response choices, etc.) <u>too complex, i.e., beyond the developmental readiness</u> , for students at this grade? | No            | <b>Not Sure...*</b>                        | <b>Yes...*</b>                     |

**\*...Provide a comment or explanation for such judgements. We must have your input to make changes. Offer specific changes whenever possible. Provide your comments, suggestions and observations directly on the page with the item.**

## Reviewing Activity and Considerations

- 1 Show all edits, suggestions, comments directly on the page containing the item.
- 2 Edit items for language and readability to make an item more appropriate for the grade intended. If you can simplify the language or reduce the reading load, please make the suggested edit. However, remember, some language has been used to make the items more authentic or relevant.
- 3 Often the “correct” response has been indicated (\*). Sometimes this designation has been omitted. When you review, check the correct answer, or when not shown, mark the correct answer. Remember: WE USE A MULTIPLE CORRECT FORMAT with some items. Therefore, read the entire problem and evaluate all choices.
- 4 Review the distractors for appropriateness and adequacy. Try to provide/produce better distractors whenever possible (add a brief rationale/justification when doing so).
- 5 Upon review, if you should have a better or another idea for an item, please give us your suggestion(s). Do not be bashful about this: if you have an idea, we want to see it! Write your suggestions directly on the page, or back of the page.
- 6 AFTER you have reviewed, edited and revised all the individual items in an indicator set, think about the items as a collection. You are to make an overall judgement about the adequacy of the collection of items as a set toward measuring the specific indicator. These questions are shown on each cover page that lists the indicator for the set of items. Respond to the two questions on the page after you have completed your review of the individual items for that indicator. (see attached)

### Remember the following:

- Not all indicators are being measured, therefore you will not see some items representing a curriculum you might expect to find; nor are you reviewing all available items. Continuing with this thought...
- We are creating items for the specific indicator; therefore, while you may desire to see a greater range of content or skill tested, or perhaps desired there to be greater or a different focus, remember that what is to be tested is solely defined and determined by the exact wording of the specific indicator. We cannot stray from this statement which is the self contained definition of the knowledge and behavior to be assessed.
- The information tested is intended to represent what all Kansas students should be able to do, and not necessarily what is being taught at this time.
- For the purpose of this review, consider only the regular education student population
- **Test security must be maintained and is essential.** Under no condition are these items to be copied, notes taken about specific questions or groups of items, or the specific content of questions shared with others. Materials are NOT to leave the reviewing room.