

TECHNICAL REPORT

2001 Kansas Assessments in Science and Social Studies*

Prepared by:

Douglas R. Glasnapp and John P. Poggio
Center for Educational Testing and Evaluation
University of Kansas

Table of Contents

| | |
|--|---------|
| Introduction and Background | 2001-2 |
| Development of the Assessments and Content-Related Validity Evidence..... | 2001-7 |
| Differential Item Functioning | 2001-18 |
| Test Equating | 2001-33 |
| Student Performance Level Classification Cut Scores | 2001-47 |
| Science Assessment Test Characteristics..... | 2001-47 |
| Social Studies Assessment Test Characteristics | 2001-47 |
| References..... | 2001-68 |

*See the Year 2000 Technical Report for complete information on the Kansas Reading, Writing and Mathematics Assessments

Introduction and Orientation to the Kansas Assessments

This technical manual provides information on the psychometric properties of the year 2001 *Kansas Assessments in Science and Social Studies*. For a reporting of the technical characteristics of the assessments in reading, writing and mathematics first administered in 2001, the reader is referred to the *Technical Report: Y2000 Kansas Assessments in Mathematics, Reading and Writing* (Glasnapp, Poggio & Omar, 2000).

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 and created beginning in May 1999 and then administered in the Spring 2000. The reading and mathematics tests are administered on an annual basis and were given to students again in the Spring 2001. Grade 5, 8 and 11 students participate in the reading and writing assessments. Grade 4, 7 and 10 students participate in the mathematics assessments. New Kansas assessments in science and social studies were planned and created beginning in May 2000 and then administered in the Spring 2001. Grade 4, 7 and 10 students participate in the science assessments. Grade 6, 8 and 11 students participate in the social studies assessments.

Students to have been tested in each of the four content areas included regular education students, gifted students, students with disabilities and English language learners (ELL). Few students at the designated grade levels are 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 the new cycle of state assessments in reading, writing and mathematics. For science and social studies, the Spring 2001 administration serves as the baseline year. The assessments administered were all 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.

As each new round of assessments were implemented, important changes from pre-Y2000 Kansas assessments were incorporated. Curriculum standards were changed and targets for the assessments restricted, performance assessments were formally abandoned by the Board of Education as part of the formal state assessment program, and test specifications were revised. In effect, no comparison to student, building, district or state performance prior to Y2000 can be made. To achieve a long term assessment and accountability system projected to be in place for a minimum of five academic years, in Y2000 there were four different parallel forms of both the reading and mathematics tests created and administered at each grade level. A single grade level form (one of the four) in reading and mathematics was chosen and administered to all students at the designated grade levels in Spring 2001 -- this plan is then to be repeated annually for three subsequent, that is through spring 2004 testing with the option to use one intact form in spring 2005. In science and social studies a comparable plan was introduced -- there were two different parallel forms created and administered in the baseline year (spring 2001), with each form to be uniquely used on a single occasion in spring 2003 and spring 2005. As was done with reading and mathematics, in the baseline year (spring 2001) the science and social studies assessments were distributed and administered such that equivalent groups of students within classrooms, buildings, districts and across the state took each form of a grade level test. As noted, in subsequent years according to the exposure of forms specified plan, different intact science and social studies forms will be

cycled through the assessment to afford comparisons for growth over time at the school, district and state levels¹. To assure comparability of scores across the different forms of the tests in science and social studies, 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 (2001). 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. The procedures used for equating the science and social studies test forms followed those used in Y2000 for equating the reading and mathematics test forms. Information on equating for the science and social studies test forms is provided in a later section of this technical report.

The science and social studies 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 each content area 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 for the science and social studies assessments mirrors that of the Y2000 technical report for the reading, writing and mathematics assessments. It is organized by first providing information on the item/test development procedures used by CETE to maximize the content validity of 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

¹ The plan for continuing use of a single baseline form (i.e., annual or biennial use of a single form through spring 2005) has been abandoned. Beginning with the spring 2002 administration, KSDE has directed use of all available forms annually. The impact of this decision on student scores, trend information, and local schools’ use of test result information (repeated and frequent exposure of intact forms) are to be monitored and will be addressed in Technical Reports and papers in later years.

first psychometric results provide information on the results from the differential item functioning (DIF) analyses on the science and social studies items. 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 science and social studies 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, Aschbacker, & Winters, 1992). For use in low stakes decisions, a minimum reliability of .70 is used as the criterion for test scores (Nunnally, 1978; Herman, Aschbacker, & Winters, 1992; Feldt & Brennan, 1989; Reckase, 1997). Whether used for high or low stakes educational decisions, there is a strong professional practice belief that a test score alone should not be used without other corroborating information or evidence to make decisions about individuals (AERA, APA, NCME Standards, 1999).

The next section presents information addressing the construction of the Kansas science and social studies 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. In science, 26 indicators are targeted for assessment at grade 4, 28 at grade 7 and 24 at grade 10. In social studies, 30 indicators are targeted at each of the grade levels. 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 is 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 later 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 afar).

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. 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.

Science Assessments: Much state, indeed national attention, was given to the Kansas Curriculum Standards in Science and the coverage of the content Standards on the state's science assessments. To ready the science assessments for spring 2001 administration, starting in late June 2000, items were created by select members of the item development teams to reflect what were the Kansas Board of Education rejected (September 1999) specific science standards and indicators. This development was unsponsored by KSDE or the Board, but rather was carried out by the test development contractor (CETE) using its resources so as to anticipate and plan for what could occur in late 2000. As it turned out, a newly configured Board did revert to the originally recommended science Standards (January 2001) and the science assessments at grades 7 and 10 were amended to provide for coverage of the original skills without compromising the original specifications for the science assessments. The shifting, inclusion and exclusion of science standards at grade 4 had no impact on these particular assessments as originally planned.

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 science and social study items 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 disability 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 science and social studies 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 size 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 5 through 7.

Items. The science items and social studies items from alternate test forms at all grade levels were analyzed for DIF. In science, each form at grades 4, 7 and 10 had 57, 66, and 70 items, respectively. In social study, each form at grades 6, 8 and 11 had 60 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, two items in each of the grade 4 science forms and two items in form 72 and one item in form 81 at grade 6 and one item in form 60 at grade 8 were dropped as poorly functioning items.

Statistical Methods. The main procedure used was the Mantel-Haenszel (MH) technique. 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.

RESULTS

A sample output for a science DIF analysis at grade 4, form 14 is given in Table 1. Ten items were deleted arbitrarily for space conservation. From the DELTA_E column of Table 1, item 4 appeared to show DIF for the female versus male comparison. Chi-squared extreme area probability for this item was less than 0.001. This item seemed to disadvantage females.

Tables 2 through 4 gives a summary of science items flagged by the Mantel-Haenszel procedure for each DIF comparison by form at each of the three grade levels for science and social studies, respectively. Table 2 shows that 1 item was flagged at grade 4 with negative DIF. The item was flagged for the gender DIF comparisons. Table 3 showed 1 flagged item at grade 7, which was positive DIF for the gender comparisons. Table 4 showed 2 flagged items at grade 10. Both were flagged showing negative DIF. One was flagged for gender comparisons while the other for ethnic comparisons.

Tables 5 through 7 summarize these DIF results for social studies items at grade levels 6, 8 and 11. Table 5 shows that 4 items were flagged at grade 6, all but one of which exhibited negative DIF. All items were flagged for ethnic comparisons. Table 6 shows 9 flagged items at grade 8. All of them were flagged for ethnic comparisons. Two of them showed positive DIF while all others exhibited negative DIF. Six out of those nine items were flagged for the comparisons between Asians and White. Table 7 shows 7 flagged items at grade 11, in which 4 items had positive DIF and 3 had negative DIF. All but one were flagged for ethnic comparisons.

DIF SUMMARY

Fewer items were found to exhibit DIF for science than for social studies. 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. 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 Science

MANTEL-HAENSZEL DIF ANALYSIS FOR 2001 science grade 4 Kansas gender dif form 14 (male=2, female=1)

NUMBER OF ITEMS = 57 & CHK = .000

| ITEM | P-2 | PB-2 | P-1 | PB-1 | P-2+1 | PB-2+1 | CHI-I | APLHA-I | DELTA-I | CHI-E | APLHA-E | DELTA-E |
|------|-----|------|-----|------|-------|--------|-------|---------|---------|-------|---------|---------|
| 1 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 |
| 2 | .66 | .35 | .66 | .32 | .66 | .33 | .56 | .94 | .15 | .37 | .95 | .12 |
| 3 | .66 | .16 | .73 | .17 | .70 | .16 | 14.08 | .74 | .72 | 13.51 | .74 | .70 |
| 4 | .74 | .37 | .60 | .40 | .66 | .39 | 62.68 | 1.97 | -1.59 | 66.67 | 1.99 | -1.61 |
| 5 | .81 | .33 | .80 | .33 | .80 | .33 | .01 | 1.01 | -.03 | .11 | 1.04 | -.09 |
| 6 | .71 | .41 | .65 | .39 | .68 | .40 | 7.20 | 1.26 | -.55 | 8.14 | 1.28 | -.57 |
| 7 | .80 | .36 | .80 | .30 | .80 | .33 | .06 | 1.03 | -.07 | .07 | 1.03 | -.07 |
| 8 | .47 | .31 | .44 | .32 | .46 | .31 | .44 | 1.06 | -.13 | .63 | 1.07 | -.15 |
| 9 | .75 | .44 | .72 | .44 | .74 | .44 | 1.39 | 1.12 | -.27 | 1.34 | 1.11 | -.26 |
| 10 | .75 | .38 | .76 | .30 | .75 | .34 | 1.23 | .90 | .25 | .99 | .91 | .22 |
| 11 | .35 | .25 | .36 | .28 | .35 | .27 | 3.13 | .87 | .34 | 2.11 | .89 | .27 |
| 12 | .69 | .53 | .67 | .47 | .68 | .50 | .05 | 1.03 | -.06 | .22 | 1.05 | -.11 |
| 13 | .32 | .26 | .35 | .31 | .34 | .28 | 8.76 | .78 | .58 | 7.64 | .80 | .53 |
| 14 | .89 | .42 | .92 | .39 | .90 | .40 | 6.37 | .70 | .85 | 5.62 | .72 | .79 |
| 15 | .80 | .41 | .85 | .39 | .83 | .40 | 17.66 | .63 | 1.07 | 17.55 | .64 | 1.05 |
| 16 | .75 | .47 | .75 | .44 | .75 | .45 | .45 | .93 | .16 | .39 | .94 | .15 |
| 17 | .80 | .33 | .83 | .27 | .81 | .30 | 5.21 | .79 | .54 | 5.23 | .80 | .53 |
| 18 | .74 | .44 | .74 | .46 | .74 | .45 | 2.37 | .86 | .35 | 2.05 | .87 | .32 |
| 19 | .74 | .37 | .76 | .34 | .75 | .36 | 7.4 | .89 | .27 | 1.51 | .90 | .24 |
| 20 | .91 | .37 | .92 | .32 | .91 | .35 | 2.37 | .79 | .54 | 2.41 | .80 | .53 |
| 21 | .53 | .31 | .56 | .29 | .54 | .30 | 6.57 | .82 | .47 | 5.34 | .84 | .42 |
| 22 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 |
| 23 | .76 | .33 | .66 | .37 | .71 | .35 | 31.59 | 1.63 | -1.15 | 32.58 | 1.63 | -1.15 |
| 24 | .84 | .40 | .84 | .36 | .84 | .38 | .01 | .99 | .03 | .00 | 1.01 | -.03 |
| 25 | .67 | .29 | .67 | .22 | .67 | .25 | .00 | .99 | .02 | .00 | 1.00 | .01 |
| 26 | .62 | .30 | .53 | .30 | .58 | .30 | 19.33 | 1.41 | -.81 | 20.58 | 1.42 | -.82 |
| 27 | .90 | .43 | .90 | .41 | .90 | .42 | .10 | .95 | .13 | .00 | .98 | .04 |
| 28 | .80 | .37 | .76 | .36 | .78 | .37 | 9.50 | 1.34 | -.69 | 9.60 | 1.34 | -.69 |
| 29 | .78 | .48 | .78 | .43 | .78 | .45 | .07 | .97 | .07 | .09 | .97 | .08 |
| 30 | .48 | .30 | .53 | .31 | .50 | .30 | 12.23 | .76 | .64 | 11.16 | .78 | .59 |
| 41 | .78 | .45 | .80 | .41 | .79 | .43 | 6.24 | .78 | .60 | 5.91 | .79 | .57 |
| 42 | .74 | .44 | .75 | .39 | .75 | .41 | 1.60 | .89 | .28 | 1.68 | .89 | .29 |
| 43 | .84 | .33 | .81 | .35 | .82 | .34 | 3.31 | 1.21 | -.45 | 3.58 | 1.22 | -.46 |
| 44 | .77 | .44 | .74 | .38 | .76 | .41 | 4.03 | 1.21 | -.45 | 3.27 | 1.18 | -.40 |
| 45 | .59 | .42 | .56 | .38 | .58 | .40 | 1.24 | 1.10 | -.22 | 1.18 | 1.09 | -.21 |
| 46 | .66 | .44 | .64 | .45 | .65 | .45 | .03 | 1.02 | -.04 | .08 | 1.03 | -.06 |
| 47 | .87 | .30 | .87 | .22 | .87 | .26 | .02 | 1.02 | -.05 | .10 | 1.04 | -.09 |
| 48 | .44 | .17 | .41 | .14 | .43 | .15 | 1.15 | 1.09 | -.20 | 1.20 | 1.09 | -.20 |
| 49 | .49 | .32 | .48 | .34 | .48 | .33 | .10 | .97 | .07 | .00 | 1.00 | .01 |
| 50 | .54 | .41 | .53 | .41 | .53 | .41 | .30 | .95 | .11 | .13 | .97 | .08 |
| 51 | .65 | .48 | .65 | .48 | .65 | .48 | .62 | .93 | .17 | .70 | .93 | .18 |
| 52 | .90 | .51 | .92 | .38 | .91 | .45 | 2.69 | .77 | .61 | 2.66 | .78 | .58 |
| 53 | .80 | .47 | .82 | .44 | .81 | .46 | 5.60 | .77 | .60 | 6.47 | .76 | .63 |
| 54 | .50 | .27 | .50 | .27 | .50 | .27 | .03 | .98 | .04 | .00 | 1.00 | .00 |
| 55 | .80 | .49 | .80 | .48 | .80 | .48 | .76 | .91 | .23 | .81 | .91 | .23 |
| 56 | .46 | .28 | .40 | .33 | .43 | .30 | 6.31 | 1.22 | -.47 | 8.24 | 1.25 | -.52 |
| 57 | .65 | .34 | .65 | .30 | .65 | .32 | .22 | .96 | .10 | .26 | .96 | .10 |

** SUMMARY DATA **

| | MEAN | SD | CASES | KR-20 |
|---------|--------|-------|-------|-------|
| GROUP-2 | 38.357 | 8.892 | 1490. | .883 |
| GROUP-1 | 37.833 | 8.498 | 1577. | .869 |
| TOTAL | 38.088 | 8.695 | 3067. | .876 |

Table 2. Summary of differentially functioning items for Grade 4 Science forms

| Form | Reference | DIF Group | | n | DIF items | n | |
|---------|-----------|-------------|-------|--------------|-----------|-----|---|
| | | N | Focal | | | + | - |
| 14 Male | | 1490 Female | | 1577 NP114 - | 0 +; | 1 - | |
| White | | 1838 Asians | | 290 | | | |
| | | Blacks | | 1001 | | | |
| | | Hispanics | | 881 | | | |
| | | Natives Dn | | 558 | | | |
| | | Natives Sn | | 209 | | | |
| | | | | | 0 +; | 0 - | |
| 23 Male | | 1517 Female | | 1553 | 0 +; | 0 - | |
| White | | 1787 Asians | | 286 | | | |
| | | Blacks | | 1033 | | | |
| | | Hispanics | | 884 | | | |
| | | Natives Dn | | 549 | | | |
| | | Natives Sn | | 209 | | | |
| | | | | | 0 +; | 0 - | |
| | | | | | 0 +; | 1 - | |

Table 3. Summary of differentially functioning items for Grade 7 Science forms

| Form | DIF Group | | | DIF items | n | |
|---------|-----------|-------------|------------|---------------|------|-----|
| | Reference | N | Focal | | | |
| 37 Male | | 1562 Female | | 1637 NP2I12 + | 1 +; | 0 - |
| White | | 1981 | Asians | 324 | | |
| | | | Blacks | 854 | | |
| | | | Hispanics | 851 | | |
| | | | Natives Dn | 780 | | |
| | | | Natives Sn | 151 | | |
| | | | | | 0 +; | 0 - |
| 43 Male | | 1525 Female | | 1664 | 0 +; | 0 - |
| White | | 1891 | Asians | 317 | | |
| | | | Blacks | 835 | | |
| | | | Hispanics | 848 | | |
| | | | Natives Dn | 750 | | |
| | | | Natives Sn | 136 | | |
| | | | | | 0 +; | 0 - |
| | | | | | 1 +; | 0 - |

Table 4. Summary of differentially functioning items for Grade 10 Science forms

| Form | DIF Group | | | DIF items | n | |
|---------|-------------|---|---------------|-----------|------|-----|
| | Reference | n | Focal | | | |
| 56 Male | 1539 Female | | 1591 | | 0 +; | 0 - |
| White | 1576 Asians | | 306 | | | |
| | Blacks | | 785 | | | |
| | Hispanics | | 757 | | | |
| | Natives Dn | | 462 | | | |
| | Natives Sn | | 131 | | | |
| 68 Male | 1451 Female | | 1614 NP3I22 - | | 0 +; | 0 - |
| White | 1660 Asians | | 329 NP1I1 - | | 0 +; | 1 - |
| | Blacks | | 759 | | | |
| | Hispanics | | 756 | | | |
| | Natives Dn | | 437 | | | |
| | Natives Sn | | 123 | | | |
| | | | | | 0 +; | 2 - |

Table 5. Summary of differentially functioning items for Grade 6 Socail study forms

| Form | DIF Group | | | DIF items | n | |
|---------|-------------|------------|-------|-----------|---------|---------|
| | Reference | n | Focal | | | |
| 72 Male | 1597 Female | | 1613 | | 0 +; | 0 - |
| White | 2043 | Asians | 339 | NP1I2 - | NP2I2 - | NP3I2 + |
| | | Blacks | 975 | | | |
| | | Hispanics | 890 | | | |
| | | Natives Dn | 712 | | | |
| | | Natives Sn | 169 | | | |
| 81 Male | 1534 Female | | 1619 | | 1 +; | 2 - |
| White | 1951 | Asians | 278 | NP3I1 - | 0 +; | 1 - |
| | | Blacks | 939 | | 1 +; | 3 - |
| | | Hispanics | 854 | | | |
| | | Natives Dn | 724 | | | |
| | | Natives Sn | 174 | | | |

Table 6. Summary of differentially functioning items for Grade 8 Social study forms

| Form | DIF Group | | | DIF items | n |
|---------|-------------|------------|-------|-------------------------|----------|
| | Reference | n | Focal | | |
| 60 Male | 1573 Female | | 1608 | | 0 +; 0 - |
| White | 1794 | Asians | 301 | NP1I2 - NP2I1 - NP3I2 - | |
| | | Blacks | 865 | | |
| | | Hispanics | 819 | | |
| | | Natives Dn | 562 | | |
| | | Natives Sn | 114 | NP1I1 - | |
| 94 Male | 1559 Female | | 1621 | | 0 +; 4 - |
| White | 1891 | Asians | 277 | NP3I2 - NP3I6 - NP3I9 + | |
| | | Blacks | 844 | NP3I14 ^a + | |
| | | Hispanics | 874 | NP1I1 - | |
| | | Natives Dn | 589 | | |
| | | Natives Sn | 121 | | |
| | | | | | 2 +; 3 - |
| | | | | | 2 +; 7 - |

a: This item has ETS delta value that equals to 1.50, but chi-squared extreme area probability (p) far less than 0.001 (p=1.13799E-08).

Table 7. Summary of differentially functioning items for Grade 11 Social study forms

| Form | DIF Group | | | DIF items | n |
|---------|-------------|------------|-------|-----------|----------|
| | Reference | n | Focal | | |
| 28 Male | 1404 Female | | 1453 | | 0 +; 0 - |
| White | 1363 | Asians | 285 | NP3I15 + | |
| | | Blacks | 640 | NP2I11 - | NP3I1 - |
| | | Hispanics | 585 | | |
| | | Natives Dn | 347 | | |
| | | Natives Sn | 98 | | |
| | | | | | 1 +; 2 - |
| 39 Male | 1403 Female | | 1431 | NP2I14 + | 1 +; 0 - |
| White | 1308 | Asians | 287 | NP1I6 - | NP2I15 + |
| | | Blacks | 629 | NP3I6 + | |
| | | Hispanics | 544 | | |
| | | Natives Dn | 347 | | |
| | | Natives Sn | 105 | | |
| | | | | | 2 +; 1 - |
| | | | | | 4 +; 3 - |

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 8 summarizes the recommendations given by the impacted groups review panel with respect to the negatively functioning items in science and social studies.

Based on these recommendations, no items were dropped during scoring on the Science tests and a total of four items were dropped and not scored on the Social Studies test forms, one item at grade 6, two items at grade 8 and one item at grade 11. 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 few (4) of the Social Studies items identified as negative DIF items. Negatively functioning items were dropped as recommended.

Table 8. Recommendations made by impacted group on negatively functioning items

| Subject | Grade | Form | Part | Item | Impacted Group | Recommendations | |
|--------------|-------|----------|------|-----------|----------------|-----------------|-------------|
| Science | 4 | 14 | 1 | 4 | Female | retain item | |
| | | 23 | | | | | |
| | 7 | 37 43 | 2 | 12 | Female | retain item | |
| Social Study | 10 | 56 | 3 | 22 | Female | retain item | |
| | | 68 | | | Asians | retain item | |
| | | 68 | | | 1 | 1 | Asians |
| | 6 | 72 | 1 | 2 | Asians | Drop item | |
| | | 72 | 2 | 2 | Asians | retain item | |
| | | 72 | 3 | 2 | Asians | retain item | |
| | | 81 | 3 | 1 | Blacks | retain item | |
| | | 8 | 60 | 1 | 2 | Asians | retain item |
| | | | 60 | 2 | 1 | Asians | retain item |
| | | | 60 | 3 | 2 | Asians | retain item |
| 60 | | | 1 | 1 | Natives Sn | retain item | |
| 94 | | | 3 | 2 | Asians | drop item | |
| 94 | | | 3 | 6 | Asians | retain item | |
| 94 | | 3 | 9 | Asians | retain item | | |
| 94 | | 3 | 14 | Blacks | retain item | | |
| 94 | | 1 | 1 | Hispanics | drop item | | |
| 11 | | 28 | 3 | 15 | Asians | retain item | |
| | | 28 | 2 | 11 | Blacks | retain item | |
| | | 28 | 3 | 1 | Blacks | drop item | |
| | | 39 | 2 | 14 | Female | retain item | |
| | 39 | 1 | 6 | Asians | retain item | | |
| | 39 | 2 | 15 | Asians | retain item | | |
| | 39 | 3 | 6 | Blacks | retain item | | |

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 science and social studies, 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 2000, two parallel forms in science 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 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 science, two independent forms (no common items) were developed with similar 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. The numbers of items for the fourth, seventh, and tenth grades were 55, 66, and 70, respectively. All items in science were multiple-choice in format.

In social studies, the same procedure was used to develop two independent forms (no common items). Due to printing errors, item analyses, or differential item functioning analyses, the two forms for social studies ended up with slightly different total number of items between forms. For example, the number of items for the sixth, eighth, and eleventh grade were 57, 59, and 59, respectively for one form, but 59, 58, and 60, respectively, for another form.

Equating Design. The data collection method for the equating was the Random Groups design. The design was implemented by spiraling two different forms at each grade level in both science and social studies during test administration in Kansas classrooms. With approximately 30,000 – 32,000 regular education students taking the test at each grade level, about 15000 - 16000 students took each of the two test forms. This represents a number that is more than adequate for a random groups equating.

Table 9 shows the percentages of students taking alternate forms of the science and social studies assessments across schools in Kansas. For the values in Table 9, 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, percentages based on demographic information supports the equivalence of groups obtained through this data collection design. In other word, data in Table 9 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 science and social studies and observed score frequency

distributions were obtained by summing the compound binomial (or multinomial) distribution across all values of theta. Then, the equi-percentile method of equating was used on the obtained observed score frequency distributions.

In both science and social studies, examinee scores were equated at the process skills (knowledge and process for science, knowledge and application for social studies) and the total score level and expressed in the percent correct metric.

Equating Criteria. Because both classical and IRT test equating were examined, comparisons between several competing methods were possible. These equating methods had to be reviewed and deliberate decisions had to be made as to which method 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 9. 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 |
| Science | 4 | 892 | 14 | 49.8 | 6.5 | 51.2 | 48.8 | 74.5 | 25.5 | 89.9 | 10.1 |
| | | | 23 | 50.2 | 6.5 | 51.1 | 48.9 | 74.7 | 25.3 | 89.2 | 10.8 |
| | | | All | | | 51.2 | 48.8 | 74.6 | 25.4 | 89.6 | 10.4 |
| | 7 | 508 | 37 | 50.7 | 6.1 | 51.2 | 48.8 | 74.3 | 25.7 | 89.1 | 10.9 |
| | | | 43 | 49.3 | 6.1 | 50.5 | 49.5 | 75.0 | 25.0 | 90.4 | 9.6 |
| | | | All | | | 50.9 | 49.1 | 74.7 | 25.3 | 89.7 | 10.3 |
| | 10 | 387 | 56 | 50.3 | 4.7 | 51.7 | 48.3 | 79.1 | 20.9 | 92.0 | 8.0 |
| | | | 68 | 49.8 | 4.7 | 50.9 | 49.1 | 78.7 | 21.3 | 91.2 | 8.8 |
| | | | All | | | 51.3 | 48.7 | 78.9 | 21.1 | 91.6 | 8.4 |
| Social Study | 6 | 658 | 72 | 50.2 | 7.4 | 51.5 | 48.5 | 74.3 | 25.7 | 89.4 | 10.6 |
| | | | 81 | 49.9 | 7.4 | 51.5 | 48.5 | 74.7 | 25.3 | 89.7 | 10.3 |
| | | | All | | | 51.5 | 48.5 | 74.5 | 25.5 | 89.6 | 10.4 |
| | 8 | 509 | 94 | 50.5 | 6.3 | 51.2 | 48.8 | 76.0 | 24.0 | 99.9 | 10.1 |
| | | | 60 | 49.5 | 6.3 | 51.0 | 49.0 | 76.3 | 23.7 | 90.1 | 9.9 |
| | | | All | | | 51.1 | 48.9 | 76.1 | 23.9 | 90.0 | 10.0 |
| | 11 | 384 | 28 | 49.8 | 6.7 | 50.6 | 49.4 | 81.6 | 18.4 | 93.2 | 6.8 |
| | | | 39 | 50.2 | 6.7 | 50.4 | 49.6 | 81.6 | 18.4 | 92.6 | 7.4 |
| | | | All | | | 50.5 | 49.5 | 81.6 | 18.4 | 92.9 | 7.1 |

RESULTS

Table 10 shows a descriptive summary of the equating samples obtained in science. The bolded numbers in the table describe the characteristics of the base form at each grade level. For science, the two test forms were comparable at each grade level (i.e., each had the same number of items, etc.), thus the selection of which form to use as the base form was arbitrary. Table 10 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 10. Descriptive Statistics for Equating Samples for Science Assessments by Test Form and Grade Level

| Grade | Form | N | Knowledge | | | | Process | | | | Total | | | |
|-------|-----------|--------------|-----------|--------------|-------------|-------------|-----------|--------------|-------------|-------------|-----------|--------------|--------------|-------------|
| | | | N | M | SD | Rxx' | n | M | SD | Rxx' | n | M | SD | Rxx' |
| 4 | 14 | 14389 | 20 | 13.37 | 3.38 | 0.70 | 35 | 25.09 | 5.73 | 0.83 | 55 | 38.46 | 8.57 | 0.87 |
| | 23 | 14520 | 20 | 14.34 | 3.32 | 0.71 | 35 | 24.25 | 6.00 | 0.83 | 55 | 38.58 | 8.80 | 0.88 |
| 7 | 37 | 14687 | 32 | 19.87 | 5.46 | 0.80 | 34 | 20.72 | 6.04 | 0.82 | 66 | 40.60 | 10.87 | 0.89 |
| | 43 | 14904 | 32 | 18.54 | 5.44 | 0.78 | 34 | 20.38 | 5.53 | 0.78 | 66 | 38.93 | 10.32 | 0.88 |
| 10 | 56 | 14481 | 49 | 24.30 | 7.40 | 0.82 | 21 | 12.82 | 3.97 | 0.74 | 70 | 37.12 | 10.66 | 0.88 |
| | 68 | 14349 | 49 | 26.03 | 7.36 | 0.81 | 21 | 11.97 | 4.04 | 0.75 | 70 | 37.99 | 10.72 | 0.87 |

Table 11 shows a descriptive summary of the equating samples in social studies. Again, bolded numbers in the table describe characteristics of the base forms. Because the final social studies forms had different numbers of scored items, the forms selected as the base forms were those having the greater number of items. While the application subscale of form 72 in grade 6 had a reliability estimate of .63, the Total scale had a reliability estimate of .83. Since most of the other subscales had reliability indices in the .70s and low .80s, equating at the process skill level did not appear objectionable.

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

| Grade | Form | N | n | Knowledge | | | Application | | | | Total | | | |
|-------|-----------|--------------|-----------|---------------|--------------|-------------|-------------|---------------|--------------|-------------|-----------|---------------|---------------|-------------|
| | | | | M | SD | Rxx' | n | M | SD | Rxx' | n | M | SD | Rxx' |
| 6 | 72 | 14984 | 35 | 20.404 | 5.354 | 0.77 | 22 | 13.011 | 3.295 | 0.63 | 57 | 33.414 | 8.014 | 0.83 |
| | 81 | 14937 | 35 | 21.033 | 5.600 | 0.79 | 24 | 13.552 | 3.999 | 0.71 | 59 | 34.585 | 8.941 | 0.86 |
| 8 | 60 | 14882 | 27 | 15.603 | 4.368 | 0.74 | 32 | 20.552 | 5.583 | 0.81 | 59 | 36.155 | 9.348 | 0.88 |
| | 94 | 15083 | 27 | 15.848 | 4.518 | 0.75 | 31 | 19.744 | 5.426 | 0.80 | 58 | 35.593 | 9.331 | 0.87 |
| 11 | 28 | 13546 | 23 | 12.203 | 3.885 | 0.69 | 36 | 22.279 | 6.355 | 0.83 | 59 | 34.482 | 9.584 | 0.88 |
| | 39 | 13404 | 24 | 12.371 | 3.939 | 0.69 | 36 | 23.701 | 6.939 | 0.87 | 60 | 36.072 | 10.218 | 0.89 |

An example of selected parts of an equating output is given in Table 12 for the science forms at grade 7. The base form for this equating is form 37, the form to be equated is form 43, and the scale to be equated is the science process subscale with 34 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 12 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 the first three moments most dissimilar to that of the base distribution while linear equating provided last moments most dissimilar to that of the base distribution.

Table 12. Moments of the Equated Form 43
by Equating Method

| Test Form/Method | mean | S.D | Skewness | Kurtosis |
|---|---------|--------|----------|----------|
| Old Form: 7F37P n= 14687; New Form: 7F43P n= 14904; | | | | |
| Raw Scores | | | | |
| 7F37P | 20.7224 | 6.0426 | -0.3084 | 2.3631 |
| 7F43P | 20.3831 | 5.5260 | -0.2795 | 2.5306 |
| 7F43P equated to 7F37P | | | | |
| unsmoo | 20.7250 | 6.0311 | -0.3092 | 2.3561 |
| s=0.01 | 20.7220 | 6.0391 | -0.3100 | 2.3609 |
| s=0.05 | 20.7219 | 6.0382 | -0.3107 | 2.3652 |
| s=0.10 | 20.7222 | 6.0354 | -0.3115 | 2.3728 |
| s=0.20 | 20.7226 | 6.0308 | -0.3125 | 2.3847 |
| s=0.30 | 20.7229 | 6.0274 | -0.3132 | 2.3937 |
| s=0.50 | 20.7232 | 6.0221 | -0.3139 | 2.4078 |
| s=0.75 | 20.7234 | 6.0170 | -0.3144 | 2.4216 |
| s=1.00 | 20.7234 | 6.0128 | -0.3146 | 2.4330 |
| linear | 20.7224 | 6.0426 | -0.2795 | 2.5306 |
| irt obs | 20.7895 | 5.9723 | -0.2785 | 2.3809 |

Based on the plots in Figure 1, the equi-percentile equating method appears to provide reasonable conversions at all score levels and is the preferred method. The linear procedure appears to be the least preferred.

Table 13 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 except the linear equating method. Table 14 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 unsmoothed equi-percentile method gave the most parsimonious yet reasonable conversion.

Figure 1. Process Grade 7 Form43 equated to Base Form 37

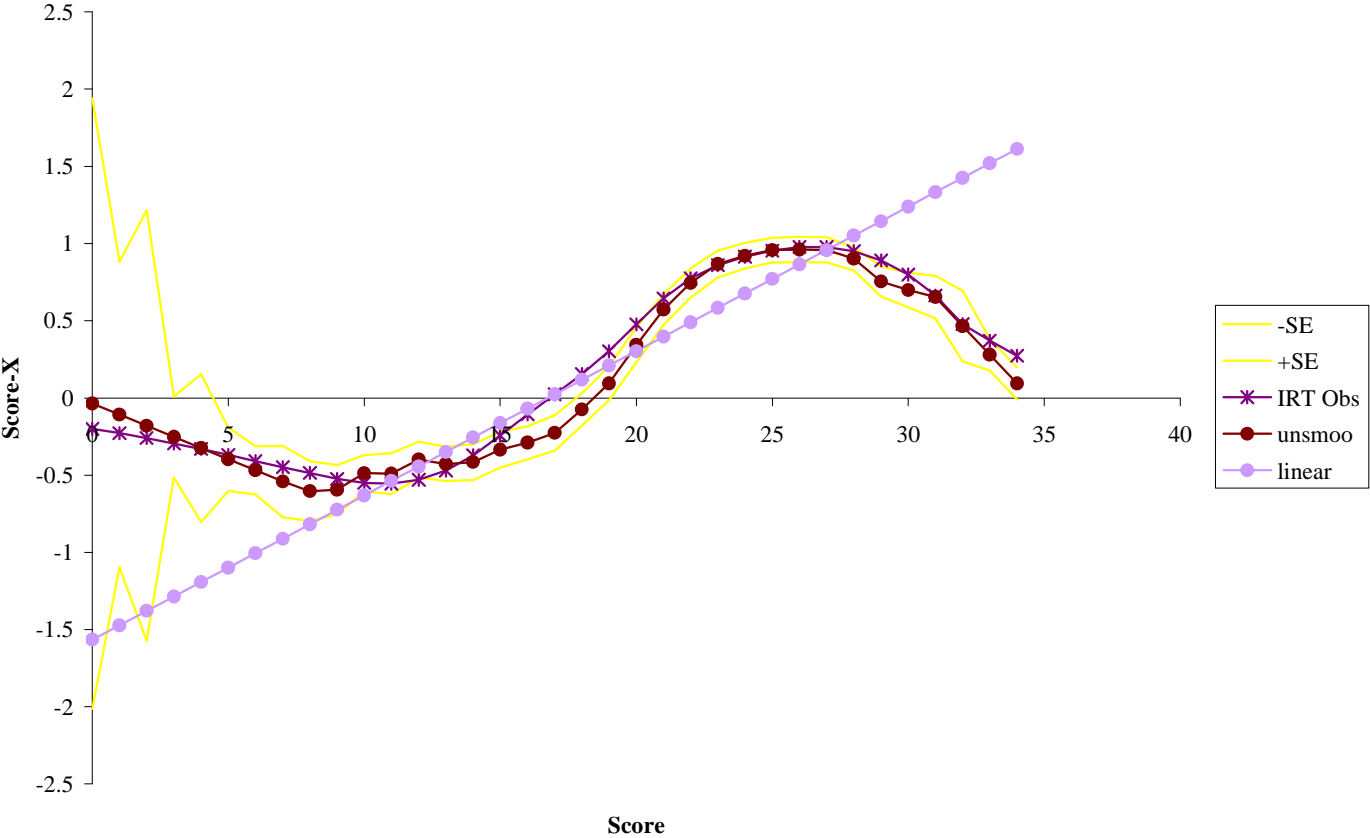


Table 13. Conversion Table for Competing Methods

| 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 | -0.2004 | -0.036 | -0.041 | -0.043 | -0.043 | -0.044 | -0.045 | -0.046 | -0.048 | -0.05 | -1.566 | |
| 1 | 0.7708 | 0.892 | 0.878 | 0.871 | 0.87 | 0.869 | 0.866 | 0.861 | 0.856 | 0.851 | -0.473 | 0 |
| 2 | 1.7394 | 1.82 | 1.797 | 1.785 | 1.784 | 1.781 | 1.777 | 1.769 | 1.759 | 1.751 | 0.621 | 2 |
| 3 | 2.7053 | 2.747 | 2.716 | 2.699 | 2.698 | 2.694 | 2.688 | 2.676 | 2.663 | 2.652 | 1.714 | 2 |
| 4 | 3.6689 | 3.675 | 3.635 | 3.614 | 3.611 | 3.606 | 3.599 | 3.584 | 3.567 | 3.552 | 2.808 | 8 |
| 5 | 4.6306 | 4.603 | 4.554 | 4.528 | 4.525 | 4.519 | 4.509 | 4.491 | 4.47 | 4.452 | 3.901 | 19 |
| 6 | 5.591 | 5.531 | 5.473 | 5.442 | 5.438 | 5.431 | 5.42 | 5.399 | 5.374 | 5.353 | 4.995 | 113 |
| 7 | 6.5517 | 6.459 | 6.402 | 6.374 | 6.371 | 6.363 | 6.35 | 6.325 | 6.297 | 6.271 | 6.088 | 113 |
| 8 | 7.515 | 7.396 | 7.375 | 7.395 | 7.397 | 7.389 | 7.378 | 7.358 | 7.336 | 7.317 | 7.182 | 206 |
| 9 | 8.4765 | 8.406 | 8.396 | 8.423 | 8.425 | 8.416 | 8.407 | 8.392 | 8.376 | 8.364 | 8.275 | 0 |
| 10 | 9.4492 | 9.512 | 9.464 | 9.461 | 9.455 | 9.445 | 9.438 | 9.427 | 9.419 | 9.412 | 9.369 | 266 |
| 11 | 10.4447 | 10.509 | 10.517 | 10.499 | 10.487 | 10.476 | 10.471 | 10.466 | 10.464 | 10.463 | 10.462 | 315 |
| 12 | 11.4698 | 11.601 | 11.557 | 11.533 | 11.518 | 11.509 | 11.507 | 11.509 | 11.514 | 11.519 | 11.556 | 375 |
| 13 | 12.5299 | 12.572 | 12.562 | 12.561 | 12.55 | 12.548 | 12.551 | 12.559 | 12.571 | 12.582 | 12.649 | 431 |
| 14 | 13.6275 | 13.584 | 13.586 | 13.593 | 13.59 | 13.597 | 13.605 | 13.621 | 13.638 | 13.654 | 13.743 | 561 |
| 15 | 14.7562 | 14.665 | 14.647 | 14.639 | 14.646 | 14.662 | 14.675 | 14.697 | 14.72 | 14.739 | 14.836 | 589 |
| 16 | 15.893 | 15.71 | 15.702 | 15.704 | 15.722 | 15.748 | 15.766 | 15.792 | 15.818 | 15.838 | 15.93 | 652 |
| 17 | 17.0238 | 16.774 | 16.782 | 16.801 | 16.829 | 16.86 | 16.88 | 16.908 | 16.933 | 16.952 | 17.023 | 769 |
| 18 | 18.1546 | 17.926 | 17.919 | 17.941 | 17.971 | 18.002 | 18.02 | 18.044 | 18.065 | 18.08 | 18.117 | 847 |
| 19 | 19.303 | 19.094 | 19.102 | 19.124 | 19.146 | 19.168 | 19.181 | 19.197 | 19.21 | 19.219 | 19.21 | 937 |
| 20 | 20.4767 | 20.344 | 20.337 | 20.337 | 20.342 | 20.349 | 20.353 | 20.358 | 20.361 | 20.362 | 20.303 | 956 |
| 21 | 21.6449 | 21.572 | 21.563 | 21.548 | 21.536 | 21.527 | 21.522 | 21.515 | 21.507 | 21.501 | 21.397 | 0 |
| 22 | 22.7742 | 22.744 | 22.744 | 22.725 | 22.704 | 22.684 | 22.672 | 22.655 | 22.639 | 22.625 | 22.49 | 995 |
| 23 | 23.8595 | 23.866 | 23.862 | 23.851 | 23.831 | 23.808 | 23.792 | 23.769 | 23.747 | 23.73 | 23.584 | 1045 |
| 24 | 24.9148 | 24.921 | 24.923 | 24.926 | 24.914 | 24.892 | 24.875 | 24.851 | 24.828 | 24.81 | 24.677 | 994 |
| 25 | 25.9534 | 25.957 | 25.956 | 25.962 | 25.954 | 25.937 | 25.923 | 25.901 | 25.881 | 25.864 | 25.771 | 971 |
| 26 | 26.9767 | 26.961 | 26.966 | 26.966 | 26.96 | 26.948 | 26.938 | 26.922 | 26.907 | 26.894 | 26.864 | 901 |
| 27 | 27.9774 | 27.959 | 27.956 | 27.939 | 27.935 | 27.929 | 27.925 | 27.917 | 27.91 | 27.903 | 27.958 | 763 |
| 28 | 28.9492 | 28.901 | 28.891 | 28.883 | 28.885 | 28.889 | 28.89 | 28.893 | 28.895 | 28.897 | 29.051 | 674 |
| 29 | 29.8902 | 29.753 | 29.783 | 29.809 | 29.82 | 29.833 | 29.842 | 29.856 | 29.869 | 29.881 | 30.145 | 522 |
| 30 | 30.7977 | 30.699 | 30.715 | 30.736 | 30.751 | 30.772 | 30.787 | 30.812 | 30.837 | 30.858 | 31.238 | 373 |
| 31 | 31.6625 | 31.653 | 31.68 | 31.669 | 31.681 | 31.708 | 31.731 | 31.767 | 31.803 | 31.834 | 32.332 | 263 |
| 32 | 32.4779 | 32.466 | 32.507 | 32.496 | 32.504 | 32.523 | 32.538 | 32.564 | 32.588 | 32.609 | 33.425 | 219 |
| 33 | 33.3715 | 33.28 | 33.304 | 33.298 | 33.302 | 33.314 | 33.323 | 33.338 | 33.353 | 33.366 | 34.519 | 21 |
| 34 | 34.2734 | 34.093 | 34.101 | 34.099 | 34.101 | 34.105 | 34.108 | 34.113 | 34.118 | 34.122 | 35.612 | 2 |

EQUATING DECISIONS

Each of the nine equating analyses in science and the nine equatings in social studies was performed and subjected to the criteria previously listed. The selected equating for each scale in social studies and science 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 consist 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 14. 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 | -0.59 | -0.11 | -0.12 | -0.13 | -0.13 | -0.13 | -0.13 | -0.14 | -0.14 | -0.15 | -4.61 |
| 1 | 2.27 | 2.62 | 2.58 | 2.56 | 2.56 | 2.56 | 2.55 | 2.53 | 2.52 | 2.50 | -1.39 |
| 2 | 5.12 | 5.35 | 5.29 | 5.25 | 5.25 | 5.24 | 5.23 | 5.20 | 5.17 | 5.15 | 1.83 |
| 3 | 7.96 | 8.08 | 7.99 | 7.94 | 7.94 | 7.92 | 7.91 | 7.87 | 7.83 | 7.80 | 5.04 |
| 4 | 10.79 | 10.81 | 10.69 | 10.63 | 10.62 | 10.61 | 10.59 | 10.54 | 10.49 | 10.45 | 8.26 |
| 5 | 13.62 | 13.54 | 13.39 | 13.32 | 13.31 | 13.29 | 13.26 | 13.21 | 13.15 | 13.09 | 11.47 |
| 6 | 16.44 | 16.27 | 16.10 | 16.01 | 15.99 | 15.97 | 15.94 | 15.88 | 15.81 | 15.74 | 14.69 |
| 7 | 19.27 | 19.00 | 18.83 | 18.75 | 18.74 | 18.71 | 18.68 | 18.60 | 18.52 | 18.44 | 17.91 |
| 8 | 22.10 | 21.75 | 21.69 | 21.75 | 21.76 | 21.73 | 21.70 | 21.64 | 21.58 | 21.52 | 21.12 |
| 9 | 24.93 | 24.72 | 24.69 | 24.77 | 24.78 | 24.75 | 24.73 | 24.68 | 24.64 | 24.60 | 24.34 |
| 10 | 27.79 | 27.98 | 27.84 | 27.83 | 27.81 | 27.78 | 27.76 | 27.73 | 27.70 | 27.68 | 27.56 |
| 11 | 30.72 | 30.91 | 30.93 | 30.88 | 30.84 | 30.81 | 30.80 | 30.78 | 30.78 | 30.77 | 30.77 |
| 12 | 33.73 | 34.12 | 33.99 | 33.92 | 33.88 | 33.85 | 33.84 | 33.85 | 33.86 | 33.88 | 33.99 |
| 13 | 36.85 | 36.98 | 36.95 | 36.94 | 36.91 | 36.91 | 36.91 | 36.94 | 36.97 | 37.01 | 37.20 |
| 14 | 40.08 | 39.95 | 39.96 | 39.98 | 39.97 | 39.99 | 40.01 | 40.06 | 40.11 | 40.16 | 40.42 |
| 15 | 43.40 | 43.13 | 43.08 | 43.06 | 43.08 | 43.12 | 43.16 | 43.23 | 43.29 | 43.35 | 43.64 |
| 16 | 46.74 | 46.21 | 46.18 | 46.19 | 46.24 | 46.32 | 46.37 | 46.45 | 46.52 | 46.58 | 46.85 |
| 17 | 50.07 | 49.34 | 49.36 | 49.41 | 49.50 | 49.59 | 49.65 | 49.73 | 49.80 | 49.86 | 50.07 |
| 18 | 53.40 | 52.72 | 52.70 | 52.77 | 52.86 | 52.95 | 53.00 | 53.07 | 53.13 | 53.18 | 53.29 |
| 19 | 56.77 | 56.16 | 56.18 | 56.25 | 56.31 | 56.38 | 56.41 | 56.46 | 56.50 | 56.53 | 56.50 |
| 20 | 60.23 | 59.84 | 59.81 | 59.81 | 59.83 | 59.85 | 59.86 | 59.88 | 59.89 | 59.89 | 59.71 |
| 21 | 63.66 | 63.45 | 63.42 | 63.38 | 63.34 | 63.31 | 63.30 | 63.28 | 63.26 | 63.24 | 62.93 |
| 22 | 66.98 | 66.89 | 66.89 | 66.84 | 66.78 | 66.72 | 66.68 | 66.63 | 66.59 | 66.54 | 66.15 |
| 23 | 70.18 | 70.19 | 70.18 | 70.15 | 70.09 | 70.02 | 69.98 | 69.91 | 69.84 | 69.79 | 69.36 |
| 24 | 73.28 | 73.30 | 73.30 | 73.31 | 73.28 | 73.21 | 73.16 | 73.09 | 73.02 | 72.97 | 72.58 |
| 25 | 76.33 | 76.34 | 76.34 | 76.36 | 76.34 | 76.29 | 76.24 | 76.18 | 76.12 | 76.07 | 75.80 |
| 26 | 79.34 | 79.30 | 79.31 | 79.31 | 79.29 | 79.26 | 79.23 | 79.18 | 79.14 | 79.10 | 79.01 |
| 27 | 82.29 | 82.23 | 82.22 | 82.17 | 82.16 | 82.14 | 82.13 | 82.11 | 82.09 | 82.07 | 82.23 |
| 28 | 85.14 | 85.00 | 84.97 | 84.95 | 84.96 | 84.97 | 84.97 | 84.98 | 84.99 | 84.99 | 85.44 |
| 29 | 87.91 | 87.51 | 87.60 | 87.67 | 87.71 | 87.74 | 87.77 | 87.81 | 87.85 | 87.89 | 88.66 |
| 30 | 90.58 | 90.29 | 90.34 | 90.40 | 90.44 | 90.51 | 90.55 | 90.62 | 90.70 | 90.76 | 91.88 |
| 31 | 93.13 | 93.10 | 93.18 | 93.14 | 93.18 | 93.26 | 93.33 | 93.43 | 93.54 | 93.63 | 95.09 |
| 32 | 95.52 | 95.49 | 95.61 | 95.58 | 95.60 | 95.66 | 95.70 | 95.78 | 95.85 | 95.91 | 98.31 |
| 33 | 98.15 | 97.88 | 97.95 | 97.94 | 97.95 | 97.98 | 98.01 | 98.05 | 98.10 | 98.14 | 101.53 |
| 34 | 100.80 | 100.27 | 100.30 | 100.29 | 100.30 | 100.31 | 100.32 | 100.33 | 100.35 | 100.36 | 104.74 |

Table 15 shows the type of equating decisions taken for science. Except for the linear method, all methods required the use of conversion tables. These conversion tables are not provided here to conserve space. After comparing the several methods, a smoothed equi-percentile equating method was chosen for six of the equatings. The unsmoothed equi-percentile equating method was chosen three times. Conversion tables related to these methods were saved and used for the equating subroutines implemented for the 2001 assessment year.

Table 15. Summary of Equating Decisions for Science

| Grade | Form | | Scale ^a | Equating Decision |
|-------|------|---------|--------------------|----------------------------------|
| | Base | Equated | | |
| 4 | 14 | 23 | Knowledge | Unsmoothed Equipercentile |
| | | | Process | Smoothed Equipercentile (s=0.05) |
| | | | Total | Smoothed Equipercentile (s=0.01) |
| 7 | 37 | 43 | Knowledge | Unsmoothed Equipercentile |
| | | | Process | Unsmoothed Equipercentile |
| | | | Total | Smoothed Equipercentile (s=0.10) |
| 10 | 56 | 68 | Knowledge | Smoothed Equipercentile (s=0.05) |
| | | | Process | Smoothed Equipercentile (s=0.01) |
| | | | Total | Smoothed Equipercentile (s=0.05) |

^a All scales were equated at the raw score level and then expressed in the percent correct metric.

Table 16 shows the types of equating chosen for social studies. After comparing several methods, the unsmoothed equi-percentile equating method was chosen four times and a smoothed equi-percentile equating method chosen in the other five instances. Conversion tables related to these methods were saved and used for the equating subroutines implemented for the 2001 assessment year.

Table 16. Summary of Equating Decisions for Social Study

| Grade | Form | | Scale ^a | Equating Decision |
|-------|------|---------|--------------------|----------------------------------|
| | Base | Equated | | |
| 6 | 81 | 72 | Knowledge | Unsmoothed Equipercentile |
| | | | Process | Unsmoothed Equipercentile |
| | | | Total | Smoothed Equipercentile (s=0.10) |
| 8 | 60 | 94 | Knowledge | Unsmoothed Equipercentile |
| | | | Process | Unsmoothed Equipercentile |
| | | | Total | Smoothed Equipercentile (s=0.01) |
| 11 | 39 | 28 | Knowledge | Smoothed Equipercentile (s=0.01) |
| | | | Process | Smoothed Equipercentile (s=0.20) |
| | | | Total | Smoothed Equipercentile (s=0.10) |

^a All scales were equated at the raw score level and then expressed in the percent correct metric.

SUMMARY AND DISCUSSION

The reliabilities for science process skills (Knowledge and Process) and Total scores appeared acceptable for equating purposes. Although one application subscale at grade 6 in social studies had a reliability value below the .70 cut-off, all other scales were adequate for equating. Therefore, equating of social studies at the subscale level did not appear objectionable. In addition, data collected from the spring 2001 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 science and social studies 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, a few items were lost after the spring 2001 administration of the tests. A few items were dropped after item analyses, a couple of items were lost due to printing errors, and a few 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 revised or replaced with like items and maintained for future administration so that the equating relationship derived here would still apply.

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, substantial 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 Classification Decision Points

As part of legislative and State Board of Education mandates, performance category criteria were to be determined for each of the assessments given by the state. For students taking any of the state's assessments, five categories to designate student performance have been established by KSBE: Advanced, Proficient, Satisfactory, Basic, and Unsatisfactory.

To assign student scores attained on the state's assessments in science and social studies to these categories, test score decision points were identified to set the rules for classification of students. Decision points for each of the science and social studies tests were determined by the Kansas State Department of Education so that the percentage distribution in categories would mirror the percentage classifications for reading and mathematics at a grade level as closely as possible.

Table 17 gives the proficiency level decision point scores for each grade level test in science and social studies 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 decision points indicated.

Table 17. Classification Decision Point Scores Based on Total Percent Correct Scores

| Content Area and Grade Level | Advanced | Proficient | Satisfactory | Basic | Unsatisfactory |
|-------------------------------------|-----------------|-------------------|---------------------|--------------|-----------------------|
| Science | | | | | |
| Grade 4 | 81 | 68 | 59 | 42 | <42 |
| Grade 7 | 71 | 63 | 53 | 42 | <42 |
| Grade 10 | 71 | 63 | 53 | 42 | <42 |
| Social Studies | | | | | |
| Grade 6 | 71 | 63 | 53 | 42 | <42 |
| Grade 8 | 71 | 63 | 53 | 42 | <42 |
| Grade 11 | 71 | 63 | 53 | 42 | <42 |
| | | | | | |

As examples, students who had total scores of 81 percent correct or higher on the 4th grade science test would be classified as “Advanced”; students who had total scores of 68 percent correct or higher, but below 81 percent correct would be classified as “Proficient”; students who had total scores of 59 percent correct or above, but below 68 percent correct would be classified as “Satisfactory”; students who had total scores of 42 percent correct or above, but below 59 percent correct would be classified as “Basic”; and students with percent correct total scores below 42 would be classified as “Unsatisfactory.”

Science Assessment Test Characteristics

As identified previously, the science assessment at the 4th, 7th, and 10th grades consisted of three administered parts which contained objective items in a single-correct multiple-choice format. Each administered part was designed for a testing period of 30-40 minutes. In addition, at each grade level two parallel forms of the assessments were administered in a spiraled design in classrooms in Kansas.

Two subskill scores (Knowledge and Process) and a total score were reported for the Kansas Science Assessment. The reported score for each student was the percentage of total points attained based on points available for each score. The total percent correct score is arrived at by adding the scores on all knowledge and process items and expressing this score as a percentage of the total number of items on knowledge and process combined. On the grade 4 test forms, there was an approximate 30 – 70 percent split of knowledge and process items, respectively. On the grade 7 test forms, there was an approximate 50 – 50 percent split of knowledge and process items, respectively. On the grade 10 test forms, there was an approximate 70 – 30 percent split of knowledge and process items, respectively. The exact number of items of each type on each grade level form is given in Table 19. For a descriptions of the indicators that determined the item focus of each score, the Kansas Science Curricular Standards (February, 2001) should be referenced.

Tables 18 and 19 report summary psychometric findings for the science assessment. Table 18 identifies the mean student percent correct across the state at each grade level for the two subskill scores (Knowledge and Process) and the science Total score.

Table 18
Science Assessment Descriptive Statistics of Percent Correct^a
Mean (Standard Deviation) for Spring 2001 Scores Of General Assessment Students

| Grade | Number Tested ^b | Subscale Scores | | Total |
|-------|----------------------------|-----------------|---------------|---------------|
| | | Knowledge | Process | |
| 4 | 32933 | 65.64 (17.37) | 70.41 (16.97) | 68.64 (16.15) |
| 7 | 33035 | 61.08 (17.52) | 59.73 (18.15) | 60.40 (16.93) |
| 10 | 31900 | 49.02 (15.14) | 60.25 (19.15) | 52.34 (15.41) |

^a Values are mean equated percent of points available

^b 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. Within grade levels, there appears to be little difference in the performance on the Knowledge and Process items for grade 4 and grade 7 students. However, the Knowledge items appear to be far more difficult for grade 10 students than are the Process items.

Table 19 provides student and building level reliability coefficients for the 2001 science 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 science Total scores show evidence of a high level of reliability for the intended purposes of the testing program with coefficients at .87, .88 or .89 for all forms and grade levels. The Knowledge and Process scores show satisfactory reliability as well with coefficients ranging from .70 to .82 for the Knowledge subscales and from .74 to .83 for the Process subscales.

Table 19
Science Assessment
Student^a and Building^b Level Reliabilities for Spring 2001 Scores

| Grade | Form | Mean class size | Scores | | | | | | | | |
|-------|------|-----------------|------------|-----------------------|-----------------------|------------|-----------------------|-----------------------|------------|-----------------------|-----------------------|
| | | | Knowledge | | | Process | | | Total | | |
| | | | N of Items | Student Reliabilities | Building ^c | N of Items | Student Reliabilities | Building ^c | N of Items | Student Reliabilities | Building ^c |
| 4 | 14 | 18 | 20 | 0.70 | 0.83 | 35 | 0.83 | 0.82 | 55 | 0.87 | 0.84 |
| | 23 | 19 | 20 | 0.71 | 0.84 | 35 | 0.83 | 0.84 | 55 | 0.88 | 0.85 |
| | all | 37 | | - | 0.90 | | - | 0.90 | | - | 0.91 |
| 7 | 37 | 33 | 32 | 0.80 | 0.88 | 34 | 0.82 | 0.89 | 66 | 0.89 | 0.89 |
| | 43 | 32 | 32 | 0.78 | 0.88 | 34 | 0.78 | 0.88 | 66 | 0.88 | 0.89 |
| | all | 65 | | - | 0.92 | | - | 0.93 | | - | 0.93 |
| 10 | 56 | 42 | 49 | 0.82 | 0.90 | 21 | 0.74 | 0.90 | 70 | 0.88 | 0.91 |
| | 68 | 41 | 49 | 0.81 | 0.90 | 21 | 0.75 | 0.90 | 70 | 0.87 | 0.90 |
| | all | 83 | | - | 0.94 | | - | 0.94 | | - | 0.95 |

^a Student level reliabilities are estimated using coefficient alpha.

^b 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.

^c 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 (37 for grade 4, 65 for grade 7 and 83 for grade 10). These reliability values are all quite high with all being .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-half of the students in the average size building. Thus, building score reliabilities are reported for building mean scores for small class sizes of 18-19 at grade 4, 32-33 at grade 7 and 41-42 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. One would expect the classification reliability to generally be greater for the extreme advanced

and unsatisfactory categories as there is more opportunity for misclassification in the middle categories that have levels both below and above them. However, this is not always the case for these data, particularly for the basic category classification that has some of the higher coefficients.

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. All coefficients for this classification decision are in the .90s with one exception in the upper .80s, thus providing confidence that appropriate decisions are being made for students classified as Unsatisfactory.

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

| Performance Level Classification | Grade 4 Tests | | Grade 7 Tests | | Grade 10 Tests | |
|-------------------------------------|---------------|----------|---------------|----------|----------------|----------|
| | Accuracy | Consist. | Accuracy | Consist. | Accuracy | Consist. |
| Advanced | .61 | .46 | .78 | .62 | .80 | .68 |
| Proficient | .59 | .49 | .65 | .54 | .59 | .47 |
| Satisfactory | .61 | .49 | .60 | .50 | .58 | .46 |
| Basic | .70 | .60 | .84 | .79 | .57 | .46 |
| Unsatisfactory | .84 | .72 | .71 | .46 | .86 | .77 |
| Overall | .65 | .54 | .73 | .63 | .68 | .58 |

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

| Dichotomous Classification | Grade 4 Tests | | Grade 7 Tests | | Grade 10 Tests | |
|-------------------------------|---------------|----------|---------------|----------|----------------|----------|
| | Accuracy | Consist. | Accuracy | Consist. | Accuracy | Consist. |
| Advanced vs Others | .91 | .88 | .93 | .90 | .95 | .92 |
| Adv,Prof vs Others | .88 | .84 | .90 | .87 | .91 | .88 |
| Adv,Prof,Sat vs Other | .90 | .86 | .91 | .87 | .90 | .86 |
| Others vs Unsatisfact | .95 | .93 | .98 | .97 | .91 | .87 |

Social Studies Assessment Test Characteristics

As with science, the social studies assessment at the 6th, 8th, and 11th grades consisted of three administered parts which contained objective items in a single-correct multiple-choice format. Each administered part was designed for a testing period of 30-40 minutes. In addition, at each grade level two parallel forms of the assessments were administered in a spiraled design in classrooms in Kansas.

Two subskill scores (Knowledge and Applications) and a total score were reported for the Kansas Social Studies Assessment. The reported score for each student was the percentage of total points attained based on points available for each score. The total percent correct score is arrived at by adding the scores on all knowledge and applications items and expressing this score as a percentage of the total number of items on knowledge and applications combined. On the grade 6 test forms, there was an approximate 65 – 35 percent split of knowledge and applications items, respectively. On the grade 8 test forms, there was an approximate 50 – 50 percent split of knowledge and applications items, respectively. On the grade 11 test forms, there was an approximate 735– 65 percent split of knowledge and applications items, respectively. The exact number of items of each type on each grade level form is given in Table 23. For a descriptions of the indicators that determined the item focus of each score, the Kansas Curricular Standards for Civics-Government, Economics, Geography and History should be referenced.

Tables 22 and 23 report summary psychometric findings for the social studies assessment. Table 22 identifies the mean student percent correct across the state at each grade level for the two subskill scores (Knowledge and Applications) and the social studies Total score.

Table 22
Social studies Assessment Descriptive Statistics of Percent Correct^a
Mean (Standard Deviation) for Spring 2001 Scores Of General Assessment Students

| Grade | Number Tested ^b | Subscale Scores | | Total |
|-------|----------------------------|-----------------|---------------|---------------|
| | | Knowledge | Applications | |
| 4 | 32978 | 59.29 (16.25) | 55.89 (16.70) | 57.91 (15.37) |
| 7 | 32844 | 57.12 (16.38) | 63.47 (17.75) | 60.57 (16.14) |
| 10 | 29393 | 51.14 (16.48) | 65.20 (19.49) | 59.57 (17.18) |

^a Values are mean equated percent of points available

^b 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 22.

However, the Application items appear to be easier than Knowledge items for students at grades 8 and 11 while there is a slight tendency for the reverse to be true at grade 6.

Table 23 provides student and building level reliability coefficients for the 2001 social studies assessment percent correct scores for each test form at each of the three grade levels. Based on the values in Table 23, all score reliabilities achieved acceptable levels. Student level social studies 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 .83 to a high of .89 across all forms and grade levels. The Knowledge and Applications scores generally show satisfactory reliability as well with coefficients ranging from .69 to .79 for the Knowledge subscales and from .63 to .87 for the Applications subscales. Caution should be used when interpreting or making decisions based on the score values for the forms with reliability estimates in the .60s.

Table 23
Social Studies Assessment
Student^a and Building^b Level Reliabilities for Spring 2001 Scores

| Grade | Form | Mean class size | Scores | | | | | | | | |
|-------|------|-----------------|-----------------------|---------------|---------|-----------------------|---------------|---------|-----------------------|---------------|------|
| | | | Knowledge | | | Application | | | Total | | |
| | | | N of Items | Reliabilities | | N of Items | Reliabilities | | N of Items | Reliabilities | |
| | | Student | Building ^c | | Student | Building ^c | | Student | Building ^c | | |
| 6 | 72 | 25 | 35 | 0.77 | 0.86 | 22 | 0.63 | 0.84 | 57 | 0.83 | 0.86 |
| | 81 | 25 | 35 | 0.79 | 0.87 | 24 | 0.71 | 0.84 | 59 | 0.86 | 0.87 |
| | All | 50 | - | - | 0.92 | - | - | 0.90 | - | - | 0.92 |
| 8 | 60 | 32 | 27 | 0.74 | 0.90 | 32 | 0.81 | 0.89 | 59 | 0.88 | 0.90 |
| | 94 | 33 | 27 | 0.75 | 0.92 | 31 | 0.80 | 0.90 | 58 | 0.87 | 0.92 |
| | All | 65 | - | - | 0.95 | - | - | 0.94 | - | - | 0.95 |
| 11 | 28 | 38 | 23 | 0.69 | 0.88 | 36 | 0.83 | 0.88 | 59 | 0.88 | 0.89 |
| | 39 | 38 | 24 | 0.69 | 0.87 | 36 | 0.87 | 0.89 | 60 | 0.89 | 0.89 |
| | all | 76 | - | - | 0.92 | - | - | 0.93 | - | - | 0.93 |

^a Student level reliabilities are estimated using coefficient alpha.

^b 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.

^c 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 23 for each grade level set of information is the reliability of building means for a building with an average class size for the state (50 for grade 6, 65 for grade 8 and 76 for grade 11). These reliability values are all quite high with all being .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-half of the students in the average size building. Thus, building score reliabilities are reported for building mean scores for smaller class sizes of 25 at grade 6, 32-33 at grade 8 and 38 at grade 10. These coefficients may be used to provide guidance to smaller school buildings in judging the reliability of their building mean scores.

As for science, performance level classifications are also made for each student based on their total social studies percent correct score. Estimates of the classification reliability were made applying the same procedure as used for the science 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 tends to be 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. All coefficients for this classification decision are in the .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 social studies test total scores.

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

| Performance Level Classification | Grade 5 Tests | | Grade 8 Tests | | Grade 11 Tests | |
|-------------------------------------|---------------|----------|---------------|----------|----------------|----------|
| | Accuracy | Consist. | Accuracy | Consist. | Accuracy | Consist. |
| Advanced | .72 | .54 | .72 | .58 | .82 | .64 |
| Proficient | .67 | .55 | .70 | .58 | .64 | .53 |
| Satisfactory | .49 | .41 | .54 | .44 | .62 | .51 |
| Basic | .68 | .57 | .64 | .54 | .65 | .55 |
| Unsatisfactory | .80 | .63 | .84 | .70 | .82 | .71 |
| Overall | .64 | .53 | .67 | .56 | .68 | .57 |

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

| Dichotomous Classification | Grade 5 Tests | | Grade 8 Tests | | Grade 11 Tests | |
|-------------------------------|---------------|----------|---------------|----------|----------------|----------|
| | Accuracy | Consist. | Accuracy | Consist. | Accuracy | Consist. |
| Advanced vs Others | .94 | .92 | .94 | .91 | .95 | .93 |
| Adv,Prof vs Others | .87 | .83 | .89 | .85 | .91 | .87 |
| Adv,Prof,Sat vs Other | .87 | .83 | .90 | .85 | .89 | .85 |
| Others vs Unsatisfact | .94 | .91 | .93 | .90 | .93 | .90 |

References

- AERA, APA, NCME. Standard for Educational and Psychological Testing. 1999.
- Feldt, L.S., & Brennan, R.L. (1989). Reliability. In R. Linn (Ed.), Educational Measurement. Phoenix, AZ: Oryx Press.
- Glasnapp, D. R., Poggio, J. P., & Omar, H. (2000). Technical Report: Y2000 Kansas Assessments in Mathematics, Reading and Writing. Lawrence, KS, Center for Educational Testing and Evaluation, The University of Kansas.
- 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.