Kansas Assessments in Science

2008

TECHNICAL MANUAL

for the

Kansas General Assessments

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

with assistance of
Jorge Carvajal, Pui Chi Chiu,
and Brooke L. Nash

Center for Educational Testing and Evaluation
The University of Kansas

2009

Table of Contents

The Kansas Assessments in Science

PURPOSE OF THE TECHNICAL REPORT

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

Information is provided to address the technical quality of the assessments developed to measure science learning outcomes specific to population of Kansas students. The Kansas General Assessments are intended for administration to students in general or regular education classes whose educational programs are not regulated by IEPs. The main body of this report addresses technical aspects, focusing on scores from the Kansas General Assessments.

The remainder of this report first presents an overview of the 2008 Kansas Assessment Program to provide a context for reviewing information. Next, the test development procedures aimed at maximizing the validity of the assessments as measures of the targeted indicators in the state's Curricular Standards are presented. Then, results from various psychometric analyses are presented in the sequence in which they were conducted for decision-making. The first psychometric results provide information from the differential item functioning (DIF) analyses. These analyses were initially conducted 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 general assessment test forms are presented. The equating results are followed by a discussion of the standard setting analyses and procedures implemented to determine score ranges for classifying students into one of five performance levels defined by the state. Information on score and performance classification reliability follows the section on standard setting. The final section presents evidence from a variety of validity studies, providing information on both internal and external sources of score validity.

Section 1

INTRODUCTION AND ORIENTATION

This technical manual provides information on the psychometric properties of the 2008 Kansas Science Assessments. The purposes of these assessments are to:

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

As background, new Kansas Science Assessments were planned and developed, and then administered for the first time in Spring 2008. WestEd served as the contractor for the development of test items based on test specifications provided by the Kansas State Department of Education (KSDE). The Center for Educational Testing and Evaluation (CETE) at The University of Kansas served as the contractor for all other aspects of the program. Students in grades 3-8 and 11 participated in the assessments. Regular education students, gifted students, students with disabilities, and English language learners (ELL) were all to be tested. Some students at the designated grade levels were exempted from participating in the state assessment programs based on guidelines set 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 2008 administration of the Kansas assessments serves as the baseline for the new cycle of state assessments. The assessments administered were all newly developed to measure the new targeted indicators (learning outcomes) in the most recent editions of the Kansas Curricular Standards for the content areas. These documents should be referenced when examining and evaluating any of the information resulting from the state assessment program. The Curricular Standards serve as the basis for what is assessed by the tests, and any interpretation and subsequent action based on student or group performance on these tests must focus on the assessed standards, benchmarks, and indicators. Copies of the Kansas Curricular Standards are available from the KSDE website at www.ksde.org.

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

To achieve a long term assessment and accountability system projected to be in place for a minimum of five academic years, four different parallel forms of the science general assessment tests were created and administered at each grade level. The tests were distributed and administered so that score equating across forms could occur using an equivalent random groups design. In subsequent years, different intact forms will be cycled through the assessment to compare performance over time at the school district and state level. To assure comparability of scores across the different forms of the science tests, 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 (2008). 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 a grade's test forms for differing difficulties and score variability. Information on equating is provided in a later section of this technical report (see Test Equating, Section 4, Page 19).

The Kansas assessments are planned and created to reflect and otherwise operationalize certain grade level learning outcomes that should serve as curricular and instructional targets in Kansas K-12 schools. As in previous years, the assessments have provided information to contribute to ongoing school accreditation status, and results from the reading and mathematics assessments have a primary role in monitoring annual yearly progress (AYP) as part of the federally mandated *No Child Left Behind* assessment requirements. As related to the accountability demands, cut scores on each test were determined in order to classify students into one of five performance categories (Exemplary, Exceeds Standard, Meets Standard, Approaches Standard, and Academic Warning). The proportion of students classified in these categories becomes a primary source of information in determining AYP for schools, districts, and the state. Section 5 of this report provides additional details on the procedures put in place to set the specific test score criteria used to classify students into one of the performance categories established by the state.

As a final important aspect of the Kansas Assessment Program, administration of the tests are offered under one of two modes on a voluntary basis, a paper and pencil (P&P) test administration mode or an online administration using the Kansas Computerized Assessment (KCA) system developed by the Center for Educational Testing and Evaluation at The University of Kansas. Documentation describing the KCA system may be found at www.kca.cete.us. Approximately 80 percent of the eligible students in grades 3 through 8 and 85% of high school students took the 2008 science tests online using KCA. Studies addressing issues of mode comparability have been on-going and continue as part of the program. Results of initial studies may be found in Poggio, Glasnapp, Yang, & Poggio (2005) and Poggio, Glasnapp, Yang, Beauchamp, & Dunham (2005). These studies are not included as part of this Technical Manual.

Section 2

TEST DEVELOPMENT AND CONTENT REPRESENTATION

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

Test Specifications

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

Item Type

The multiple choice item type is the only item type used on the Kansas General Science Assessments. For all multiple choice items appearing on any general assessment test form, students select the one best answer from among four choices provided.

Item Development

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

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

Item Delivery and Tryouts

All science items that were approved were delivered via electronic upload to the CETE server. Items received were subjected to reviews by CETE staff prior to being assembled onto pilot forms that would be administered in field tests to representative samples of Kansas students. From CETE reviews, where gaps or shortages in the item pool were identified based on the table of specifications, specific requests were made for additional items at the indicator level so that multiple operational test forms at a grade level in a content area could ultimately be constructed.

All Kansas schools were encouraged and invited to participate in the science assessment pilot testing. Due to the large number of science items to be piloted, a fourth test session was added to 2007 state testing of math and reading at all grade levels. For grades 4, 7, and high school, the students took science pilot items, and for the remaining grades, students took history and government, mathematics, or reading items. All items in the science item pool supplied to CETE were piloted. Science pilot tests were administered via the KCA and P&P delivery modes. For the P&P pilot forms, items were randomly selected and assigned to forms. Ten forms per grade level were created and administered via P&P mode. The remaining items also randomly selected and assigned to forms and were piloted via computer (KCA). At 4^{th} grade, there were 18 forms; at 7^{th} grade, 24 forms; and at high school, 23 forms. When the students logged into the KCA system, they were randomly assigned a test form. As pilot forms were randomly assigned and administered via computer (KCA), it was possible for each student in a class to take a different pilot test and see a different set of items. Thus, pilot forms were randomly distributed to test takers ensuring that each test item was administered to a random group of students representative of the student population subgroups in Kansas. The number of students responding to an item ranged from a minimum of 70 students to a maximum of 505 students for a few items.

Pilot Item Analysis

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

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

discrimination information, for example). The rules or criteria for identifying poorly functioning items were the following.

Items were flagged for review if:

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

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

Test Form Development

Content area forms within a grade level were constructed to be parallel and have the same number of items per indicator and as a total. In Science, a sufficient number of items were available to build four operational forms at each of the three grade levels.

For the Science forms, items surviving the review of the pilot data were compiled at each grade level and grouped by measured student learning outcome classification (standard, benchmark, indicator, sub-indicator). Items were ordered on the basis of item difficulty from low to high (value) and placed on one of four forms. In some cases, more items existed in the pool for a given indicator than called for by the test specifications, so not all items were used during form construction. After all forms were initially constructed in this manner at a grade level, content and statistical reviews of each form were conducted. All items corresponding to an indicator across forms at a grade level were examined to ensure adequate content coverage. In places where there was overlap on a form or content gaps, items were deliberately moved across forms in an attempt to ensure content representation and reduce content overlap within a form. Statistical reviews were then executed, whereby average difficulty values were calculated at the test and benchmark level across forms. Items were moved across forms to ensure statistical similarity in terms of difficulty at the benchmark and overall form level with consideration given to content representation.

For the Spring 2008 administration, all operational test forms were administered on KCA using random assignment of test forms for purposes of equating test scores across forms (see Section 4). Due to delays in the development process, only one form was available at each grade to be printed in time for distribution into the field. Thus, only one form of any grade level test was made available to be administered in the traditional P&P modality.

Following the administration of the first operational forms of the Kansas Science Assessments in Spring 2008, analysis work commenced employing classical and IRT methods. Traditional item analysis studies were conducted on each test form to reconfirm the pilot test results that items selected for operational use were functioning adequately and as expected. As sufficient numbers of students in impacted subgroups do not exist in Kansas for examining differential item functioning (DIF) during the pilot testing phase of item development and selection, DIF analyses were performed on all items across all content area forms using spring administration test data. A Bias/Equity Review Committee was formed to review all items flagged as showing DIF (see Section 3 of this report). Test form equating was performed (see Section 4) following the DIF studies. Before AYP reporting could occur, standard setting activities needed to be implemented to establish score ranges on the tests that would define levels of test score performance needed for students to be classified into one of the five performance level categories established by the state (Exemplary, Exceeds Standard, Meets Standard, Approaches Standard, and Academic Warning). See Section 5 of this report for descriptions of the standard setting activities implemented. Based on the standards recommended by KSDE and approved by the Kansas State Board of Education, final results for the Kansas Science Assessments were reported in September 2008.

Elementary Science Tested Indicators		
Indicator #	**Text of Indicator**	**# of items**
S.4.1.1.1	asks questions that he/she can answer by investigating	2
S.4.1.1.2	plans and conducts a simple investigation	2
S.4.1.1.3	employs appropriate equipment, tools, and safety procedures to gather data	2
S.4.1.1.4	begins developing the abilities to communicate, critique, analyze his/her own investigations, and interprets the work of other students	2
S.4.2.1.1	observes properties of objects and measures those properties using appropriate tools	2
S.4.2.1.2	describes and *classifies* objects by more than one property	2
S.4.2.1.3	observes and records how one object *interacts* with another object	2
S.4.2.1.4	recognizes and describes the differences between solids, liquids, and gases	2
S.4.2.2.1	moves objects by pushing, pulling, throwing, spinning, dropping, and rolling; and describes the motion	2
S.4.2.3.1	identifies that the source of sound is vibrations	2
S.4.2.4.1	demonstrates that magnets attract and repel	2
S.4.2.4.3	constructs a *simple circuit*	2
S.4.3.1.1	observes different organisms and compares and contrasts how similar functions are served by different structural characteristics	2
S.4.3.1.2	compares basic needs of different organisms in their environment	2
S.4.3.2.1	compares, contrasts, and asks questions about life cycles of various organisms	2
S.4.4.1.1	collects, observes *properties*, and *classifies* a variety of *earth materials* in his/her *environment*	2
S.4.4.1.3	describes *properties* of water and process of the water cycle	2
S.4.4.2.3	discusses that the sun provides light and heat (electro-magnetic radiation) to maintain the temperature of the earth	2
S.4.4.3.1	describes changes in the surface of the earth	2
S.4.4.3.2	observes, describes, and records daily and seasonal weather changes	2
S.4.5.1.1	identifies a simple *design problem* (designs a plan, implements the plan, evaluates the results, makes changes to improve the product, and communicates the results)	2
S.4.6.1.1	discusses the nutritional value of various foods and their contribution to health	2

7th Grade Science Tested Indicators		
Indicator #	**Text of Indicator**	**# of items**
7.1.1.1	identifies questions that can be answered through scientific investigations.	2
7.1.1.2	designs and conducts *scientific investigations* safely using appropriate tools, mathematics, *technology*, and techniques to gather, analyze, and interpret data.	2
7.1.1.3	identifies the relationship between evidence and logical conclusions.	2
7.1.1.4	communicates scientific procedures, results and explanations.	2
7.1.3.2	evaluates the work of others to determine evidence which scientifically supports or contradicts the results, identifying faulty reasoning or conclusions that go beyond evidence and/or are not supported by data.	2
7.2.1.1	compares and classifies the states of matter; solids, liquids, gases, and plasma	2
7.2.2.1	understands the relationship of atoms to elements and elements to compounds.	2
7.2.2.2	measures and graphs the effects of temperature on matter.	2
7.2.3.2	describes, measures, and represents data on a graph showing the motion of an object (position, direction of motion, speed).	2
7.2.3.3	recognizes and describes examples of Newton's Laws of Motion.	2
7.2.3.4	investigates and explains how simple machines multiply force at the expense of distance.	2
7.2.4.2	understands that when work is done energy transforms from one form to another, including mechanical, heat, light, sound, electrical, chemical, and nuclear energy, yet is conserved.	2
7.2.4.3	observes and communicates how light (electromagnetic) energy interacts with matter: transmitted, reflected, refracted, and absorbed.	2
7.2.4.4	understands that heat energy can be transferred from hot to cold by radiation, convection, and conduction.	2
7.3.1.1	will understand the cell theory; that all organisms are composed of one or more cells, cells are the basic unit of life, and that cells come from other cells.	2
7.3.1.2	relates the structure of cells, organs, tissues, organ systems, and whole organisms to their functions	2

7.3.2.1	differentiates between asexual and sexual reproduction of organisms.	2
7.3.3.1	understands that internal and/or environmental conditions affect an organism's behavior and/or response in order to maintain and regulate stable internal conditions to survive in a continually changing environment.	2
7.3.4.1	recognizes that all populations living together (biotic resources) and the physical factors (abiotic resources) with which they interact compose an ecosystem.	2
7.3.4.3	traces the energy flow from the sun (source of radiant energy) to producers (via photosynthesis – chemical energy) to consumers and decomposers in food webs.	2
7.3.5.2	understands that adaptations of organisms (changes in structure, function, or behavior that accumulate over successive generations) contribute to biological diversity.	2
7.3.5.3	associates extinction of a species with environmental changes and insufficient adaptive characteristics.	2
7.4.1.1	identifies properties of the solid earth, the oceans and fresh water, and the atmosphere.	2
7.4.1.2	models earth's cycles, constructive and destructive processes, and weather systems.	2
7.4.2.1	understands that earth processes observed today (including movement of lithospheric plates and changes in atmospheric conditions) are similar to those that occurred in the past; earth history is also influenced by occasional catastrophes, such as the impact of a comet or asteroid.	2
7.4.3.1	compares and contrasts the characteristics of stars, planets, moons, comets, and asteroids.	2
7.4.4.1	demonstrates and models object/space/time relationships that explain phenomena such as the day, the month, the year, seasons, phases of the moon, eclipses and tides.	2
7.6.1.1	identifies individual nutrition, exercise, and a rest needs based on science and uses a scientific approach to thinking critically about personal health, lifestyle choices, risks and benefits.	2
7.6.2.1	investigates the effects of human activities on the environment and analyzes decisions based on the knowledge of benefits and risks.	2
7.7.2.1	recognizes that new knowledge leads to new questions and new discoveries, replicates historic experiments to understand principles of science, and relates contributions of men and women to the fields of science.	2

Indicator #	Text of Indicator	Physical or Life Science Test Component	# of items
S.HS.1.1.2	actively engages in investigations, including developing questions, gathering and analyzing data, and designing and conducting research	P	2
S.HS.1.1.3	actively engages in using technological tools and mathematics in their own scientific investigations.	P	2
S.HS.2A.1.1	understands atoms, the fundamental organizational unit of matter, are composed of subatomic particles. Chemists are primarily interested in the protons, electrons, and neutrons found in the atom.	P	2
S.HS.2A.2.1	understands chemists use kinetic and potential energy to explain the physical and chemical properties of matter on earth that may exist in any of these three states: solids, liquids, and gases.	P	2
S.HS.2A.2.2	understands the periodic table lists elements according to increasing atomic number. This table organizes physical and chemical trends by groups, periods, and sub-categories.	P	2
S.HS.2A.2.3	understands chemical bonds result when valence electrons are transferred or shared between atoms. Breaking a chemical bond requires energy. Formation of a chemical bond releases energy. Ionic compounds result from atoms transferring electrons. Molecular compounds result from atoms sharing electrons. For example, carbon atoms can bond to each other in chains, rings, and branching networks. Branched network and metallic solids also result from bonding.	P	2
S.HS.2A.3.1	understands a chemical reaction occurs when one or more substances (reactants) react to form a different chemical substance(s) (products). There are different types of chemical reactions all of which demonstrate the Law of Conservation of Matter and Energy.	P	2
S.HS.2B.1.1	understands Newton's Laws and the variables of time, position, velocity, and acceleration can be used to describe the position and motion of particles.	P	2
S.HS.2B.2.2	understands the first law of thermodynamics states the total internal energy of a substance (the sum of all the kinetic and potential energies of its constituent molecules) will change only if heat is exchanged with the environment or work is done on or by the substance. In any physical interaction, the total energy in the universe is conserved.	P	2
S.HS.2B.3.2	understands waves have energy and can transfer energy when they interact with matter.	P	2
S.HS.2B.3.5	understands electromagnetic waves resulty when a charged particle is accelerated or decelerated.	P	2
S.HS.3.1.2	understands cell functions involve specific chemical reactions.	L	2

S.HS.3.2.1	understands living organisms contain DNA or RNA as their genetic material, which provides the instructions that specify the characteristics of organisms.	L	2
S.HS.3.2.3	understands hereditary information is contained in genes, located in the chromosomes of each cell.	L	2
S.HS.3.3.1	understands biological evolution, descent with modification, is a scientific explanation for the history of the diversification of organisms from common ancestors.	L	2
S.HS.3.3.3	understands biological evolution is used to explain the earth's present day biodiversity: the number, variety and variability of organisms.	L	2
S.HS.3.3.4	understands organisms vary widely within and between populations. Variation allows for natural selection to occur.	L	2
S.HS.3.4.1	understands atoms and molecules on the earth cycle among the living and nonliving components of the biosphere.	L	2
S.HS.3.4.3	understands the distribution and abundance of organisms and populations in ecosystems are limited by the carrying capacity.	L	2
S.HS.3.5.2	understands the sun is the primary source of energy for life through the process of photosynthesis.	L	2
S.HS.3.5.3	understands food molecules contain biochemical energy, which is then available for cellular respiration.	L	2
S.HS.3.6.1	understands animals have behavioral responses to internal changes and to external stimuli.	L	2
S.HS.3.7.2	understands that homeostasis is the dynamic regulation and balance of an organisms internal environment to maintain conditions suitable for survival.	L	2
S.HS.3.7.3	understands that living things change following a specific pattern of developmental stages called life cycles.	L	2
S.HS.4.1.2	understands the theory of Plate Tectonics explains that internal energy drives the earth's ever changing structure.	P	2
S.HS.4.2.1	understands geological time is used to understand the earth's past.	L	2
S.HS.4.3.2	understands the relationship between the earth, moon, and sun explains the seasons, tides and moon phases.	P	2
S.HS.4.4.1	understands stellar evolution.	P	2
S.HS.5.1.1	understands technology is the application of scientific knowledge for functional purposes.	P	2
S.HS.6.3.1	understands natural resources from the lithosphere and ecosystems are required to sustain human populations.	L	2

Summary Statistics Spring 2008 Administration

The summary statistics displayed in the Spring 2008 Science Administration tables were created using scores from students with an AYP status code equal to 1 and based on end of year data. The Differential Item Functioning Analyses, Test Equating, Standard Setting, Reliability Analyses, Validity Information and the Comparability Study used all scores regardless of AYP status and were based on the end of testing data. The available data at the close of a testing window for KCA usually does not differ from the data that is available at the end of year. However, the scanned paper and pencil data typically requires a considerable amount of effort to clean up and verify the results which could account for some of the differences found between the end of year and end of testing data. In any case, the main difference in the results will be due to the AYP status of the test takers. Below is the list of AYP Status Codes for 2008.

2008 AYP Status Codes

-4	Incomplete test (fewer than one third of the items tried) so student did not participate in testing.
-3	Test does not represent a reasonable measure of the student's ability (a KAMM or Alternate assessment was given to a student with no SPED code, the student was tested at a grade level different from their KIDS grade, the student did not try, cheating happened) so student did not participate in testing.
-2	No test was received (includes students whose scanned tests had bad identifying information) and no Special Circumstance code was provided so student did not participate in testing.
-1	No test was received and a Special Circumstance code was provided, but the circumstance did not exempt the student from testing so student did not participate in testing.
0	The student was exempted from testing by a Special Circumstance code so student was not included in AYP reporting.
1	The student was included in participation and % proficient calculations.
2	The student was represented by an OTL test from a previous year and was included in participation and % proficient calculations.
3	The student counted only in participation calculations as result of a Special Circumstance code.
4	The student counted only in participation calculations because they arrived after September 20th or were recently arrived in USA.
5	The student was exempt from testing because they arrived later than the date chosen by KSDE after which testing is not required so was not included in AYP reporting.
6	The student was included in the AYP calculation of a building that is not currently the AYP building, having scored proficient at a previous AYP building.
7	The student has ambiguous AYP building information in KIDS and will be considered not tested if no correction is provided.

2008 Grade 4 Science Spring Testing Results

	Forms				
	435 (P&P)	**435 (KCA)**	**271**	**618**	**967**
Number of items	44	44	44	44	44
Sample Size (N)	6,749	7,670	6,419	6,405	6,393
Mean Raw Score	31.065	32.529	32.697	33.540	33.137
SD Raw Score	6.866	6.602	6.412	6.216	6.540
Mean Scaled Score	70.603	73.933	75.641	75.573	75.654
SD Scaled Score	15.606	15.004	14.476	14.411	14.567
Reliability (alpha)	0.847	0.847	0.841	0.828	0.847
SEM Raw Score	2.683	2.583	2.557	2.577	2.555
SEM Scaled Score	6.098	5.870	5.773	5.974	5.692
Non-Mastery	12.21%	8.42%	6.28%	6.59%	6.12%
Mastery	87.79%	91.58%	93.72%	93.41%	93.88%
Academic Warning	0.90%	0.74%	0.59%	0.59%	0.75%
Approaches Standard	11.31%	7.68%	5.69%	6.00%	5.37%
Meets Standard	40.58%	34.97%	30.77%	30.20%	33.87%
Exceeds Standard	33.53%	37.50%	43.67%	46.82%	37.17%
Exemplary	13.68%	19.11%	19.29%	16.39%	22.85%

2008 Grade 7 Science Spring Testing Results

	Forms				
	726 (P&P)	**726 (KCA)**	**236**	**589**	**892**
Number of items	60	60	60	60	60
Sample Size (N)	7,033	7,311	6,300	6,329	6,318
Mean Raw Score	36.064	37.917	39.746	39.114	38.772
SD Raw Score	10.241	9.789	9.254	9.461	9.564
Mean Scaled Score	60.106	63.191	65.144	65.158	65.204
SD Scaled Score	17.071	16.316	15.749	15.828	15.862
Reliability (alpha)	0.891	0.883	0.876	0.877	0.879
SEM Raw Score	3.385	3.343	3.261	3.314	3.326
SEM Scaled Score	5.643	5.572	5.549	5.544	5.517
Non-Mastery	22.18%	16.81%	12.84%	13.21%	11.78%
Mastery	77.82%	83.19%	87.16%	86.79%	88.22%
Academic Warning	4.41%	2.76%	1.81%	2.01%	2.26%
Approaches Standard	17.77%	14.05%	11.03%	11.20%	9.51%
Meets Standard	38.67%	36.23%	36.02%	38.06%	38.49%
Exceeds Standard	26.26%	31.87%	36.54%	31.55%	33.32%
Exemplary	12.88%	15.09%	14.60%	17.17%	16.41%

2008 High School Physical Science Spring Testing Results

	Forms				
	916 (P&P)	**916 (KCA)**	**449**	**657**	**714**
Number of items	30	30	30	30	30
Sample Size (N)	6,081	9,247	8,672	8,721	8,716
Grade 9	334	1110	1024	1039	1060
Grade 10	475	1474	1378	1384	1379
Grade 11	5269	6663	6269	6294	6276
Grade 12	3	0	1	4	1
Mean Raw Score	16.703	16.264	15.914	16.660	16.043
SD Raw Score	5.398	5.257	4.461	5.266	4.694
Mean Scaled Score	55.679	54.212	54.909	54.939	55.297
SD Scaled Score	18.002	17.525	17.091	16.984	17.133
Reliability (alpha)	0.796	0.783	0.696	0.785	0.717
SEM Raw Score	2.439	2.451	2.461	2.442	2.499
SEM Scaled Score	8.135	8.172	9.429	7.876	9.122

2008 High School Life Science Spring Testing Results

	Forms				
	118 (P&P)	**118 (KCA)**	**249**	**369**	**875**
Number of items	30	30	30	30	30
Sample Size (N)	7,688	10,724	10,201	10,227	10,226
Grade 9	1,048	788	795	795	802
Grade 10	1,352	3,278	3,124	3,132	3,125
Grade 11	5,287	6,656	6,280	6,299	6,296
Grade 12	1	2	2	1	3
Mean Raw Score	17.536	17.688	17.148	17.225	17.560
SD Raw Score	5.409	5.213	4.714	5.207	4.890
Mean Scaled Score	58.457	58.961	59.253	59.605	59.502
SD Scaled Score	18.035	17.379	17.055	17.334	17.035
Reliability (alpha)	0.792	0.775	0.726	0.781	0.748
SEM Raw Score	2.465	2.472	2.467	2.438	2.456
SEM Scaled Score	8.218	8.240	8.926	8.114	8.556

Physical and Life Science all Form Combinations Spring Testing Results

All Form Combinations

Number of items	60
Sample Size (N)	36170
Mean Scaled Score	56.996
SD Scaled Score	15.994
Reliability (Stratified alpha)	.824 to .879
Non-Mastery	14.89%
Mastery	85.11%
Academic Warning	1.77%
Approaches Standard	13.11%
Meets Standard	53.11%
Exceeds Standard	24.54%
Exemplary	7.46%

Section 3

DIFFERENTIAL ITEM FUNCTIONING (DIF) ANALYSES

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

Several implementation issues, essential for appropriate DIF analysis, were considered for the Kansas Science Assessments given for the first time in Spring 2008. As with any statistical procedure, sample sizes of the comparison groups have a direct impact on the power of the DIF procedure. With very small samples of reference or focal groups, results from the DIF analysis might not be trustworthy. Based on the sample size recommendations in the literature, sample sizes for both reference and focal groups were examined before the DIF analysis.

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

Procedures

Samples

Taking the above into account, the DIF analysis procedures and criteria put in place emphasized sufficient sample sizes and curriculum matching as a basis for making decisions and recommendations. In 2008, analyses were conducted for each science test form using gender and racial/ethnic groups. To control for the effects of different curriculum/instructional approaches, samples of Caucasian test takers were drawn from schools that had minority groups. Separate samples of Caucasians were drawn for each minority group.

In 2008, there were four science test forms per grade administered for equating purposes, which led to smaller sample sizes for minority groups taking any one test form than would have occurred if only one form had been administered. The sample size issue becomes particularly relevant for Asian Americans and Native Americans. Such sample sizes are consistently less than 200, a number suggested by the literature as the minimal sufficient sample size for conducting DIF studies. Therefore, for racial/ethnic group comparison, DIF studies were conducted only on African Americans and Hispanic Americans as focal groups, using same district/building sampled Caucasian students as the reference group.

Items

The science items from general assessment test forms at all grade levels were analyzed for DIF. There were four test forms per grade with an equal number of items across test forms. For each of the four grade level tests in science (Grades 4, 7, 11 Life, and 11 Physical), the number of items on each test form at a given grade was 44, 60, 30, and 30, respectively. Thus, the total number of science items involved in the DIF analyses ranged from 176 at grade 4 to 240 at grades 7 and 11. All items were in the multiple choice format and thus were scored dichotomously.

Statistical Methods

The procedure used was the Mantel-Haenszel (MH) technique. The criteria used in these analyses were: (1) the absolute delta value larger than 1.5 and (2) the absolute delta value statistically significantly larger than 1.0. Using a statistical significance level of 0.01, the second criterion is equivalent to a MH chi-squared value of 12.7866. Items with negative delta values created a disadvantage for the focal group while positive values created an advantage for the focal group in comparison to the reference group.

Results

A sample output for a science DIF analysis at Grade 4, Form 271 is given in Table 3.1. As shown in Table 3.1., Item 35 appeared to show DIF for the Hispanic group versus the Caucasian comparison group. This item has an absolute Delta value of 1.76 and MH chi-squared value of 24.82. This item seemed to advantage Caucasians.

Tables 3.2. through 3.5 give summaries of items flagged by the Mantel-Haenszel procedure for each DIF comparison by form at each of the four grade levels, respectively. In each of the tables, information about the flagged DIF items for each specific comparison performed on each form at a given grade is grouped into four parts. The test form number is given in the first column of each table (under the title *Form*). In the second part (under the title *DIF Group*), both the reference and the focal group in each of the three comparisons performed on a test form, as well as their corresponding sample sizes, are given. It should be noted that different samples of Caucasians were drawn for Hispanic/Caucasian and Black/Caucasian comparisons, respectively. In the third part (under the title *DIF Items*), ID numbers for items that are showing DIF are presented in the table. Specifically, the ID number for each item is a unique number in the CETE test system that makes it possible to track all changes and decisions made for the item. For each item ID number, a “A” or “D” indicates the direction of the DIF that the item shows. As mentioned earlier, items with “D” were seen to disadvantage the focal group while items with “A” advantaged the focal group in comparison to the reference group. The last part of each table gives the total counts of the number of flagged items for each test form (*Total*). Items that advantage or disadvantage the focal group were tallied separately.

As an example, Table 3.2. presents the results of DIF analyses for the four science test forms at grade 4 (i.e., Forms 271, 435, 618, and 967). For each form, DIF analyses were conducted for each of the three comparisons (i.e., Male vs. Female, Caucasian vs. Hispanic, and

Caucasian vs. Black). Table 3.2 shows that four items were flagged at grade 4 from a total of 528 comparisons (176 items and three group comparisons), all of them with negative DIF. All of the items were flagged for ethnic comparisons. For the science DIF analyses across all grade levels, 1,968 comparisons were made, and a total of five items were flagged showing positive DIF and 22 items were flagged showing negative DIF. The number of flagged items represent slightly over one percent (1.4%) of the total number of statistical comparisons made.

Judgmental Review of DIF Items

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

Equity Review Committee

An equity review committee was convened by the Kansas State Department of Education to review potentially biased items on the Kansas Assessments. The committee, composed of representatives from affected minority or gender groups, was formed to judgmentally review test items for sensitivity and fairness that were flagged as showing differential item functioning (DIF) during the statistical DIF analyses. This review was conducted by KSDE during the week of June 2, 2008 in Topeka, Kansas, and focused on the review of items that evidenced DIF in the statistical analysis of the items for students belonging to the respective ethnic or gender groupings. Each committee member was provided sets of flagged items from the science content area tests; committee members reviewed only items evidencing DIF for students in their same ethnic or gender grouping.

An overview of the bias review process was presented by KSDE staff to start the proceedings. After the training session, committee members began the judgment procedure. Panelists were directed to review each item independently in terms of fairness, focusing specifically on content, language, offensiveness, or stereotypes that may have been present in the respective items. After the independent item review was completed by committee members, panelists discussed each item under review. The review criteria presented to the committee during the training session required committee members representing a group to reach consensus regarding each item. For items that the review committees detected bias, a description or explanation of the source of the bias was required. KSDE was supplied with the item feedback from committee members and made the final decision regarding an item's deletion or retention. Based on the committees' review and feedback, KSDE decided to retain all of the items on the science assessments.

Table 3.1
Example DIF Output for Grade 4 Science Form 271

MANTEL-HAENSZEL DIP ANALYSIS FOR DIF, G4 Science, form 271,
Caucasian(4,reference) vs. Hispanic (3, focal)

NUMBER OF ITEMS = 44 & CHK = .000

ITEM	P-4	PB-4	P-3	PB-3	P-4+3	PB-4+3	CHI-I	APLHA-I	DELTA-I	CHI-E	APLHA-E	DELTA-E
1	.88	.38	.74	.49	.81	.47	2.12	1.27	-.56	3.53	1.35	-.70
2	.76	.38	.68	.47	.73	.43	5.38	.72	.77	2.66	.80	.53
3	.61	.48	.40	.34	.51	.45	4.19	1.29	-.59	8.65	1.42	-.82
4	.74	.36	.68	.32	.71	.34	4.31	.76	.65	1.50	.85	.38
5	.96	.19	.93	.31	.95	.26	.78	.75	.66	.02	.93	.18
6	.81	.42	.71	.42	.76	.43	.33	.91	.21	.00	1.00	-.00
7	.77	.39	.67	.40	.73	.41	.86	.88	.31	.09	.95	.11
8	.45	.34	.35	.32	.40	.35	.20	.94	.14	.07	1.04	-.09
9	.66	.38	.48	.40	.58	.43	2.11	1.20	-.42	6.05	1.34	-.69
10	.81	.42	.74	.40	.77	.41	4.21	.74	.71	2.14	.81	.50
11	.93	.29	.88	.45	.91	.39	.69	.82	.46	.19	.90	.26
12	.75	.45	.64	.45	.70	.46	1.66	.83	.43	.47	.91	.23
13	.84	.29	.74	.38	.80	.36	.30	1.09	-.21	1.07	1.17	-.38
14	.20	.06	.21	.07	.20	.06	1.78	.82	.46	.28	.92	.19
15	.71	.53	.49	.52	.61	.56	2.46	1.24	-.51	5.85	1.37	-.74
16	.56	.32	.43	.31	.50	.34	.30	1.07	-.17	1.93	1.18	-.39
17	.91	.41	.87	.36	.89	.38	3.44	.69	.88	3.16	.70	.83
18	.76	.35	.62	.34	.69	.37	1.52	1.18	-.39	3.36	1.26	-.55
19	.86	.44	.75	.46	.81	.46	.11	.94	.15	.02	1.04	-.08
20	.79	.41	.68	.42	.73	.43	.32	.92	.20	.01	1.02	-.06
21	.81	.30	.76	.39	.79	.35	3.80	.74	.70	1.97	.81	.50
22	.63	.54	.41	.48	.53	.54	1.94	1.21	-.44	5.15	1.33	-.68
23	.80	.35	.67	.44	.74	.42	.22	1.08	-.17	1.75	1.21	-.44
24	.78	.26	.72	.30	.76	.29	1.18	.85	.37	.13	.95	.13
25	.66	.37	.59	.44	.63	.41	9.40	.68	.91	3.54	.79	.55
26	.78	.27	.79	.09	.79	.17	8.10	.68	.92	3.68	.77	.61
27	.94	.37	.84	.42	.89	.42	.93	1.24	-.50	3.13	1.43	-.83
28	.90	.20	.81	.39	.85	.33	1.46	1.24	-.51	2.20	1.29	-.61
29	.57	.41	.44	.26	.51	.36	.04	.97	.07	.73	1.11	-.25
30	.67	.28	.58	.28	.62	.29	.13	.95	.11	.37	1.08	-.18
31	.95	.28	.90	.37	.93	.35	.00	1.02	-.04	.01	1.04	-.10
32	.76	.32	.64	.33	.70	.35	.37	1.09	-.20	2.11	1.21	-.44
33	.90	.43	.78	.44	.84	.45	.84	1.19	-.40	2.10	1.29	-.60
34	.59	.38	.45	.37	.52	.40	.10	1.05	-.11	.89	1.12	-.27
35	.88	.34	.68	.38	.79	.41	24.82	2.11	-1.76	30.20	2.23	-1.88
36	.79	.44	.67	.39	.73	.43	.03	1.03	-.07	.59	1.12	-.26
37	.61	.28	.51	.27	.56	.29	.03	1.03	-.06	1.11	1.13	-.29
38	.49	.39	.35	.22	.42	.33	.97	1.13	-.29	3.44	1.25	-.52
39	.64	.49	.44	.38	.54	.47	3.73	1.28	-.57	8.00	1.40	-.80
40	.96	.30	.90	.31	.93	.32	.59	1.24	-.50	.71	1.25	-.53
41	.90	.43	.82	.34	.86	.40	.03	1.04	-.10	.10	1.07	-.15
42	.58	.27	.49	.17	.54	.24	.07	1.04	-.09	1.42	1.15	-.32
43	.81	.25	.73	.22	.77	.25	.58	1.12	-.26	2.26	1.23	-.49
44	.60	.34	.49	.33	.55	.35	.21	.94	.14	.32	1.07	-.17

** SUMMARY DATA **

	MEAN	SD	CASES	KR-20
GROUP-4	32.778	6.301	792.	0.835
GROUP-3	28.137	6.977	721.	0.838
TOTAL	30.566	7.025	1513.	0.854

Table 3.2
Summary of Differentially Functioning Items for Grade 4 Science Forms

Form	DIF Group				DIF Items	Total
	Reference	N	Focal	N		
271	Male	3331	Female	3306		
	Caucasian	792	Hispanic	721	32446 D	
		479	Black	403		
						0 A, 1 D
435	Male	3348	Female	3321		
	Caucasian	861	Hispanic	767	32421 D	
		506	Black	403		
						0 A, 1 D
618	Male	3390	Female	3242		
	Caucasian	744	Hispanic	744	32248 D	
		555	Black	418		
						0 A, 1 D
967	Male	3315	Female	3320		
	Caucasian	790	Hispanic	729	32273 D	
		559	Black	433		
						0 A, 1 D
Total						A=0, D=4

D = disadvantaged the focal group
A = advantaged the focal group

Table 3.3
Summary of Differentially Functioning Items for Grade 7 Science Forms

Form	DIF Group				DIF Items			Total
	Reference	N	Focal	N				
236	Male	3256	Female	3284	32586 A	32682 D	32749 A	
	Caucasian	768	Hispanic	686				
		542	Black	369				
								2 A, 1 D
589	Male	3236	Female	3272	32703 D			
	Caucasian	726	Hispanic	682				
		431	Black	345	37412 D	32808 D		
								0 A, 3 D
726	Male	3335	Female	3217	32952 A	32878 D		
	Caucasian	740	Hispanic	658	32846 D			
		515	Black	370				
								1 A, 2 D
892	Male	3285	Female	3227	32949 A	32968 D	32707 D	
	Caucasian	774	Hispanic	681				
		411	Black	309				
								1 A, 2 D
Total								A=4, D=8

D = disadvantaged the focal group
A = advantaged the focal group

Table 3.4
Summary of Differentially Functioning Items for Grade 11 Life Science Forms

Form	DIF Group				DIF Items	Total
	Reference	N	Focal	N		
118	Male	3278	Female	3223	33426 D	
	Caucasian	693	Hispanic	569		
		583	Black	396		
						0 A, 1 D
249	Male	3267	Female	3213	33428 D	
	Caucasian	753	Hispanic	552		
		567	Black	410		
						0 A, 1 D
369	Male	3209	Female	3269		
	Caucasian	827	Hispanic	595		
		504	Black	370		
						0 A, 0 D
875	Male	3363	Female	3121	33429 D	
	Caucasian	726	Hispanic	543	33196 D	
		526	Black	402	33429 D	
						0 A, 3 D
Total						A=0, D=5

D = disadvantaged the focal group
A = advantaged the focal group

Table 3.5
Summary of Differentially Functioning Items for Grade 11 Physical Science Forms

Form	DIF Group				DIF Items		Total
	Reference	N	Focal	N			
449	Male	3270	Female	3185			
	Caucasian	706	Hispanic	525			
		489	Black	381	33390 D	33323 D	
							0 A, 2 D
657	Male	3232	Female	3233			
	Caucasian	778	Hispanic	576			
		537	Black	399			
							0 A, 0 D
714	Male	3337	Female	3125	33087 A		
	Caucasian	754	Hispanic	574	33394 D		
		539	Black	379	33399 D		
							1 A, 2 D
916	Male	3243	Female	3251	33348 D		
	Caucasian	774	Hispanic	574			
		571	Black	411	33348 D		
							0 A, 2 D
Total							A=1, D=6

D = disadvantaged the focal group
A = advantaged the focal group

Section 4

TEST EQUATING

When multiple forms of a test are built and administered, test equating is an essential component to the scoring process. Test equating ensures that all examinees receive a score on the same scale, regardless of the test form the examinee was administered. Three important properties of test equating are equity, symmetry, and identical test specifications (Kolen and Brennan, 2004, pp.10-13). Simply stated, equity requires that, if multiple forms of a test exist for the same ability level, it should make no difference to examinees which form they are administered. Symmetry requires that examinee scores would be consistent (relative to other examinees) regardless of which form is chosen as the base and which forms are equated to it. Identical test specifications require that every form is built with the same content constraints and statistical indicators in mind. Without these three properties or assumptions, a test form cannot be said to be satisfactorily equated even if sophisticated methods were applied.

For the Kansas Science Assessments administered in Spring 2008, scores from parallel test forms administered to different groups needed to be equated to ensure the equitability of scores for every examinee. As detailed in the Technical Manual under the section "Forms Development," test forms were pre-equated based on pilot data to ensure that those test forms were constructed to be classically parallel, an important prerequisite for equating scores across multiple test forms. This section summarizes the description of the equating design and methods, as well as addresses issues of equating multiple forms of the Kansas Assessments.

Procedures

Equating Design and Data Configuration

The Spring 2008 administration of the Kansas Science Assessments allowed schools to select the mode of administration (paper-and-pencil or computer) for individual students. Thus, a school could voluntarily elect to test none, part, or all of its students on the computer.

Four parallel forms, using items configured from pilot test data, were available in science at each grade level. All test forms were made available on the computer (KCA) and were randomly assigned to students when students were registered for KCA. Only one paper-and-pencil (P&P) form for each grade level was available. All other forms for a grade level content area were available only on the computer (KCA). Thus, the basic configuration for test administration is as follows:

✓ P& P			
✓ Self-selected			
Form A	**Form B**	**Form C**	**Form D**
✓ KCA	✓ KCA	✓ KCA	✓ KCA
✓ Random G1	✓ Random G2	✓ Random G3	✓ Random G4

The above configuration provided a randomized, equivalent groups design that could be used to equate test form scores using only the KCA tested students. A potential problem is the volunteer-nature of the KCA group and the possibility that it may not sufficiently reflect the complete distribution of ability and performance of all students in the state. Table 4.1 shows that the raw score distribution means and standard deviations are not very different for KCA and P&P.

Table 4.1
Grade Level Means and Standard Deviations for Matched Groups of Students Taking Tests in the P&P or KCA Mode

Grade Test Form	Mode of Testing	Sample Size	Mean	Standard Deviation	Percent Proficient
Grade 4	P&P	5536	72.33	15.19	90.0
	KCA	5536	74.05	15.00	91.5
Grade 7	P&P	5714	62.35	16.77	81.8
	KCA	5714	63.65	16.33	83.5
HS Life Science	P&P	5026	58.33	17.56	85.2
	KCA	5026	57.42	17.12	84.8
HS Physical Science	P&P	4967	56.06	17.66	82.1
	KCA	4967	54.70	17.07	80.8

Also, a large majority of students took the Kansas Science Assessments using KCA and thus were included in the equating data. From the approximately 30,000-35,000 regular education students taking the test at each grade level, approximately 20 percent took the test using the P&P mode and received the single P&P form available at a grade level. For the remaining 80 percent of the students who were administered the test using KCA, each of the science test forms was administered to approximately 6,500 students per grade level. As the percent of students in the state taking KCA increases, the sample size of the P&P available for equating decreases. At some point, the standard error of equating (especially at the tails of the distribution) increases. Thus Kansas has made the policy decision not to separately equate the KCA and P&P versions.

Table 4.2 below shows the percentages of students across schools in Kansas who were administered the different KCA forms of the Kansas Science Assessments. For the values in the table, percentages of students taking each form were obtained for each school, and these percentages were summarized across schools. In addition, the table provides percentages of students taking each form by gender, race, and educational classifications. The numbers for the base form (the form given in both P&P and KCA mode) at each grade level are shown in bold. Across grade levels, the demographic percentages support the equivalence of groups using this data collection design. In other words, data in Table 4.2 suggest the equivalence of the KCA groups responding to each form, at all grades.

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

Grade	N of Schools	Form	Gender			Race		Education	
			Total	Female	Male	White	Minority	Regular	Sped
4	666	271	25.0	49.8	50.2	75.2	24.8	90.7	9.3
	668	**435**	**25.1**	**50.2**	**49.8**	**74.8**	**25.2**	**90.8**	**9.2**
	664	618	25.0	48.9	51.1	75.4	24.6	90.7	9.3
	671	967	25.0	50.0	50.0	74.9	25.1	90.6	9.4
7	407	236	25.0	50.2	49.8	77.9	22.1	93.4	6.6
	400	589	24.9	50.3	49.7	77.4	22.6	93.5	6.5
	400	**726**	**25.1**	**49.1**	**50.9**	**78.0**	**22.0**	**92.6**	**7.4**
	400	892	24.9	49.6	50.4	78.5	21.5	93.3	6.7
11 Life	**333**	**118**	**25.1**	**49.6**	**50.4**	**79.3**	**20.7**	**93.6**	**6.4**
	330	249	25.0	49.6	50.4	79.9	20.1	93.1	6.9
	331	369	25.0	50.5	49.5	79.7	20.3	93.1	6.9
	330	875	25.0	48.1	51.9	79.8	20.2	93.6	6.4
11 Physical	334	449	24.9	49.3	50.7	80.4	19.6	93.2	6.8
	331	657	25.0	50.0	50.0	79.2	20.8	93.1	6.9
	327	714	25.0	48.4	51.6	79.5	20.5	93.7	6.3
	332	**916**	**25.1**	**50.1**	**49.9**	**79.8**	**20.2**	**93.9**	**6.1**

Statistical Procedures

Using random student score samples from the KCA test forms, results for classical equipercentile test equating procedures were examined. Alternate equating procedures, including classical linear equating and item response theory methods, were considered for previous administrations. These previous examinations determined that equipercentile methods were most appropriate for these data. The equipercentile equating methods were used on the observed score frequency distributions. In science, the total score levels are expressed in the percent correct metric. The test form given in both the KCA and the P&P mode (Form A) served as the base form in all equating analyses. Scores from all other forms were transformed onto the base form score percent correct scale. Criteria for selecting the best equipercentile method for equating two specific sets of scores are presented below.

A major issue in 2008 involved equating scores between the P&P test form and the corresponding KCA form (Form A). As the assignment of test taking mode for a student was not random but rather a local decision made by districts or schools, the possibility exists that the assignment of students to KCA or P&P was related to or determined by characteristics of the students. Consequently, the two populations (students who take P&P Form A and students who take the KCA test forms) might be different in terms of proficiency for a given subject at a given

grade. Thus, the effects of test mode and population ability differences are intertwined. In a scenario with small mode effects where any difference in P&P and KCA student scores reflects primarily population ability differences, one need not equate. Rather, there would be an assumption of score value equivalency for the same two scores in both populations. This situation has been evidenced to some extent by data from prior studies in Kansas on mode effects for reading and mathematics and from other studies and reviews found in the testing literature. Based on prior studies in reading and mathematics where the mode effect size has been judged small, Kansas has decided to treat the P&P test form as equivalent to the same KCA administered test form; thus no adjustment is made in the scores for either set of data.

While Kansas continues to study the comparability issue and the need for conversion tables that might adjust scores obtained under different modes of testing, the most recent studies examining test mode differences on the science test forms support the decision to treat P&P and KCA test forms the same. To partially control for ability differences in the two populations, an equal number of test scores from each racial and free/reduced lunch group was used for both test taking modes. The means, standard deviations, and percent classified as Proficient (above the state's "Meets Standard" cut-score) for the two matched groups are reported in Table 4.3 and illustrate the distributional similarities.

Table 4.3
Grade Level Means and Standard Deviations for Matched Groups of Students Taking Tests in P&P or KCA Mode

Grade Test Form	Mode of Testing	Sample Size	Mean	Standard Deviation	Percent Proficient
Grade 4	P&P	5536	72.33	15.19	90.0
	KCA	5536	74.05	15.00	91.5
Grade 7	P&P	5714	62.35	16.77	81.8
	KCA	5714	63.65	16.33	83.5
HS Life Science	P&P	5026	58.33	17.56	85.2
	KCA	5026	57.42	17.12	84.8
HS Physical Science	P&P	4967	56.06	17.66	82.1
	KCA	4967	54.70	17.07	80.8

Equating Criteria

Because several classical equipercentile test equating methods were implemented, comparisons among competing methods were necessary. Thus, equating methods were evaluated and decisions made as to which method produced the most reasonable conversion of scores for students taking different forms of the test.

To assist in selecting the best equating conversion, the following criteria were used:

1. *Fidelity to the equated data*

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

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

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

3. *Parsimony*

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

4. *Smoothed distributional properties*

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

These criteria were used simultaneously, with the favored, and subsequently adopted, methods meeting all or most criteria. Each of the following steps refers to the comparisons and equating conducted within each grade (in each of which, three forms were equated to a base form). The descriptions of these procedures address Standards 4.10, 4.11, and 4.12 of the AERA, APA, NCME (1999) Standards for Educational and Psychological Testing, which are related to issues of score equivalence, equating methods, and equivalence of groups, respectively. It should be noted that these procedures were adopted based on recommendations found in Kolen and Brennan's (2004) book on test score equating.

1. *Comparison of raw score distributions*

   The four histograms of raw score distributions were visually compared for comparability. The first four moments of the raw score distributions were likewise compared. This provided evidence of pre-existing form comparability (pre-equated parallel forms) as well as evidence to support the randomness of form administration.

2. *Check for evidence of random administration*

   Randomness of form administration was further considered by comparing demographic profiles of examinees taking each of the four forms. Forms were very balanced with regard to gender, race, and education status (regular vs. SPED). Previously referenced Table 4.2 contains this evidence.

3. *Conduct equipercentile equating*

   Equipercentile equating with cubic spline post-smoothing was conducted using the RAGE-RGEQUATE program by Kolen, et al. (discussed and used in the Kolen and Brennan, 2004). This program provides output for different spline sizes ("S"). Values of S for each equating were chosen based on procedures demonstrated in Kolen & Brennan, 2004. Generally, potential values of S range between 0 (no smoothing) and 1 (much smoothing). Larger values of S therefore result in smoother equating functions but also tend to change the raw score distributions more.

   In choosing S values for each equating, the following decision rules were followed. The largest S value possible that created a smooth equating function was chosen, as long as

   - the first four moments of the base form distribution were "preserved" (i.e., the equated scores' moments are very similar to or the same as 2 to 4 decimal places).
   - across the raw score distribution, the differences between the equated and raw scores are no larger than +/- 1 standard error (SE) of the unsmoothed equipercentile equating solution. Note that all S values led to some conversions outside +/- 1 SE, for very infrequent or non-occurring score points (e.g., total scores below chance-level success), but these are negligible because no examinees were actually affected.

Results

Table 4.4 shows a descriptive summary of the equating samples obtained for science. The numbers for the base form at each grade level are shown in bold. All forms at a grade level were constructed to the same content and statistical specifications. Table 4.4 below presents total scores on forms in terms of average number correct and average percent correct. Also included is reliability information for each of the test forms. Table 4.4 shows that all the forms across grade levels had sufficient reliability for equating purposes.

Table 4.4
Descriptive Statistics for Equating Samples for Science by Test Form

Grade	Form	N Items	N	Reliability (α)	Mean Raw Score	SD Raw Score	Mean Percent Correct	SD Percent Correct
4	271	44	6637	0.84	32.61	6.49	74.12	14.74
	435	**44**	**6669**	**0.85**	**33.18**	**6.45**	**75.40**	**14.66**
	618	44	6632	0.83	33.44	6.31	76.01	14.33
	967	44	6635	0.85	33.05	6.57	75.12	14.93
7	236	60	6540	0.88	39.58	9.33	65.96	15.55
	589	60	6508	0.88	38.95	9.51	64.92	15.86
	726	**60**	**6552**	**0.88**	**38.93**	**9.53**	**64.88**	**15.89**
	892	60	6512	0.88	38.61	9.58	64.35	15.96
11 Life	**118**	**30**	**6501**	**0.77**	**17.39**	**5.13**	**57.96**	**17.09**
	249	30	6480	0.73	16.80	4.71	56.01	15.71
	369	30	6478	0.78	16.71	5.15	55.70	17.16
	875	30	6484	0.75	17.11	4.90	57.04	16.34
11 Physical	449	30	6455	0.71	15.96	4.52	53.18	15.08
	657	30	6465	0.80	16.73	5.38	55.77	17.95
	714	30	6462	0.73	16.00	4.76	53.34	15.86
	916	**30**	**6494**	**0.78**	**16.55**	**5.21**	**55.16**	**17.36**

Example Equating Output

In total, 12 separate equating analyses were conducted (3 forms equated to a base form for each of 4 grade level tests by 1 subject). Since these produce a large quantity of output, only one example equating is demonstrated here. However, similar output is available for all other equating analyses. The following example is based on Grade 4 Science. For this equating, Form 271 was equated to Form 435 (the base form) on the percent correct scale. Both of these forms contain 44 items. Employing the standards listed in the "Equating Criteria" section, the first four moments of the equated scores from several competing methods were compared to the base form. Additionally, raw score distributions were inspected to determine comparability. The histograms for Base Form 435 and equated Form 271 are contained in Figures 4.1 and 4.2, respectively. These figures show that distributions of scores were generally very similar, indicating the appropriateness of equating these forms.

Figure 4.1. Raw score distribution for base Form 435, Grade 4 Science

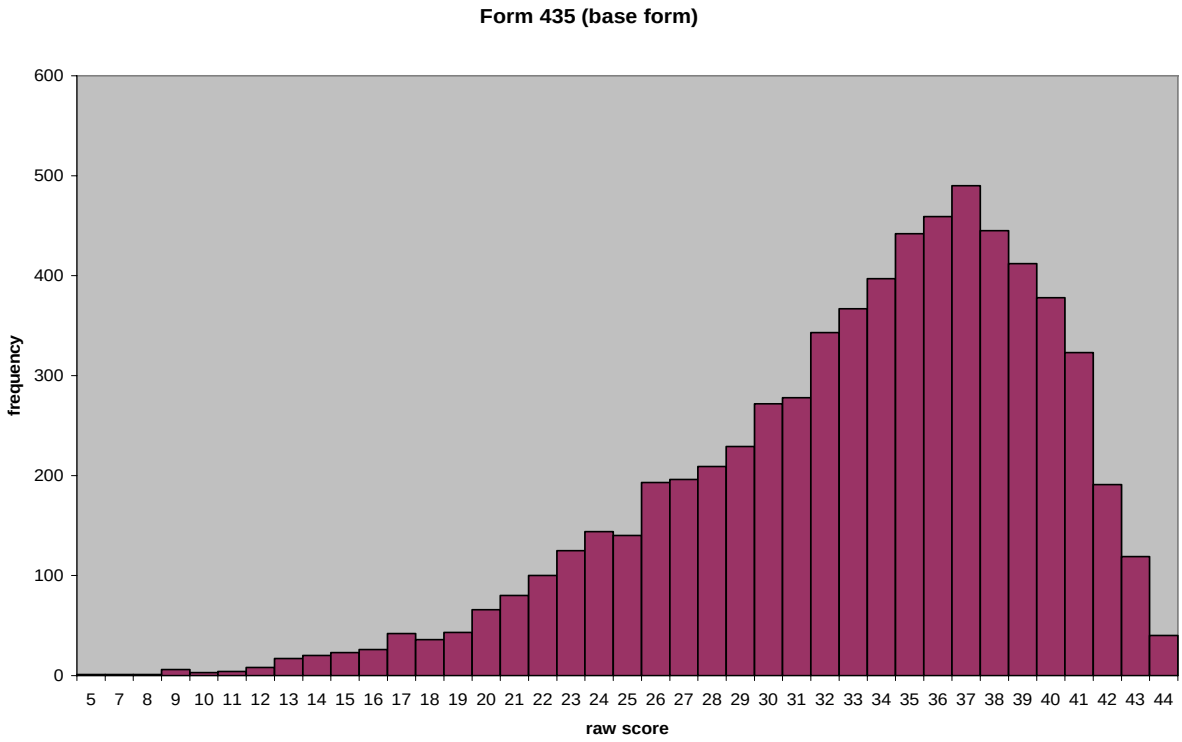


Figure 4.2. Raw score distribution for equated Form 271, Grade 4 Science

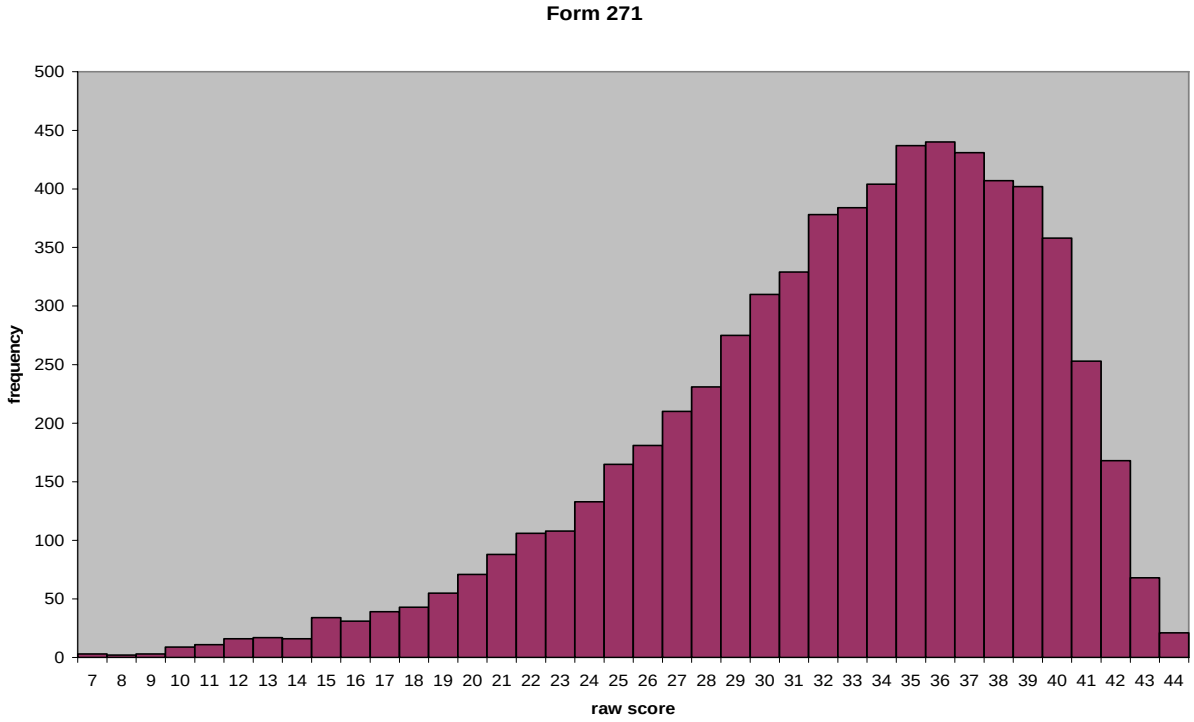


Table 4.5 shows that the equipercentile methods yielded equated number-correct scores with very close approximations of the first four moments of the distribution for the base form, as expected. The values in bold indicate the final equating solution (S = 0.1) chosen for this equating.

Table 4.5
Moments of the Base Form (Form 435, N = 6669) and Equated Form (Form 271, N = 6637) by Equating Method

Test Form/Method	Mean	SD	Skewness	Kurtosis
Raw Scores				
Form 435	33.1753	6.4489	-0.8175	3.3681
Form 271	32.6120	6.4859	-0.8168	3.4497
Form 271 equated to Base Form 435				
Unsmoothed	33.1743	6.4386	-0.8204	3.3668
S=0.01	33.1756	6.4464	-0.8141	3.3457
S=0.05	33.1760	6.4460	-0.8136	3.3516
S=0.10	**33.1762**	**6.4453**	**-0.8134**	**3.3599**
S=0.20	33.1760	6.4437	-0.8148	3.3829
S=0.30	33.1752	6.4412	-0.8178	3.4101
S=0.40	33.1741	6.4381	-0.8212	3.4367
S=0.50	33.1723	6.4345	-0.8237	3.4593
S=0.75	33.1697	6.4306	-0.8229	3.4673
S=1.00	33.1697	6.4306	-0.8229	3.4673

The mean difference between base Form 435 and Form 271 was 0.56. After equating, the mean difference between the forms was -0.0009. Figure 4.3 illustrates the conversion by smoothing method compared to an unsmoothed solution across the total score distribution. The figure indicates that the smoothed equipercentile equating method selected provides a reasonable conversion across score levels, particularly in the area of the distribution where the majority of the data congregated (i.e., above chance-level success, which is a score of 11 or greater on this test).

Figure 4.3. Grade 4 Science Form 271 equated to base Form 435 (number-correct scale)

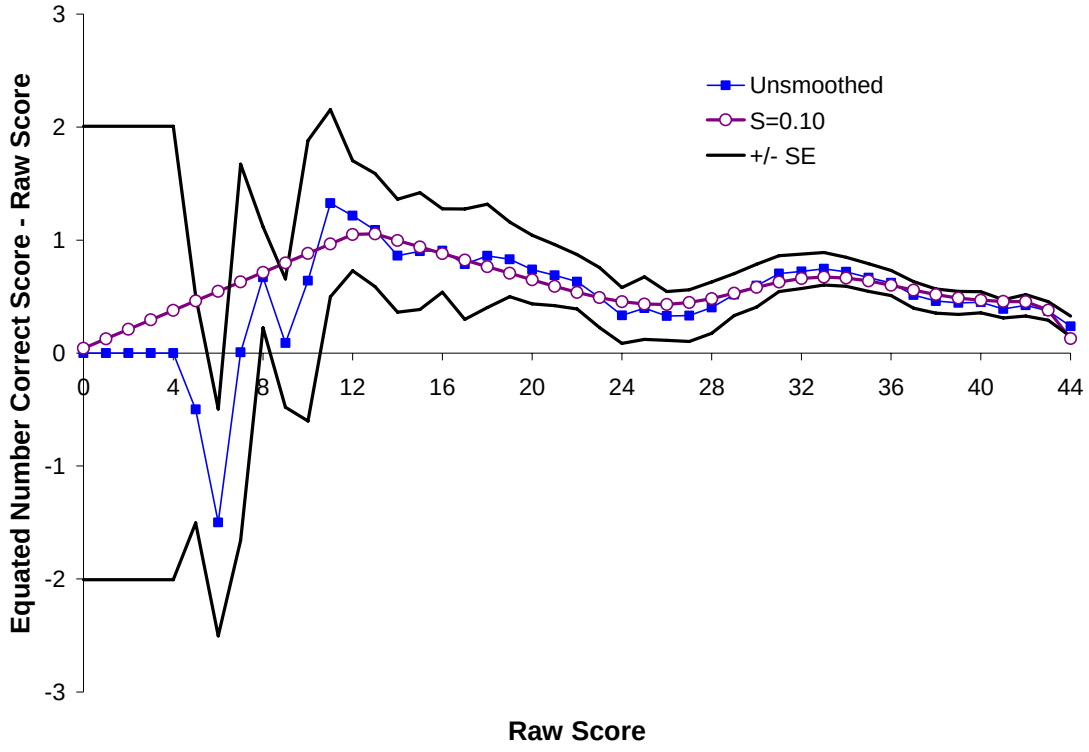


Table 4.6 shows the respective conversion table for all competing methods. Most conversions showed reasonable progression of equated scores through the raw score scale. The values in bold indicate the final equating solution (S = 0.10) chosen for this equating.

Table 4.6
Conversion Table for Various Methods

Raw Score	Unsmoothed	S=0.01	S=0.05	**S=0.10**	S=0.20	S=0.30	S=0.40	S=0.50	S=0.75	S=1.00
0	0.00	0.04	0.04	**0.04**	0.04	0.04	0.03	0.03	0.03	0.03
1	1.00	1.13	1.13	**1.13**	1.12	1.11	1.10	1.09	1.08	1.08
2	2.00	2.21	2.21	**2.21**	2.20	2.18	2.16	2.14	2.14	2.14
3	3.00	3.30	3.30	**3.29**	3.27	3.25	3.22	3.20	3.20	3.20
4	4.00	4.38	4.38	**4.38**	4.35	4.32	4.29	4.26	4.25	4.25
5	4.50	5.47	5.47	**5.46**	5.43	5.39	5.35	5.32	5.31	5.31
6	4.50	6.56	6.55	**6.55**	6.51	6.46	6.41	6.38	6.37	6.37
7	7.01	7.64	7.63	**7.63**	7.59	7.53	7.48	7.43	7.42	7.42
8	8.67	8.73	8.72	**8.71**	8.67	8.60	8.54	8.49	8.48	8.48
9	9.09	9.81	9.80	**9.80**	9.74	9.67	9.60	9.55	9.53	9.53
10	10.64	10.90	10.89	**10.88**	10.82	10.74	10.67	10.61	10.59	10.59
11	12.33	11.98	11.97	**11.97**	11.90	11.81	11.73	11.66	11.65	11.65
12	13.22	13.07	13.06	**13.05**	12.98	12.88	12.79	12.72	12.70	12.70
13	14.09	14.04	14.06	**14.06**	13.99	13.89	13.81	13.74	13.72	13.72
14	14.86	14.97	15.01	**15.00**	14.94	14.86	14.79	14.73	14.72	14.72
15	15.90	15.92	15.95	**15.94**	15.88	15.82	15.77	15.72	15.71	15.71
16	16.91	16.88	16.90	**16.88**	16.83	16.78	16.74	16.71	16.70	16.70
17	17.79	17.85	17.85	**17.82**	17.78	17.75	17.72	17.69	17.69	17.69
18	18.86	18.83	18.79	**18.76**	18.73	18.71	18.69	18.68	18.69	18.69
19	19.83	19.80	19.73	**19.71**	19.68	19.67	19.67	19.67	19.68	19.68
20	20.74	20.74	20.67	**20.65**	20.64	20.64	20.65	20.66	20.67	20.67
21	21.69	21.67	21.61	**21.59**	21.60	21.61	21.63	21.65	21.66	21.66
22	22.63	22.58	22.54	**22.54**	22.56	22.58	22.61	22.63	22.65	22.65
23	23.49	23.48	23.48	**23.49**	23.53	23.56	23.60	23.62	23.64	23.64
24	24.33	24.39	24.43	**24.45**	24.50	24.55	24.58	24.62	24.63	24.63
25	25.40	25.34	25.40	**25.43**	25.49	25.54	25.58	25.61	25.63	25.63
26	26.33	26.32	26.39	**26.43**	26.49	26.53	26.57	26.60	26.62	26.62
27	27.33	27.34	27.41	**27.45**	27.50	27.54	27.57	27.60	27.61	27.61
28	28.40	28.40	28.45	**28.48**	28.52	28.55	28.57	28.59	28.60	28.60
29	29.52	29.49	29.51	**29.53**	29.55	29.56	29.58	29.59	29.59	29.59
30	30.60	30.60	30.58	**30.58**	30.58	30.58	30.58	30.59	30.58	30.58
31	31.70	31.68	31.65	**31.63**	31.61	31.59	31.59	31.58	31.57	31.57
32	32.72	32.73	32.69	**32.66**	32.62	32.60	32.59	32.58	32.57	32.57
33	33.75	33.74	33.70	**33.67**	33.63	33.61	33.59	33.57	33.56	33.56
34	34.72	34.72	34.69	**34.66**	34.63	34.60	34.58	34.56	34.55	34.55
35	35.67	35.67	35.65	**35.64**	35.61	35.59	35.57	35.56	35.54	35.54
36	36.62	36.61	36.60	**36.60**	36.59	36.58	36.56	36.55	36.53	36.53
37	37.51	37.53	37.55	**37.56**	37.56	37.56	37.55	37.53	37.52	37.52
38	38.46	38.47	38.50	**38.52**	38.53	38.53	38.53	38.52	38.51	38.51
39	39.44	39.45	39.47	**39.49**	39.50	39.51	39.51	39.51	39.51	39.51
40	40.45	40.45	40.46	**40.47**	40.48	40.49	40.50	40.50	40.50	40.50
41	41.39	41.45	41.45	**41.46**	41.47	41.47	41.48	41.48	41.49	41.49
42	42.42	42.47	42.46	**42.45**	42.45	42.46	42.46	42.47	42.48	42.48
43	43.37	43.39	43.39	**43.38**	43.37	43.38	43.38	43.39	43.40	43.40
44	44.24	44.13	44.13	**44.13**	44.12	44.13	44.13	44.13	44.14	44.14

Table 4.7 shows the same conversion table when transformed onto the percent correct metric for the same sample of score values. In this percent correct metric, differences in distributional smoothness throughout the scale are not immediately apparent. However, using these conversion tables in accord with the moments output from the various methods and the graphical representation of the conversions for each method is useful. Considering the multiple criteria for making equating decisions, the equipercentile method with a smoothing parameter applied (S = 0.10) gave the most reasonable conversion.

Table 4.7

Conversion Table for Various Methods Expressed in Percent Correct Metric

Raw Score	Unsmoothed	S=0.01	S=0.05	**S=0.10**	S=0.20	S=0.30	S=0.40	S=0.50	S=0.75	S=1.00
0	0	0	0	**0**	0	0	0	0	0	0
1	2	3	3	**3**	3	3	2	2	2	2
2	5	5	5	**5**	5	5	5	5	5	5
3	7	8	7	**7**	7	7	7	7	7	7
4	9	10	10	**10**	10	10	10	10	10	10
5	10	12	12	**12**	12	12	12	12	12	12
6	10	15	15	**15**	15	15	15	14	14	14
7	16	17	17	**17**	17	17	17	17	17	17
8	20	20	20	**20**	20	20	19	19	19	19
9	21	22	22	**22**	22	22	22	22	22	22
10	24	25	25	**25**	25	24	24	24	24	24
11	28	27	27	**27**	27	27	27	27	26	26
12	30	30	30	**30**	29	29	29	29	29	29
13	32	32	32	**32**	32	32	31	31	31	31
14	34	34	34	**34**	34	34	34	33	33	33
15	36	36	36	**36**	36	36	36	36	36	36
16	38	38	38	**38**	38	38	38	38	38	38
17	40	41	41	**41**	40	40	40	40	40	40
18	43	43	43	**43**	43	43	42	42	42	42
19	45	45	45	**45**	45	45	45	45	45	45
20	47	47	47	**47**	47	47	47	47	47	47
21	49	49	49	**49**	49	49	49	49	49	49
22	51	51	51	**51**	51	51	51	51	51	51
23	53	53	53	**53**	53	54	54	54	54	54
24	55	55	56	**56**	56	56	56	56	56	56
25	58	58	58	**58**	58	58	58	58	58	58
26	60	60	60	**60**	60	60	60	60	60	60
27	62	62	62	**62**	63	63	63	63	63	63
28	65	65	65	**65**	65	65	65	65	65	65
29	67	67	67	**67**	67	67	67	67	67	67
30	70	70	70	**70**	69	70	70	70	70	70
31	72	72	72	**72**	72	72	72	72	72	72
32	74	74	74	**74**	74	74	74	74	74	74
33	77	77	77	**77**	76	76	76	76	76	76
34	79	79	79	**79**	79	79	79	79	79	79
35	81	81	81	**81**	81	81	81	81	81	81
36	83	83	83	**83**	83	83	83	83	83	83
37	85	85	85	**85**	85	85	85	85	85	85
38	87	87	88	**88**	88	88	88	88	88	88
39	90	90	90	**90**	90	90	90	90	90	90
40	92	92	92	**92**	92	92	92	92	92	92
41	94	94	94	**94**	94	94	94	94	94	94
42	96	97	97	**96**	96	96	97	97	97	97
43	99	99	99	**99**	99	99	99	99	99	99
44	101	100	100	**100**	100	100	100	100	100	100

Equating Decisions

The equating methods selected for each form in science are summarized in this section of the report. A total of 12 equating analyses were performed in science, each of which was subjected to the criteria previously listed. Equating analyses in 2008 were only conducted at the total score level. For all forms, a value of zero on the raw score scale converted to a value of zero, regardless of equating method adopted. Additionally, when relevant, any equated scores that had a negative value were set to the minimum score value of zero. Similarly, equated scores that were greater than the top score on the base form were set to a percent correct value of 100%. All methods selected across grade levels in science required the use of conversion tables. These conversion tables, similar to the samples detailed in Tables 4.6 and 4.7, are not provided in this report because of space considerations.

After comparing all possible methods employing the criteria set forth in the previous section, decisions were made for each form individually. The smoothed equipercentile equating method was chosen exclusively for all 12 equating decisions in science. The smoothing parameter, S, selected for each decision was 0.01, 0.05, or 0.10. All forms were equated at the raw score level and subsequently expressed in the percent correct metric.

Table 4.8 details the equating decisions adopted for science. The smoothing parameters varied across forms and are detailed in the table.

Table 4.8
Summary of Equating Decisions for Science

Science		
Grade	**Form**	**S**
4 (Base Form 435)	271	0.10
	618	0.10
	967	0.10
7 (Base Form 726)	236	0.10
	589	0.05
	892	0.05
11 Life (Base Form 118)	249	0.10
	369	0.10
	875	0.10
11 Physical (Base Form 916)	449	0.01
	657	0.10
	714	0.01

Summary

Test forms for the 2008 Kansas Science Assessments were built to the same specifications as articulated by KSDE and were developed to the same statistical specifications. The reliability coefficients for all forms were acceptable for the purpose of equating. Further, data collected from the Spring 2008 administration show that groups administered various test forms appeared to be random.

Several classical equipercentile equating methods were considered for each equating. Certain criteria were used to select the equating method for a particular test form that would provide for the most equitable scores for the Kansas students administered the assessments. Methods that best fit the data through the criteria listed were selected.

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

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

2008 Grade 4 Science Equating Conversion Table Results

	Base Form 435		Form 271		Form 618		Form 967	
			S = 0.10	S = 0.10	S = 0.10	S = 0.10	S = 0.10	S = 0.10
Raw Score	Score	% Correct	Equated Score	Equated % Correct	Equated Score	Equated % Correct	Equated Score	Equated % Correct
0	0	0	0.04	0	-0.01	0	0.03	0
1	1	2	1.13	3	0.97	2	1.10	3
2	2	5	2.21	5	1.95	4	2.17	5
3	3	7	3.29	7	2.93	7	3.24	7
4	4	9	4.38	10	3.91	9	4.31	10
5	5	11	5.46	12	4.89	11	5.38	12
6	6	14	6.55	15	5.87	13	6.45	15
7	7	16	7.63	17	6.85	16	7.51	17
8	8	18	8.71	20	7.83	18	8.58	20
9	9	20	9.80	22	8.82	20	9.65	22
10	10	23	10.88	25	9.80	22	10.72	24
11	11	25	11.97	27	10.78	24	11.79	27
12	12	27	13.05	30	11.76	27	12.86	29
13	13	30	14.06	32	12.74	29	13.85	31
14	14	32	15.00	34	13.72	31	14.79	34
15	15	34	15.94	36	14.74	34	15.73	36
16	16	36	16.88	38	15.77	36	16.67	38
17	17	39	17.82	41	16.79	38	17.60	40
18	18	41	18.76	43	17.80	40	18.54	42
19	19	43	19.71	45	18.80	43	19.48	44
20	20	45	20.65	47	19.78	45	20.42	46
21	21	48	21.59	49	20.75	47	21.35	49
22	22	50	22.54	51	21.68	49	22.28	51
23	23	52	23.49	53	22.61	51	23.21	53
24	24	55	24.45	56	23.53	53	24.14	55
25	25	57	25.43	58	24.46	56	25.09	57
26	26	59	26.43	60	25.41	58	26.06	59
27	27	61	27.45	62	26.40	60	27.06	62
28	28	64	28.48	65	27.41	62	28.08	64
29	29	66	29.53	67	28.44	65	29.12	66
30	30	68	30.58	70	29.49	67	30.17	69
31	31	70	31.63	72	30.55	69	31.21	71
32	32	73	32.66	74	31.62	72	32.23	73
33	33	75	33.67	77	32.69	74	33.22	76
34	34	77	34.66	79	33.76	77	34.19	78
35	35	80	35.64	81	34.81	79	35.15	80
36	36	82	36.60	83	35.85	81	36.09	82
37	37	84	37.56	85	36.86	84	37.05	84
38	38	86	38.52	88	37.87	86	38.01	86
39	39	89	39.49	90	38.88	88	38.99	89
40	40	91	40.47	92	39.90	91	39.98	91
41	41	93	41.46	94	40.94	93	40.97	93
42	42	95	42.45	96	41.99	95	41.95	95
43	43	98	43.38	99	43.04	98	42.93	98
44	44	100	44.13	100	44.01	100	43.98	100

2008 Grade 7 Science Equating Conversion Table Results

	Base Form 726		Form 236		Form 589		Form 892	
Raw Score	Score	% Correct	S = 0.10 Equated Score	S = 0.10 Equated % Correct	S = 0.05 Equated Score	S = 0.05 Equated % Correct	S = 0.05 Equated Score	S = 0.05 Equated % Correct
0	0	0	0.00	0	-0.01	0	0.00	0
1	1	2	1.01	2	0.99	2	0.99	2
2	2	3	2.01	3	1.98	3	1.99	3
3	3	5	3.01	5	2.97	5	2.98	5
4	4	7	4.02	7	3.96	7	3.98	7
5	5	8	5.02	8	4.95	8	4.97	8
6	6	10	6.03	10	5.94	10	5.97	10
7	7	12	7.03	12	6.93	12	6.96	12
8	8	13	8.04	13	7.92	13	7.95	13
9	9	15	9.04	15	8.91	15	8.95	15
10	10	17	10.04	17	9.90	16	9.94	17
11	11	18	11.05	18	10.89	18	10.94	18
12	12	20	12.05	20	11.88	20	11.93	20
13	13	22	13.06	22	12.87	21	12.93	22
14	14	23	14.06	23	13.86	23	13.92	23
15	15	25	15.06	25	14.85	25	14.92	25
16	16	27	15.98	27	15.93	27	16.02	27
17	17	28	16.91	28	17.02	28	17.13	29
18	18	30	17.83	30	18.10	30	18.22	30
19	19	32	18.75	31	19.17	32	19.31	32
20	20	33	19.67	33	20.22	34	20.37	34
21	21	35	20.59	34	21.25	35	21.41	36
22	22	37	21.51	36	22.26	37	22.41	37
23	23	38	22.43	37	23.25	39	23.39	39
24	24	40	23.35	39	24.24	40	24.36	41
25	25	42	24.27	40	25.21	42	25.34	42
26	26	43	25.20	42	26.17	44	26.32	44
27	27	45	26.13	44	27.12	45	27.31	46
28	28	47	27.07	45	28.07	47	28.31	47
29	29	48	28.01	47	29.02	48	29.32	49
30	30	50	28.97	48	29.97	50	30.33	51
31	31	52	29.94	50	30.92	52	31.34	52
32	32	53	30.92	52	31.88	53	32.35	54
33	33	55	31.91	53	32.85	55	33.37	56
34	34	57	32.91	55	33.85	56	34.39	57
35	35	58	33.93	57	34.86	58	35.42	59
36	36	60	34.95	58	35.88	60	36.43	61
37	37	62	35.98	60	36.88	61	37.44	62
38	38	63	37.02	62	37.87	63	38.45	64
39	39	65	38.06	63	38.84	65	39.45	66
40	40	67	39.11	65	39.80	66	40.44	67
41	41	68	40.15	67	40.78	68	41.41	69
42	42	70	41.21	69	41.77	70	42.38	71
43	43	72	42.27	70	42.78	71	43.35	72
44	44	73	43.34	72	43.81	73	44.34	74

Raw Score	Base Form 726 Score	Base Form 726 % Correct	Form 236 S = 0.10 Equated Score	Form 236 S = 0.10 Equated % Correct	Form 589 S = 0.05 Equated Score	Form 589 S = 0.05 Equated % Correct	Form 892 S = 0.05 Equated Score	Form 892 S = 0.05 Equated % Correct
45	45	75	44.42	74	44.86	75	45.35	76
46	46	77	45.50	76	45.93	77	46.36	77
47	47	78	46.59	78	46.99	78	47.35	79
48	48	80	47.67	79	48.04	80	48.32	81
49	49	82	48.76	81	49.08	82	49.26	82
50	50	83	49.84	83	50.12	84	50.18	84
51	51	85	50.92	85	51.16	85	51.10	85
52	52	87	51.98	87	52.21	87	52.02	87
53	53	88	53.05	88	53.26	89	52.96	88
54	54	90	54.11	90	54.33	91	53.92	90
55	55	92	55.17	92	55.43	92	54.90	92
56	56	93	56.23	94	56.54	94	55.93	93
57	57	95	57.27	95	57.46	96	56.97	95
58	58	97	58.19	97	58.33	97	57.98	97
59	59	98	59.12	99	59.20	99	58.99	98
60	60	100	60.04	100	60.07	100	60.00	100

2008 High School Physical Science Equating Conversion Table Results

	Base Form 916		Form 449 S = 0.01	S = 0.01	Form 657 S = 0.10	S = 0.10	Form 714 S = 0.01	S = 0.01
Raw Score	Score	% Correct	Equated Score	Equated % Correct	Equated Score	Equated % Correct	Equated Score	Equated % Correct
0	0	0	-0.05	0	-0.03	0	-0.05	0
1	1	3	0.86	3	0.91	3	0.86	3
2	2	7	1.77	6	1.85	6	1.77	6
3	3	10	2.67	9	2.79	9	2.68	9
4	4	13	3.58	12	3.73	12	3.59	12
5	5	17	4.48	15	4.67	16	4.50	15
6	6	20	5.41	18	5.67	19	5.43	18
7	7	23	6.44	21	6.76	23	6.43	21
8	8	27	7.50	25	7.84	26	7.50	25
9	9	30	8.60	29	8.89	30	8.62	29
10	10	33	9.66	32	9.91	33	9.75	33
11	11	37	10.68	36	10.92	36	10.89	36
12	12	40	11.75	39	11.93	40	12.08	40
13	13	43	12.91	43	12.95	43	13.28	44
14	14	47	14.14	47	13.97	47	14.44	48
15	15	50	15.38	51	14.99	50	15.60	52
16	16	53	16.64	55	16.02	53	16.78	56
17	17	57	17.90	60	17.03	57	17.90	60
18	18	60	19.08	64	18.02	60	18.95	63
19	19	63	20.23	67	18.97	63	20.00	67
20	20	67	21.40	71	19.89	66	21.04	70
21	21	70	22.56	75	20.79	69	22.05	73
22	22	73	23.65	79	21.65	72	23.04	77
23	23	77	24.70	82	22.51	75	24.03	80
24	24	80	25.63	85	23.37	78	25.00	83
25	25	83	26.46	88	24.23	81	25.96	87
26	26	87	27.25	91	25.12	84	26.83	89
27	27	90	28.00	93	26.02	87	27.62	92
28	28	93	28.72	96	26.94	90	28.44	95
29	29	97	29.43	98	27.88	93	29.26	98
30	30	100	30.14	100	29.61	99	30.09	100

2008 High School Life Science Equating Conversion Table Results

	Base Form 118		Form 249		Form 369		Form 875	
			S = 0.10	S = 0.10	S = 0.10	S = 0.10	S = 0.10	S = 0.10
Raw Score	Score	% Correct	Equated Score	Equated % Correct	Equated Score	Equated % Correct	Equated Score	Equated % Correct
0	0	0	0.00	0	0.06	0	0.00	0
1	1	3	1.00	3	1.18	4	1.01	3
2	2	7	2.00	7	2.30	8	2.02	7
3	3	10	3.00	10	3.42	11	3.03	10
4	4	13	4.01	13	4.54	15	4.03	13
5	5	17	5.01	17	5.66	19	5.04	17
6	6	20	6.01	20	6.69	22	6.05	20
7	7	23	6.98	23	7.71	26	7.05	24
8	8	27	7.96	27	8.72	29	8.05	27
9	9	30	8.95	30	9.74	32	9.06	30
10	10	33	9.95	33	10.75	36	10.06	34
11	11	37	10.98	37	11.75	39	11.05	37
12	12	40	12.03	40	12.73	42	12.04	40
13	13	43	13.11	44	13.71	46	13.03	43
14	14	47	14.21	47	14.69	49	14.04	47
15	15	50	15.34	51	15.66	52	15.05	50
16	16	53	16.48	55	16.65	55	16.10	54
17	17	57	17.63	59	17.64	59	17.16	57
18	18	60	18.78	63	18.64	62	18.23	61
19	19	63	19.92	66	19.65	65	19.32	64
20	20	67	21.03	70	20.65	69	20.41	68
21	21	70	22.10	74	21.66	72	21.50	72
22	22	73	23.14	77	22.67	76	22.59	75
23	23	77	24.16	81	23.68	79	23.67	79
24	24	80	25.15	84	24.69	82	24.72	82
25	25	83	26.12	87	25.70	86	25.75	86
26	26	87	27.08	90	26.70	89	26.76	89
27	27	90	28.03	93	27.70	92	27.76	93
28	28	93	28.74	96	28.52	95	28.57	95
29	29	97	29.44	98	29.31	98	29.34	98
30	30	100	30.15	100	30.11	100	30.11	100

Section 5

STANDARD SETTING

2008 Kansas Science Performance Standards

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

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

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

1. Development of Performance Level Names

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

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

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

2. Development of Performance Level Descriptors

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

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

3. Bookmark Procedure

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

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

Preparation of Item Ordered Booklets

The following describes the creation of ordered item booklets which were prepared for Bookmark standard setting activities that took place in Summer 2008. Standard setting activities for science were conducted at grades 4, 7, and high school (life science and physical science). For each of these four grade level tests, a Kansas Assessment of Modified Measures (KAMM) was also administered, thus ordered item booklets were also created for KAMM. The 8 tests for which ordered item booklets were created are listed in Table 5.1.

Table 5.1
Overview of Tests for Which Performance Standards Were Set

Subject	Grade	# Items (General)	# Items (KAMM)
Science	4	44	42
Science	7	60	60
Life Science	High School	30	30
Physical Science	High School	30	30

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

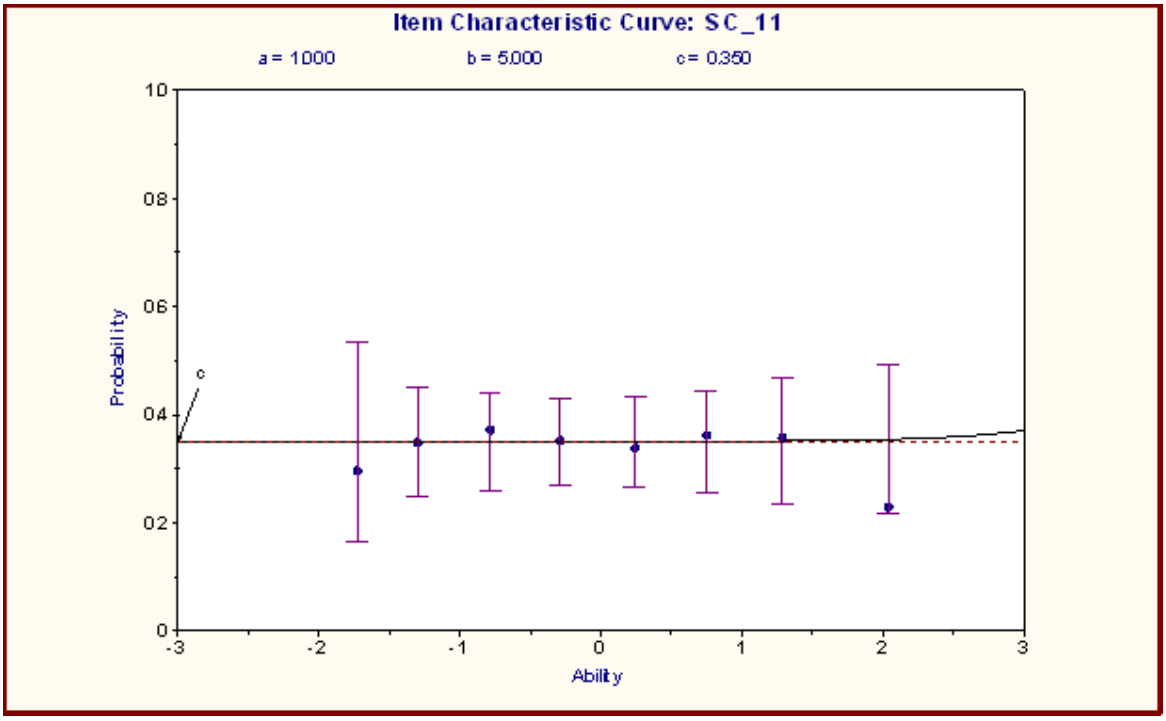
All item response data were fit to the three-parameter logistic (3-PL) IRT model (e.g., Lord, 1980) using BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 2002). The 3-PL model relates the probability of success for examinee *j* on item *i* (i.e., the item response, u_{ij} = 1 instead of 0) as a function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

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

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

For some assessments, a small number of items (no more than four items for any one assessment) with poor statistical qualities (point-biserial correlations between item response and total test score approximately equal to zero or less) caused problems with the convergence of solutions. Inspection of item-true score regressions indicated that these items were difficult enough for examinees that no ability level performed better than chance-levels of success. As a result, maximum likelihood solutions for difficulty parameters could not be obtained. In this situation, the following procedure was employed: (1) all items were calibrated while excluding the items with poor statistical qualities, (2) item parameter estimates were fixed at their estimated values, and (3) the items with poor statistical quality were placed back in the dataset, with the a-parameter (discrimination) set equal to 1, the b-parameter (difficulty) set equal to 5, and the c-parameter (lower asymptote) estimated using BILOG-MG. This procedure produced item response functions like the example in Figure 5.1. This figure illustrates that the item characteristic curve is effectively flat in the region of ability where most examinees are located (from -3 to 3 on a standardized metric). The only effect on the probability of success is the c-parameter. Items such as these, although they do not increase reliability nor the precision of ability measurement, were still included in the ordered item booklets so that no items would be eliminated. It should be noted that these items tended to be the most difficult on any assessment, and thus such items were always placed toward the very back of the ordered item booklets.

Figure 5.1. Example item characteristic curve for an item with poor statistical quality. This was item #11 on the High School KAMM Life Science Assessment.



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

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

where θ_P is the ability level for an examinee for which $P(u_{ij}=1) = P$, P is the desired level of probability (in this case, $P = 0.67$), and all other terms have been defined previously. Table 5.2 contains an example of an item ordering on the basis of PR-67 values. A table like this was created for each of the science assessments, and ordered item booklets were created on the basis of them.

Table 5.2

Item Ordering for the High School KAMM Science Assessment

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

(Note that item #11, the item demonstrated in Figure 5.1, is the last item on the list.)

Panel Participants

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

We are bringing together a group of Kansas educators to provide input toward setting performance standards (i.e., cut scores) for the new Kansas assessments in science. This is important work, and we plan to involve as many interested and available persons as possible. Below are the considerations for your building nominee.

1. *Only nominate a person with whom you have spoken and that person has agreed to attend <u>both days</u> of scheduled meetings if selected.*
2. *The meetings will be held in Kansas City (meeting location was later changed to Overland Park) from 1PM to 5PM, Friday, June 20th, then continue on Saturday, June 21st starting at 8:30AM and ending by 2PM.*
3. *A person may only serve in one content area, the area in which he or she is nominated by you. Participants will receive travel and meal allowances along with a $175 stipend for their participation; lodging will be paid by CETE (double occupancy in the local hotel). Nominees will be notified by June 2 if they are selected, at which time details will be provided.*

From among the nominees, participants will be chosen based on expertise, instructional/ supervisory experience, and qualifications at the specific content (science), grade (elementary, middle/jr. high, or high school), and test type (General or KAMM assessment) for the students being assessed. Other participation selection factors when considering prospective nominees will include:

- *highly regarded and respected local educators with at least three years experience;*
- *instructional/supervisory experiences with students who have disabilities, students with limited English proficiency, and other subgroups;*
- *balanced regional representation;*

- *access to an email address to receive communication (even after schools close); and,*
- *building or district administrators with qualifications are also eligible to be nominated.*

*While we would like to involve as many educators as may be interested in this process, that is not possible. Based on KSDE information and random selection principles, we request that you nominate one educator to represent your building who meets the qualifications and condition identified above **at grade GG in science with the TEST**. We invite you to nominate one person who, in your judgment, meets the experience, expertise, and training criteria to serve with other Kansas educators. Your nominee(s) must be highly qualified, held in esteem by peers, and have at least three years experience teaching at the grade and content area as a general or special educator (note: special educators may be nominated for the general assessment slots if desired, as many of their students take the general assessments). We rely on your professional judgment to nominate an individual who can help guide Kansas education and expectations for the future.*

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

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

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

Each of 46 Kansas educators participated on one of 7 panels as part of the Bookmark process. Table 5.3 presents the number of participants on each panel.

Table 5.3

Number of Participants on Each Benchmark Panel

Grade and Subject	General
Grade 4 Science	8
Grade 7 Science	8
HS Life Science	6
HS Physical Science	8

Of these participants, 34 were female and 12 were male; 44 were Caucasian and 2 were minorities; 3 had fewer than two years teaching experience, 7 had 3-5 years experience, 9 had 6-10 years, 16 had 11-20 years, and 11 had more than 20 years; 4 came from inner city schools, 3 from other urban, 17 from suburban and 22 from rural.

Bookmark Results

After the third round of judgments, the average of the panelist cut scores was determined and transformed to the 0 to 100 reported score scale metric. Judgments were rounded to the nearest integer value. Table 5.4 presents the average Bookmark Procedure recommended score ranges for each performance category for the science assessments.

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

Table 5.4

Science Bookmark Procedure Recommended Performance Level Score Ranges

Performance Level	General		
	4	7	HS
Academic Warning	0-27	0-36	0-26
Approaches Standard	28-43	37-50	27-43
Meets Standard	44-66	51-67	44-75
Exceeds Standard	67-86	68-81	76-90
Exemplary	87-100	82-100	91-100

Evaluation of Bookmark Procedure

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

Standard Setting Policy Advisory Group

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

Advisory Group Participants

Two members from each of the 14 Bookmark Procedure panels (14 table leaders and 14 who were nominated by their peers) were invited to participate in the policy advisory meeting to ensure that the process and results would be well represented. Of these 28 invitees, 27 were available and participated. An additional 29 people representing a variety of

constituency groups also participated. Demographically, the 56 participants included 35 classroom teachers, three building administrators, 11 district administrators, five parents/grandparents of Kansas students, and two representatives of state educational organizations. These same 56 members included 51 Caucasians, three African Americans, one Hispanic, and one Native American. Thirty-two of these participants were female and 24 were male. Rural and urban, and eastern, central, and western areas of the state were all represented.

Policy Advisory Meeting Agenda

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

- *Introductions, logistics, and agenda*.
- *Context and purpose*. Participants were provided with an overview of the state assessment program and the steps that had occurred so far. Issues of consistency were discussed as was their task of ensuring an appropriate level of consistency and recommending specific cut scores the Kansas Department of Education.
- *Review of performance level descriptors*. Performance level descriptors were reviewed to provide further grounding and to ensure that the external consistency issues they were considering (cross-grade and cross-test type) were within the context of consistency with performance level descriptors.
- *Description of Bookmark standard setting process*. The Bookmark procedures were described so that advisory group members who did not participate in the Bookmark process would nonetheless understand it.
- *Results of Bookmark process*. The Bookmark recommended cut scores were presented to the participants.
- *Recommended consistent performance standards*. Using procedures explained below, panelists were presented examples of cut scores adjusted for inconsistencies within subject but across grades and two of the test types, general and KAMM. It was stressed that this was an example and that participants could indicate they agreed with the Bookmark assigned cut scores, with an example of more consistent scores, or with a different cut score. In order to make the task reasonable, data were presented both in terms of cut scores and also percent of students who would fall into each performance level. Table 5.5 and 5.6 and Figures 5.2 present some of the kinds of information presented to the participants. At the end of this presentation, the participants made their recommendations in terms of what percent of students they believed should be in each category.

Procedure for Creating Example of More Consistent Standards

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

1. Cut scores were transformed to the z-score corresponding to the proportion of students at or below that test score. So, for example, if 84% of the sample scored at or below the recommended cut score, the corresponding z-score was 1.0
2. The weighted average of the z-scores was taken, giving 50 percent weight to the z-score for the cut score under consideration and 25% for each of the other two grades. For example, if the z-scores for grades 4, 7, and HS were 0.6, 0.8, and 0.9, respectively, then the resulting z-score for grade 4 would be (0.6 x.5) + (0.8 x .25) + (0.9 x .25), or 0.725.
3. The proportion of a population corresponding to the weighted average z-score was calculated.
4. The raw score that has a cumulative frequency closest to the proportion from step 3 was selected as the more consistent example.

Table 5.5

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

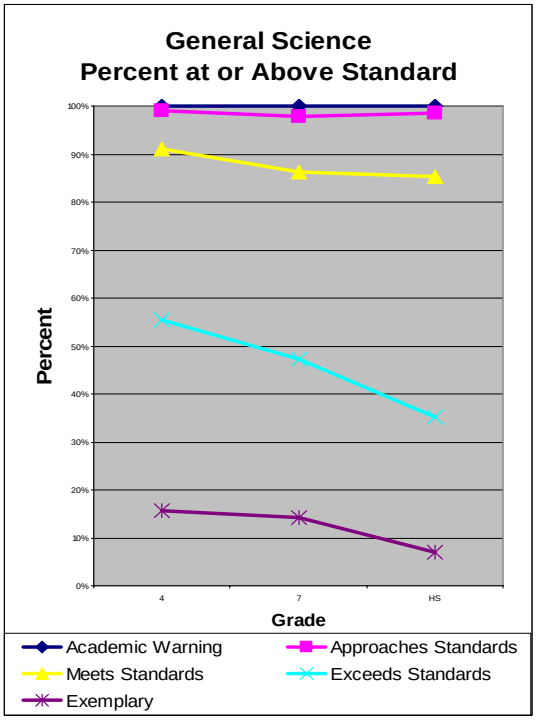
Performance Level	Bookmark Recommended Maximum Possible Scaled Score in Performance Level			Cross-Grade Consistent Maximum Possible Scaled Score in Performance Level		
	4	**7**	**HS**	**4**	**7**	**HS**
Academic Warning	27	36	26	34	30	25
Approaches Standard	43	50	43	53	46	38
Meets Standard	66	67	75	76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

Table 5.6

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

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

Figures 5.2 Examples of more consistent cut score graphs presented to participants



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

Resulting Recommendations from Standard Setting Policy Advisory Group

Table 5.7 presents the score ranges corresponding to the average recommended percent of students in each performance level from the 56 members of the Standard Setting Policy Advisory Group.

Table 5.7

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

Assessment Type	Grade	Academic Warning	Approaches Standard	Meets Standard	Exceeds Standard	Exemplary
General	4	0-31	32-50	51-73	74-88	89-100
General	7	0-30	31-45	46-66	67-81	82-100
General	HS	0-25	26-39	40-65	66-80	81-100

Evaluation of Standard Setting Policy Advisory Group Meeting

At the end, participants evaluated the meeting, and 95 percent of the participants found the training “adequate” or “more than adequate” while none found the training “not adequate.” Some participants found the task of considering consistency issues to be very difficult. Table 5.8 presents the results when participants were asked how confident they were in their final decisions. All 56 people participated in all decisions, though not all responded to each question.

Table 5.8

Percent of Participants Indicating Confidence Level for Final Cut Score Decisions

Performance Category	No Response	Not Confident	Partially Confident	Confident	Very Confident
General Science	4	2	7	39	48

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

proportion of policy advisory group participants chosen from Bookmark participants should be limited to about 20 percent.

State Board of Education Adoption of Performance Standards

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

Section 6

RELIABILITY ANALYSES

Score Reliability

Information on the reliability of test scores for each general assessment test form was provided in Section 4, Table 4.4. The information is condensed and presented below in Table 6.1. The score reliability estimates reported in the tables are Cronbach alpha coefficients. The coefficient values range from a low of 0.71 to a high of 0.88 across all the science forms. The overall general standard errors of measurement on the percent correct score scale range from 5.39 to 8.24 for scores on the science general assessment test forms.

Table 6.1

Science Form Reliabilities by Test Form

Grade	Form	(α) Reliability	SEM % Correct
4	271	0.84	5.90
	435	**0.85**	5.68
	618	0.83	5.91
	967	0.85	5.78
7	236	0.88	5.39
	589	0.88	5.49
	726	**0.88**	5.50
	892	0.88	5.53
11 Life	**118**	**0.77**	**8.20**
	249	0.73	8.16
	369	0.78	8.05
	875	0.75	8.17
11 Physical	449	0.71	8.12
	657	0.8	8.03
	714	0.73	8.24
	916	**0.78**	8.14

The reliability of the composite high school science test was estimated as 0.860 with a corresponding standard error of measurement of 6.0. The estimation was based on the formula for the reliability of a composite score.

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

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

In this analysis, the average scaled score error variance of the four Life Science forms and the four Physical Science forms were used. These error variances were calculated from the scaled, equated form variances and coefficient alpha reliability estimates. Table 6.1 shows the individual form data that were used, as well as the average variance and reliability and the SEM calculated from that average. Since each test is given a weight of 0.5 in the calculation of the composite, that weight was applied to calculate the composite SEM calculated from the Table 6.2 data. The variance of the composite score was calculated directly from the distribution and was 259.00.

Table 6.2

Data Used in Estimating the High School Science Composite Reliability

Form	**Life Science**			**Physical Science**		
	Variance	**Reliability**	**SEM**	**Variance**	**Reliability**	**SEM**
1	295.4	0.730	4.64	295.6	0.699	5.18
2	314.8	0.784	3.83	291.0	0.786	3.65
3	303.6	0.783	3.78	296.6	0.719	4.84
4	294.9	0.752	4.26	317.7	0.790	3.74
Average	302.18	0.762	4.14	300.23	0.749	4.35

Classification Consistency

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

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

Method

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

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

Procedures

Samples

Based on the 2008 administration for grades 4 and 7, base form data were used for these analyses. Four forms were administered on computer and one of those forms was also administered on paper (for those schools that could not or did not want to test on computer).

For high school, performance levels are applied to a composite score based on the life science and physical science tests. Students can take one or both of these tests at one administration or take the two tests at different administrations. With 16 different combinations of forms, the number of students taking one particular combination of forms was relatively low. Also, the composite scores only existed as the average of equated scaled scores. Thus, for the high school assessment classification, consistency and accuracy were based on the distribution of equated scores across all combinations of forms and on estimations of the reliability of the composite score from the average of the reliabilities of the four life science and four physical science forms. Table 6.3 presents descriptive statistics for the three analyses.

Table 6.3

Descriptive Statistics for Each Grade Analysis

Statistic	Grade 4	Grade 7	High School
n	6,669	6,552	31,860
Max possible	44	60	100
Mean	33.1753	38.9254	56.5115
Variance	41.5886	90.8865	258.9966
Reliability	0.8400	0.8810	0.8585

Results

Table 6.4 summarizes the consistency and accuracy results across all five performance level categories for the three grade levels.

Table 6.4

Summary of Consistency and Accuracy Results

Approach		Grade 4	Grade 7	High School
Hanson & Brennan	Consistency	0.63	Not estimable	0.74
	Accuracy	0.73	0.89	0.81
Livingston & Lewis	Consistency	0.63	0.77	0.67
	Accuracy	0.73	0.82	0.76

The Hanson and Brennan approach yielded a consistency estimate greater than one for grade 7. Since this is not a possible value, the table indicates the statistic as not estimable. It is possible that a better estimate could be obtained by replacing the program's default values with estimating the moments, but that was not attempted.

Because the most critical decision for school accountability is whether a student is correctly classified as being at or above the Meets Standard performance level, Table 6.5 presents the results for that binary classification decision.

Table 6.5

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

Approach		Grade 4	Grade 7	High School
Hanson & Brennan	Consistency	0.95	0.91	0.91
	Accuracy	0.96	0.93	0.94
Livingston & Lewis	Consistency	0.95	0.91	0.89
	Accuracy	0.96	0.94	0.92

Conditional Standard Errors of Measurement

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

Procedure

Sample

The analysis of reliability was based on samples of students who were administered the Kansas General Science Assessments in Spring 2008. In 2007, parallel test forms were constructed for grades 4, 7, and 11 (life science and physical science). There were four test forms per grade except for grade 11 which had eight test forms total, four each for life science and physical science. For each grade-level, raw scores from all test forms were scaled and equated to a common percent correct scale.

Method

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

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

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

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

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

Because the Kansas Science Assessments' results are not reported in terms of raw scores but rather in terms of equated scaled scores (the impact of the equating on the CSMEs is probably small and is not taken into account by these procedures), the raw score CSEM had to be converted to a scaled score CSEM. A scaled score CSEM is simply the raw score CSEM multiplied by 100 and then divided by the number of items (*n*) on the test.

Results

Conditional Standard Errors of Measurement (CSEM)

Both a raw score CSEM and a scaled score CSEM were estimated for each test form across the grades. The results are presented in Tables 6.6 – 6.9. For each test form and grade, the general trends are parabolic, which is concave downward. The peaking of CSEM occurs in the middle of the score range. Because the variance of binomial distribution is maximized when the probability of getting an item correct equals 0.5, the CSEM for the number of correct scores is usually greatest in this range, and scores are less reliable in this range. The distributions of scaled score CSEMs for each form and grade are summarized in Figures 6.1 – 6.4.

Table 6.6

Conditional Standard Errors of Measurement (CSEM) for Grade 4 Science by Test Form

Form 271				Form 435				Form 618				Form 967			
Raw Score	Raw Score CSEM	Reported Scaled Score	Scaled Score CSEM	Raw Score	Raw Score CSEM	Reported Scaled Score	Scaled Score CSEM	Raw Score	Raw Score CSEM	Reported Scaled Score	Scaled Score CSEM	Raw Score	Raw Score CSEM	Reported Scaled Score	Scaled Score CSEM
0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0
1	0.9	3	2.1	1	0.9	2	2.1	1	1.0	2	2.2	1	0.9	3	2.1
2	1.3	5	2.9	2	1.3	5	3.0	2	1.3	4	3.0	2	1.3	5	3.0
3	1.6	7	3.6	3	1.6	7	3.6	3	1.6	7	3.7	3	1.6	7	3.6
4	1.8	10	4.1	4	1.8	9	4.1	4	1.8	9	4.2	4	1.8	10	4.1
5	2.0	12	4.5	5	2.0	11	4.5	5	2.0	11	4.6	5	2.0	12	4.5
6	2.1	15	4.8	6	2.1	14	4.9	6	2.2	13	5.0	6	2.2	15	4.9
7	2.3	17	5.2	7	2.3	16	5.2	7	2.3	16	5.3	7	2.3	17	5.2
8	2.4	20	5.4	8	2.4	18	5.5	8	2.5	18	5.6	8	2.4	20	5.5
9	2.5	22	5.7	9	2.5	20	5.7	9	2.6	20	5.9	9	2.5	22	5.8
10	2.6	25	5.9	10	2.6	23	6.0	10	2.7	22	6.1	10	2.6	24	6.0
11	2.7	27	6.1	11	2.7	25	6.2	11	2.8	24	6.3	11	2.7	27	6.2
12	2.8	30	6.3	12	2.8	27	6.3	12	2.9	27	6.5	12	2.8	29	6.4
13	2.8	32	6.4	13	2.9	30	6.5	13	2.9	29	6.6	13	2.9	31	6.5
14	2.9	34	6.6	14	2.9	32	6.6	14	3.0	31	6.8	14	2.9	34	6.7
15	2.9	36	6.7	15	3.0	34	6.7	15	3.0	34	6.9	15	3.0	36	6.8
16	3.0	38	6.8	16	3.0	36	6.8	16	3.1	36	7.0	16	3.0	38	6.9
17	3.0	41	6.9	17	3.1	39	6.9	17	3.1	38	7.1	17	3.1	40	7.0
18	3.1	43	6.9	18	3.1	41	7.0	18	3.2	40	7.2	18	3.1	42	7.0
19	3.1	45	7.0	19	3.1	43	7.1	19	3.2	43	7.2	19	3.1	44	7.1
20	3.1	47	7.0	20	3.1	45	7.1	20	3.2	45	7.3	20	3.1	46	7.1
21	3.1	49	7.0	21	3.1	48	7.1	21	3.2	47	7.3	21	3.1	49	7.2
22	3.1	51	7.1	22	3.1	50	7.1	22	3.2	49	7.3	22	3.2	51	7.2
23	3.1	53	7.0	23	3.1	52	7.1	23	3.2	51	7.3	23	3.1	53	7.2
24	3.1	56	7.0	24	3.1	55	7.1	24	3.2	53	7.3	24	3.1	55	7.1
25	3.1	58	7.0	25	3.1	57	7.1	25	3.2	56	7.2	25	3.1	57	7.1
26	3.1	60	6.9	26	3.1	59	7.0	26	3.2	58	7.2	26	3.1	59	7.0
27	3.0	62	6.9	27	3.1	61	6.9	27	3.1	60	7.1	27	3.1	62	7.0
28	3.0	65	6.8	28	3.0	64	6.8	28	3.1	62	7.0	28	3.0	64	6.9
29	2.9	67	6.7	29	3.0	66	6.7	29	3.0	65	6.9	29	3.0	66	6.8
30	2.9	70	6.6	30	2.9	68	6.6	30	3.0	67	6.8	30	2.9	69	6.7
31	2.8	72	6.4	31	2.9	70	6.5	31	2.9	69	6.6	31	2.9	71	6.5
32	2.8	74	6.3	32	2.8	73	6.3	32	2.9	72	6.5	32	2.8	73	6.4
33	2.7	77	6.1	33	2.7	75	6.2	33	2.8	74	6.3	33	2.7	76	6.2
34	2.6	79	5.9	34	2.6	77	6.0	34	2.7	77	6.1	34	2.6	78	6.0
35	2.5	81	5.7	35	2.5	80	5.7	35	2.6	79	5.9	35	2.5	80	5.8
36	2.4	83	5.4	36	2.4	82	5.5	36	2.5	81	5.6	36	2.4	82	5.5
37	2.3	85	5.2	37	2.3	84	5.2	37	2.3	84	5.3	37	2.3	84	5.2
38	2.1	88	4.8	38	2.1	86	4.9	38	2.2	86	5.0	38	2.2	86	4.9
39	2.0	90	4.5	39	2.0	89	4.5	39	2.0	88	4.6	39	2.0	89	4.5
40	1.8	92	4.1	40	1.8	91	4.1	40	1.8	91	4.2	40	1.8	91	4.1
41	1.6	94	3.6	41	1.6	93	3.6	41	1.6	93	3.7	41	1.6	93	3.6
42	1.3	96	2.9	42	1.3	95	3.0	42	1.3	95	3.0	42	1.3	95	3.0
43	0.9	99	2.1	43	0.9	98	2.1	43	1.0	98	2.2	43	0.9	98	2.1
44	0.0	100	0.0	44	0.0	100	0.0	44	0.0	100	0.0	44	0.0	100	0.0

Table 6.7

Conditional Standard Errors of Measurement (CSEM) for Grade 7 Science by Test Form

Form 236				Form 726				Form 589				Form 892			
Raw Score	Raw Score CSEM	Reported Scaled Score	Scaled Score CSEM	Raw Score	Raw Score CSEM	Reported Scaled Score	Scaled Score CSEM	Raw Score	Raw Score CSEM	Reported Scaled Score	Scaled Score CSEM	Raw Score	Raw Score CSEM	Reported Scaled Score	Scaled Score CSEM
0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0
1	0.9	2	1.6	1	0.9	2	1.6	1	0.9	2	1.6	1	0.9	2	1.6
2	1.3	3	2.2	2	1.3	3	2.2	2	1.3	3	2.2	2	1.3	3	2.2
3	1.6	5	2.6	3	1.6	5	2.7	3	1.6	5	2.7	3	1.6	5	2.7
4	1.8	7	3.0	4	1.8	7	3.1	4	1.8	7	3.1	4	1.8	7	3.1
5	2.0	8	3.4	5	2.0	8	3.4	5	2.0	8	3.4	5	2.0	8	3.4
6	2.2	10	3.6	6	2.2	10	3.7	6	2.2	10	3.7	6	2.2	10	3.7
7	2.3	12	3.9	7	2.4	12	4.0	7	2.4	12	3.9	7	2.4	12	4.0
8	2.5	13	4.1	8	2.5	13	4.2	8	2.5	13	4.2	8	2.5	13	4.2
9	2.6	15	4.3	9	2.6	15	4.4	9	2.6	15	4.4	9	2.6	15	4.4
10	2.7	17	4.5	10	2.8	17	4.6	10	2.7	16	4.6	10	2.8	17	4.6
11	2.8	18	4.7	11	2.9	18	4.8	11	2.9	18	4.8	11	2.9	18	4.8
12	2.9	20	4.9	12	3.0	20	4.9	12	2.9	20	4.9	12	3.0	20	4.9
13	3.0	22	5.0	13	3.0	22	5.1	13	3.0	21	5.1	13	3.0	22	5.1
14	3.1	23	5.1	14	3.1	23	5.2	14	3.1	23	5.2	14	3.1	23	5.2
15	3.2	25	5.3	15	3.2	25	5.3	15	3.2	25	5.3	15	3.2	25	5.3
16	3.2	27	5.4	16	3.3	27	5.4	16	3.3	27	5.4	16	3.3	27	5.5
17	3.3	28	5.5	17	3.3	28	5.5	17	3.3	28	5.5	17	3.3	29	5.6
18	3.3	30	5.6	18	3.4	30	5.6	18	3.4	30	5.6	18	3.4	30	5.6
19	3.4	31	5.7	19	3.4	32	5.7	19	3.4	32	5.7	19	3.4	32	5.7
20	3.4	33	5.7	20	3.5	33	5.8	20	3.5	34	5.8	20	3.5	34	5.8
21	3.5	34	5.8	21	3.5	35	5.9	21	3.5	35	5.9	21	3.5	36	5.9
22	3.5	36	5.9	22	3.6	37	5.9	22	3.5	37	5.9	22	3.6	37	5.9
23	3.5	37	5.9	23	3.6	38	6.0	23	3.6	39	6.0	23	3.6	39	6.0
24	3.6	39	6.0	24	3.6	40	6.0	24	3.6	40	6.0	24	3.6	41	6.0
25	3.6	40	6.0	25	3.6	42	6.1	25	3.6	42	6.1	25	3.6	42	6.1
26	3.6	42	6.0	26	3.7	43	6.1	26	3.7	44	6.1	26	3.7	44	6.1
27	3.6	44	6.0	27	3.7	45	6.1	27	3.7	45	6.1	27	3.7	46	6.1
28	3.6	45	6.1	28	3.7	47	6.1	28	3.7	47	6.1	28	3.7	47	6.1
29	3.6	47	6.1	29	3.7	48	6.2	29	3.7	48	6.1	29	3.7	49	6.2
30	3.6	48	6.1	30	3.7	50	6.2	30	3.7	50	6.1	30	3.7	51	6.2

Table 6.7 Continued

Conditional Standard Errors of Measurement (CSEM) for Grade 7 Science by Test Form (Continued)

Form 236				Form 726				Form 589				Form 892			
Raw Score	Raw Score CSEM	Reported Scaled Score	Scaled Score CSEM	Raw Score	Raw Score CSEM	Reported Scaled Score	Scaled Score CSEM	Raw Score	Raw Score CSEM	Reported Scaled Score	Scaled Score CSEM	Raw Score	Raw Score CSEM	Reported Scaled Score	Scaled Score CSEM
31	3.6	50	6.1	31	3.7	52	6.2	31	3.7	52	6.1	31	3.7	52	6.2
32	3.6	52	6.1	32	3.7	53	6.1	32	3.7	53	6.1	32	3.7	54	6.1
33	3.6	53	6.0	33	3.7	55	6.1	33	3.7	55	6.1	33	3.7	56	6.1
34	3.6	55	6.0	34	3.7	57	6.1	34	3.7	56	6.1	34	3.7	57	6.1
35	3.6	57	6.0	35	3.6	58	6.1	35	3.6	58	6.1	35	3.6	59	6.1
36	3.6	58	6.0	36	3.6	60	6.0	36	3.6	60	6.0	36	3.6	61	6.0
37	3.5	60	5.9	37	3.6	62	6.0	37	3.6	61	6.0	37	3.6	62	6.0
38	3.5	62	5.9	38	3.6	63	5.9	38	3.5	63	5.9	38	3.6	64	5.9
39	3.5	63	5.8	39	3.5	65	5.9	39	3.5	65	5.9	39	3.5	66	5.9
40	3.4	65	5.7	40	3.5	67	5.8	40	3.5	66	5.8	40	3.5	67	5.8
41	3.4	67	5.7	41	3.4	68	5.7	41	3.4	68	5.7	41	3.4	69	5.7
42	3.3	69	5.6	42	3.4	70	5.6	42	3.4	70	5.6	42	3.4	71	5.6
43	3.3	70	5.5	43	3.3	72	5.5	43	3.3	71	5.5	43	3.3	72	5.6
44	3.2	72	5.4	44	3.3	73	5.4	44	3.3	73	5.4	44	3.3	74	5.5
45	3.2	74	5.3	45	3.2	75	5.3	45	3.2	75	5.3	45	3.2	76	5.3
46	3.1	76	5.1	46	3.1	77	5.2	46	3.1	77	5.2	46	3.1	77	5.2
47	3.0	78	5.0	47	3.0	78	5.1	47	3.0	78	5.1	47	3.0	79	5.1
48	2.9	79	4.9	48	3.0	80	4.9	48	2.9	80	4.9	48	3.0	81	4.9
49	2.8	81	4.7	49	2.9	82	4.8	49	2.9	82	4.8	49	2.9	82	4.8
50	2.7	83	4.5	50	2.8	83	4.6	50	2.7	84	4.6	50	2.8	84	4.6
51	2.6	85	4.3	51	2.6	85	4.4	51	2.6	85	4.4	51	2.6	85	4.4
52	2.5	87	4.1	52	2.5	87	4.2	52	2.5	87	4.2	52	2.5	87	4.2
53	2.3	88	3.9	53	2.4	88	4.0	53	2.4	89	3.9	53	2.4	88	4.0
54	2.2	90	3.6	54	2.2	90	3.7	54	2.2	91	3.7	54	2.2	90	3.7
55	2.0	92	3.4	55	2.0	92	3.4	55	2.0	92	3.4	55	2.0	92	3.4
56	1.8	94	3.0	56	1.8	93	3.1	56	1.8	94	3.1	56	1.8	93	3.1
57	1.6	95	2.6	57	1.6	95	2.7	57	1.6	96	2.7	57	1.6	95	2.7
58	1.3	97	2.2	58	1.3	97	2.2	58	1.3	97	2.2	58	1.3	97	2.2
59	0.9	99	1.6	59	0.9	98	1.6	59	0.9	99	1.6	59	0.9	98	1.6
60	0.0	100	0.0	60	0.0	100	0.0	60	0.0	100	0.0	60	0.0	100	0.0

Table 6.8

Conditional Standard Errors of Measurement (CSEM) for Grade 11 Life Science by Test Form

Form 249				Form 118				Form 369				Form 875			
Raw Score	Raw Score CSEM	Reported Scaled Score	Scaled Score CSEM	Raw Score	Raw Score CSEM	Reported Scaled Score	Scaled Score CSEM	Raw Score	Raw Score CSEM	Reported Scaled Score	Scaled Score CSEM	Raw Score	Raw Score CSEM	Reported Scaled Score	Scaled Score CSEM
0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0
1	0.9	3	3.1	1	1.0	3	3.2	1	0.9	4	3.1	1	0.9	3	3.2
2	1.3	7	4.4	2	1.3	7	4.5	2	1.3	8	4.4	2	1.3	7	4.4
3	1.6	10	5.3	3	1.6	10	5.4	3	1.6	11	5.3	3	1.6	10	5.3
4	1.8	13	6.0	4	1.8	13	6.1	4	1.8	15	6.0	4	1.8	13	6.0
5	2.0	17	6.5	5	2.0	17	6.7	5	2.0	19	6.5	5	2.0	17	6.6
6	2.1	20	7.0	6	2.2	20	7.2	6	2.1	22	7.0	6	2.1	20	7.0
7	2.2	23	7.4	7	2.3	23	7.6	7	2.2	26	7.4	7	2.2	24	7.4
8	2.3	27	7.8	8	2.4	27	7.9	8	2.3	29	7.8	8	2.3	27	7.8
9	2.4	30	8.0	9	2.5	30	8.2	9	2.4	32	8.0	9	2.4	30	8.1
10	2.5	33	8.3	10	2.5	33	8.4	10	2.5	36	8.3	10	2.5	34	8.3
11	2.5	37	8.4	11	2.6	37	8.6	11	2.5	39	8.5	11	2.5	37	8.5
12	2.6	40	8.6	12	2.6	40	8.8	12	2.6	42	8.6	12	2.6	40	8.6
13	2.6	44	8.7	13	2.7	43	8.9	13	2.6	46	8.7	13	2.6	43	8.7
14	2.6	47	8.7	14	2.7	47	8.9	14	2.6	49	8.8	14	2.6	47	8.8
15	2.6	51	8.8	15	2.7	50	9.0	15	2.6	52	8.8	15	2.6	50	8.8
16	2.6	55	8.7	16	2.7	53	8.9	16	2.6	55	8.8	16	2.6	54	8.8
17	2.6	59	8.7	17	2.7	57	8.9	17	2.6	59	8.7	17	2.6	57	8.7
18	2.6	63	8.6	18	2.6	60	8.8	18	2.6	62	8.6	18	2.6	61	8.6
19	2.5	66	8.4	19	2.6	63	8.6	19	2.5	65	8.5	19	2.5	64	8.5
20	2.5	70	8.3	20	2.5	67	8.4	20	2.5	69	8.3	20	2.5	68	8.3
21	2.4	74	8.0	21	2.5	70	8.2	21	2.4	72	8.0	21	2.4	72	8.1
22	2.3	77	7.8	22	2.4	73	7.9	22	2.3	76	7.8	22	2.3	75	7.8
23	2.2	81	7.4	23	2.3	77	7.6	23	2.2	79	7.4	23	2.2	79	7.4
24	2.1	84	7.0	24	2.2	80	7.2	24	2.1	82	7.0	24	2.1	82	7.0
25	2.0	87	6.5	25	2.0	83	6.7	25	2.0	86	6.5	25	2.0	86	6.6
26	1.8	90	6.0	26	1.8	87	6.1	26	1.8	89	6.0	26	1.8	89	6.0
27	1.6	93	5.3	27	1.6	90	5.4	27	1.6	92	5.3	27	1.6	93	5.3
28	1.3	96	4.4	28	1.3	93	4.5	28	1.3	95	4.4	28	1.3	95	4.4
29	0.9	98	3.1	29	1.0	97	3.2	29	0.9	98	3.1	29	0.9	98	3.2
30	0.0	100	0.0	30	0.0	100	0.0	30	0.0	100	0.0	30	0.0	100	0.0

Table 6.9

Conditional Standard Errors of Measurement (CSEM) for Grade 11 Physical Science by Test Form

Form 449				Form 657				Form 714				Form 916			
Raw Score	Raw Score CSEM	Reported Scaled Score	Scaled Score CSEM	Raw Score	Raw Score CSEM	Reported Scaled Score	Scaled Score CSEM	Raw Score	Raw Score CSEM	Reported Scaled Score	Scaled Score CSEM	Raw Score	Raw Score CSEM	Reported Scaled Score	Scaled Score CSEM
0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0
1	0.9	3	3.1	1	0.9	3	3.1	1	0.9	3	3.2	1	0.9	3	3.2
2	1.3	6	4.3	2	1.3	6	4.4	2	1.3	6	4.4	2	1.3	7	4.4
3	1.6	9	5.2	3	1.6	9	5.3	3	1.6	9	5.3	3	1.6	10	5.3
4	1.8	12	5.9	4	1.8	12	6.0	4	1.8	12	6.0	4	1.8	13	6.0
5	1.9	15	6.4	5	2.0	16	6.5	5	2.0	15	6.6	5	2.0	17	6.5
6	2.1	18	6.9	6	2.1	19	7.0	6	2.1	18	7.0	6	2.1	20	7.0
7	2.2	21	7.3	7	2.2	23	7.4	7	2.2	21	7.4	7	2.2	23	7.4
8	2.3	25	7.6	8	2.3	26	7.7	8	2.3	25	7.8	8	2.3	27	7.8
9	2.4	29	7.9	9	2.4	30	8.0	9	2.4	29	8.1	9	2.4	30	8.0
10	2.4	32	8.1	10	2.5	33	8.3	10	2.5	33	8.3	10	2.5	33	8.3
11	2.5	36	8.3	11	2.5	36	8.4	11	2.5	36	8.5	11	2.5	37	8.5
12	2.5	39	8.4	12	2.6	40	8.6	12	2.6	40	8.6	12	2.6	40	8.6
13	2.6	43	8.5	13	2.6	43	8.7	13	2.6	44	8.7	13	2.6	43	8.7
14	2.6	47	8.6	14	2.6	47	8.7	14	2.6	48	8.8	14	2.6	47	8.8
15	2.6	51	8.6	15	2.6	50	8.8	15	2.6	52	8.8	15	2.6	50	8.8
16	2.6	55	8.6	16	2.6	53	8.7	16	2.6	56	8.8	16	2.6	53	8.8
17	2.6	60	8.5	17	2.6	57	8.7	17	2.6	60	8.7	17	2.6	57	8.7
18	2.5	64	8.4	18	2.6	60	8.6	18	2.6	63	8.6	18	2.6	60	8.6
19	2.5	67	8.3	19	2.5	63	8.4	19	2.5	67	8.5	19	2.5	63	8.5
20	2.4	71	8.1	20	2.5	66	8.3	20	2.5	70	8.3	20	2.5	67	8.3
21	2.4	75	7.9	21	2.4	69	8.0	21	2.4	73	8.1	21	2.4	70	8.0
22	2.3	79	7.6	22	2.3	72	7.7	22	2.3	77	7.8	22	2.3	73	7.8
23	2.2	82	7.3	23	2.2	75	7.4	23	2.2	80	7.4	23	2.2	77	7.4
24	2.1	85	6.9	24	2.1	78	7.0	24	2.1	83	7.0	24	2.1	80	7.0
25	1.9	88	6.4	25	2.0	81	6.5	25	2.0	87	6.6	25	2.0	83	6.5
26	1.8	91	5.9	26	1.8	84	6.0	26	1.8	89	6.0	26	1.8	87	6.0
27	1.6	93	5.2	27	1.6	87	5.3	27	1.6	92	5.3	27	1.6	90	5.3
28	1.3	96	4.3	28	1.3	90	4.4	28	1.3	95	4.4	28	1.3	93	4.4
29	0.9	98	3.1	29	0.9	93	3.1	29	0.9	98	3.2	29	0.9	97	3.2
30	0.0	100	0.0	30	0.0	99	0.0	30	0.0	100	0.0	30	0.0	100	0.0

Figure 6.1. Conditional Standard Errors of Measurement (CSEM) for grade 4 science by test form

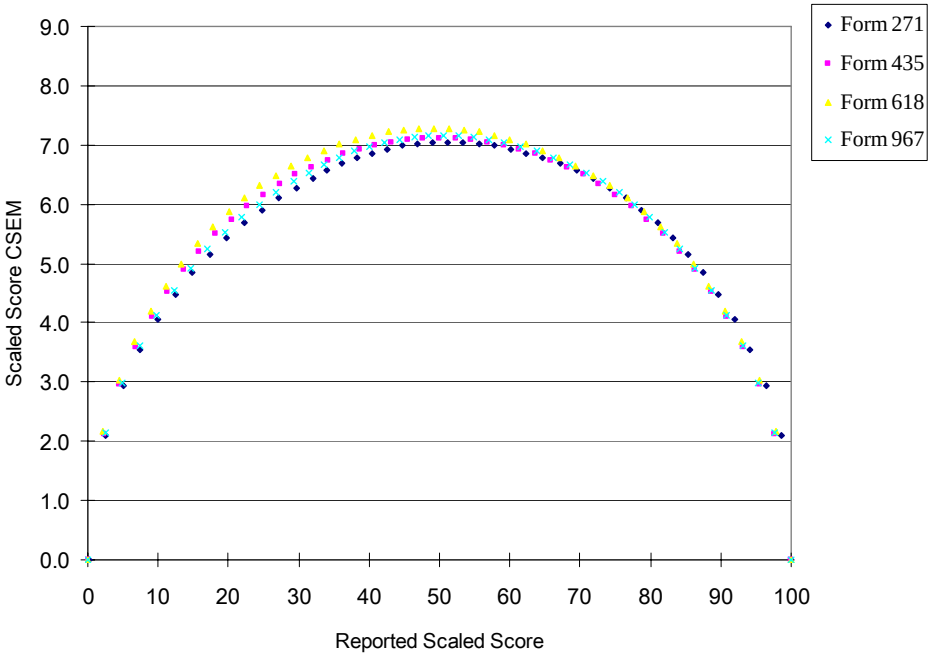


Figure 6.2. Conditional Standard Errors of Measurement (CSEM) for grade 7 science by test form

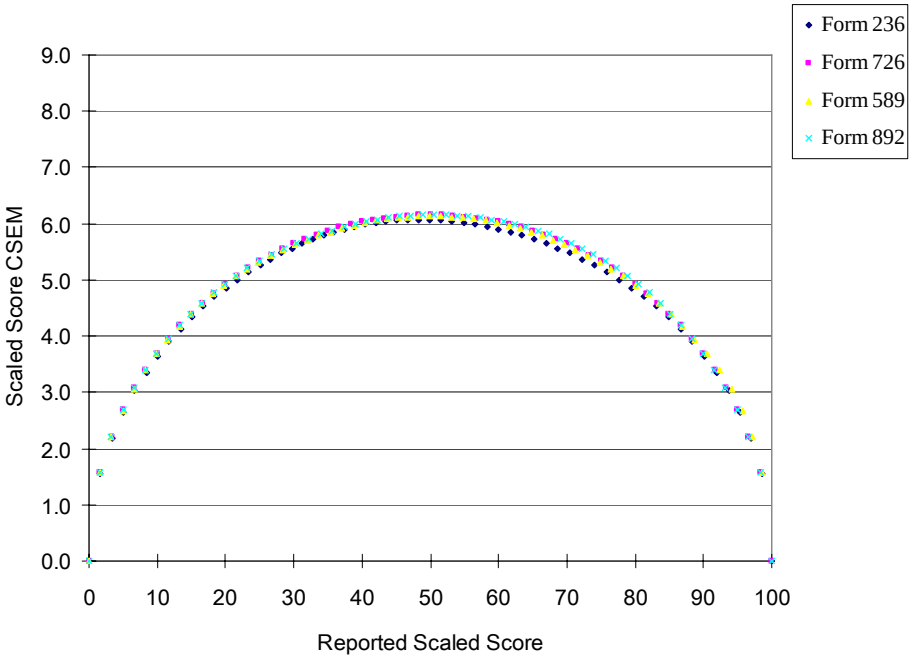


Figure 6.3. Conditional Standard Errors of Measurement (CSEM) for grade 11 life science by test form

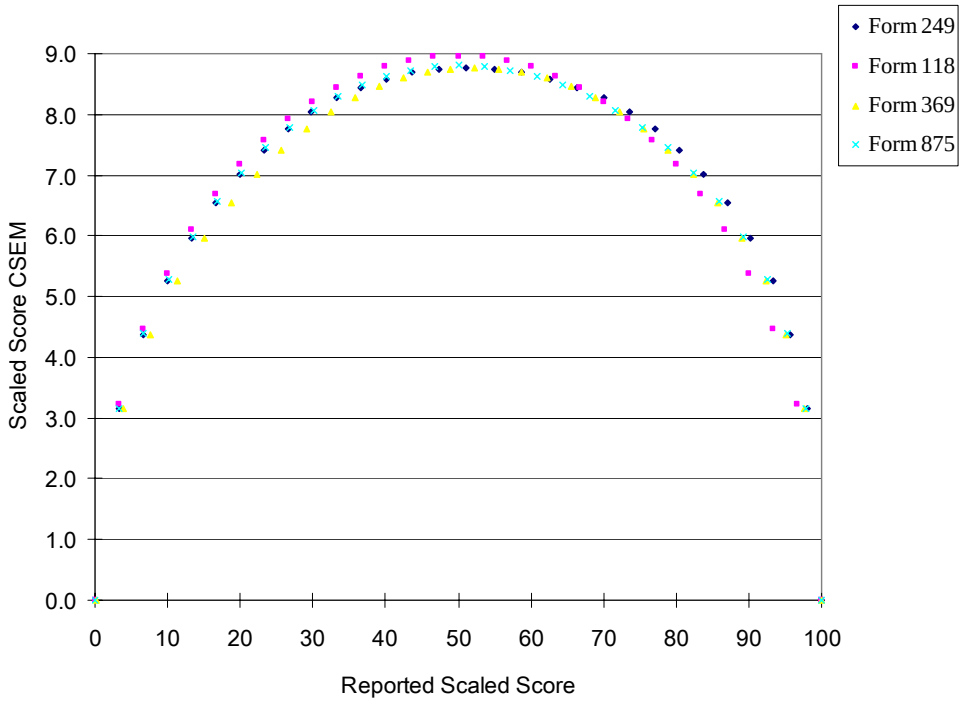
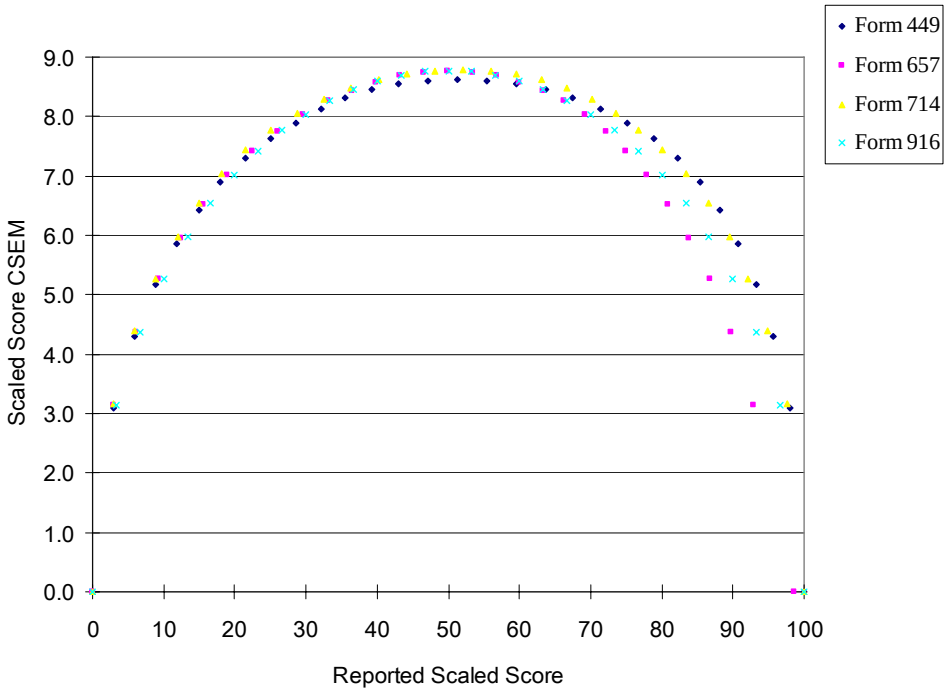


Figure 6.4. Conditional Standard Errors of Measurement (CSEM) for grade 11 physical science by test form



Section 7

VALIDITY

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

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

Validity Evidence:
Correlations among Science Sub-domain Scores at Benchmark Level

To be an effective science test, one must be able to demonstrate that the test, in fact, measures a student's knowledge of science uncontaminated by other constructs. Furthermore, if the test is measuring the curriculum as taught, one would expect a strong correlation among the sub-domains assessed. This study examines those correlations to support that the measured construct is science.

Procedure

Sample

The correlational study among sub-domain scores was based on samples of students who were administered the Kansas general assessments in science via the computer or P&P. Although there were four parallel test forms for each grade, correlational study among sub-domain scores was only conducted on the base form. For grade 11, correlations among sub-domain scores were calculated in the combined test forms of life science and physical science. Table 7.1 presents the base form number at each grade, the number of items in the base form, and the sample sizes used in the analyses.

Table 7.1
Test Length and Sample Size and Descriptive Statistics for Each Form in Science

Grade	Form	Number of Items	Sample Size
3	435	44	6,687
7	726	60	15,069
11	118 (Life)	30	8,446
11	916 (Physical)	30	

Method

In the Kansas Science Assessments, there were four benchmarks on each test form per grade. Each benchmark tests a specific content area, which include scientific processes and connections, physical science, life science, and earth/space science in benchmarks 1, 2, 3, and 4, respectively. Table 7.2 below displays the number of items per benchmark at each grade level. Pearson Product-Moment Correlations were calculated among sub-domain scores at the benchmark level.

Table 7.2
Number of Items per Benchmark by Grade Level

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

Results

Correlations among Sub-domain Scores

The correlations between sub-domain scores were calculated for the base forms across the grades. The results are presented in Tables 7.3 – 7.5. All correlations are positive and statistically significant at the 0.01 level. The average correlation ranges from 0.45 to 0.65 in grade 4, 0.58 to 0.67 in grade 7, and 0.47 to 0.64 in grade 11. There is no clear increasing or decreasing trend across grades. The greatest correlation appears between benchmarks 1 (scientific processes and connection, physical science) and 2 (physical science) in grades 4 (r = 0.65) and 7 (r = 0.67), and between benchmarks 2 (physical science) and 3 (life science) in grade 11 (r = 0.64).

Table 7.3
Correlations among Sub-domain Scores at Benchmark Level for Grade 4 Science (n = 6667)

		Scientific Processes & Connections	Physical Science	Life Science	Earth/Space Science
Scientific Processes & Connections	Pearson Correlation				
Physical Science	Pearson Correlation	.646**			
Life Science	Pearson Correlation	.448**	.469**		
Earth/Space	Pearson Correlation	.569**	.617**	.453**	

**. Correlation is significant at the 0.01 level (2-tailed).

Table 7.4

Correlations among Sub-domain Scores at Benchmark Level for Grade 7 Science (n = 15,069)

		Scientific Processes & Connections	Physical Science	Life Science	Earth/Space Science
Scientific Processes & Connections	Pearson Correlation				
Physical Science	Pearson Correlation	.669**			
Life Science	Pearson Correlation	.650**	.653**		
Earth/Space Science	Pearson Correlation	.579**	.581**	.603**	

**. Correlation is significant at the 0.01 level (2-tailed).

Table 7.5

Correlations among Sub-domain Scores at Benchmark Level for Grade 11 Science (n = 8446)

		Scientific Processes & Connections	Physical Science	Life Science	Earth/Space Science
Scientific Processes & Connections	Pearson Correlation				
Physical Science	Pearson Correlation	.576**			
Life Science	Pearson Correlation	.614**	.644**		
Earth/Space Science	Pearson Correlation	.472**	.493**	.548**	

**. Correlation is significant at the 0.01 level (2-tailed).

Validity Evidence:
Intercorrelations across Content Area Tests

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

An evaluation of the Kansas general assessment science scores' criterion validity includes assessing the relationships of total scores in the areas of science with mathematics and reading for grades 4 and 7 and the relationships of life science, physical science, and combined sciences (both life science and physical science) with mathematics and reading for grade 11.

The evaluation of the strength of test relationships was based on samples of students who were administered any form of the Kansas general assessments in a given subject administered either via the computer or paper-and-pencil. In 2008, parallel test forms of the Kansas general assessment were constructed at grades 4, 7, and 11 for both mathematics (N=35,166, N=34,710, N=40,967, respectively) and reading (N=35,035, N=34,728, N=36,484, respectively), at grades 4 and 7 for science (N=35,382, N=34,756, respectively), and at grade 11 for life science (N=50,907) and physical science (N=43,083). For grades 4 and 7 science, there were four test forms per grade, and for grade 11, there were 4 forms for life science and four forms for physical science. There were four test forms for mathematics and reading for grades 4 and 7 while only three test forms were available for grade 11 for mathematics and reading.

Pearson product-moment correlations were calculated using the total score for science, mathematics, and reading for grades 4 and 7, and correlating the total scores for life science, physical science, and combined science with the total scores for mathematics and reading for grade 11.

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

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

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

Results from the validity assessments are displayed in Tables 7.6, 7.7, and 7.8, which detail the intercorrelations for grades 4, 7, and 11, respectively, including the observed correlations and the correlations corrected for attenuation.

Intercorrelations for grade 4 ranged from .697-.727 (observed) and .796-.840 (corrected), for grade 7 from .712-.745 (observed) and .784-.827 (corrected), and for grade 11 from .645-.711 (observed) and .722-.847 (corrected).

For each grade, the science scores (including life science, physical science, and combined sciences for grade 11) have higher observed and corrected correlations with reading scores than with mathematics scores. This is somewhat surprising and suggests that the science test may have a fairly heavy reading load. If this is so, for some students, reading ability might interfere with their ability to demonstrate their science knowledge. Further study of this issue is desirable.

Table 7.6

Intercorrelations for Grade 4

Grade 4	Observed Correlation		Correlation After Correction for Attenuation	
	Mathematics	**Reading**	**Mathematics**	**Reading**
Science	0.697	0.727	0.796	0.840
Mathematics		0.722		0.801

Table 7.7
Intercorrelations for Grade 7

Grade 7	Observed Correlation		Correlation After Correction for Attenuation	
	Mathematics	**Reading**	**Mathematics**	**Reading**
Science	0.712	0.745	0.784	0.827
Mathematics		0.728		0.784

Table 7.8
Intercorrelations for Grade 11

Grade 11	Observed Correlation		Correlation After Correction for Attenuation	
	Mathematics	**Reading**	**Mathematics**	**Reading**
Life Science	0.656	0.664	0.775	0.789
Physical Science	0.645	0.648	0.765	0.773
Combined Sciences	0.687	0.711	0.813	0.847
Mathematics		0.677		0.722

Section 8

PAPER AND PENCIL VERSUS COMPUTER ADMINISTERED TEST COMPARABILITY STUDIES FOR SCIENCE

Introduction

The Kansas Assessment Program first implemented computerized (online) delivery of state-mandated tests starting with a limited pilot for the grade 7 mathematics test in spring of 2003 (approximately 3,000 students tested). Since that time, computerized testing in Kansas has grown under a volunteer dual paper and pencil (P&P) or computerized administration program so that in the spring of 2008, an overwhelming majority of the students were tested online, approximately 78 percent in reading, 80 percent in mathematics, 82 percent in science, and 81 percent in history/government. As indicated by these participation rates, the dominant administration test mode is via computer.

Whenever tests that are administered under both testing modes co-exist in a testing program, score comparability between computerized and P&P tests becomes an issue. The Center for Educational Testing and Evaluation at the University of Kansas has been conducting studies addressing the comparability issues since the initial implementation of volunteer online testing in Kansas (Glasnapp, Poggio, Yang, and Poggio (2004); Poggio, Glasnapp, Yang, and Poggio (2005); Poggio, Glasnapp, Yang, Beauchamp, and Dunham (2005); Yang, Glasnapp, and Poggio (2006)). In these latter studies, the conditions and constraints of the testing program when the studies were initiated necessitated that a "double testing" design be put in place so that the individual students served as their own controls in the repeated measures design. In this design, controlling for order of administration is a potential problem, and while best controlled through random assignment, it typically is not practical to implement such a requirement in a state-mandated testing program.

There is no doubt that the best, failsafe design for studying comparability issues is to implement a true randomized experimental design with random selection and assignment of student to test mode. It would be the design of choice. However, this design requires that schools that do not volunteer students for online testing would have to test selected students online. This would be difficult to implement because effort and advance preparation for a school are required to tool up for an online assessment implementation. Similarly, schools that volunteer students for online testing would have to test selected students P&P. This would not be as difficult to implement, but it does require participation of some students in the P&P mode which may create a problem for school administrators, teachers, and parents. The Kansas State Department of Education made a decision in 2006 that implementing such a design within the context of a federal- and state-mandated testing to study comparability was impractical and unwise. Thus, other designs were explored.

For the study of test mode administration comparability in reading and mathematics, the tested grade levels (3 through 8 mandated yearly) provide for the existence of longitudinal data such that a "matched students" quasi-experimental design can be implemented, using prior year achievement scores as matching control variables or covariates to control for potential prior achievement differences in the volunteer computer based testing (CBT) group and the selected P&P comparison group along with other matching demographic covariates. (See the similar comparability report for reading and mathematics results submitted for peer review, October 2008.) However, the NCLB mandate does not require the testing of adjacent grade students in science. In Kansas, grades 4, 7, and high school are the tested grade levels, and 2008 was the initial year for science testing. Thus, longitudinal data do not exist whereby prior year science scores might be used as a matching variable to control for achievement or ability differences in the volunteer CBT and P&P populations. Given the data available, it was decided the best analysis design to implement was still a "matched groups" quasi-experimental design, but the matching of students was not as precise as in the design implemented to study test mode comparability in reading and mathematics. Rather than match CBT and P&P student pairs specifically on prior achievement scores, proportional selection/matching at the group level was done using racial background and information on free/reduced lunch as the matching variables. These two variables were selected to serve as proxies to control for volunteer CBT and P&P group differences in achievement/ability as both variables are moderately correlated with achievement scores. This design was implemented while being cognizant that the matched groups design has its own inherent weakness (e.g., no control for differential instructional or other interventions impacting outcome scores).

The current report presents and discusses the results from these studies within the context of the Kansas program. Data sets were configured and analyses were conducted that address both construct and score distribution equivalency between P&P and computer administered tests in science for test forms at the grade levels tested in science. At grades 4 and 7, only one test form was administered in both the CBT and P&P mode. At the high school level, the test was divided into two part tests (life science and physical science) with one form of each part given in both the CBT and P&P mode.

In the presentation that follows, a general description of the "matched groups" design data sets is first presented. This is followed by separate descriptions of the methodology implemented and results of analyses addressing, first, the construct equivalency of test scores between P&P and computer administered tests and second, the score distribution equivalency between test modes.

Matched Groups Design Data Sets

Four "matched groups" data sets were configured, one at grade 4, one at grade 7, and two at the high school level (one for the life science part test form and one for the physical science part test form). Only one test form was available for administration in the P&P mode at the three grade levels. Based on school building choice, students were either administered the tests in the CBT or the P&P mode. As mentioned previously, approximately 20 percent of the students in the state received a P&P test administration. For the CBT test administration, four forms of the test were available for administration at each grade and were randomly assigned to students as part of the random groups equating design, thus approximately 20 percent of the students in the state received the same form via CBT as available for P&P test administration.

It was these two "same form" sample data sets, one sample administered the form via the CBT mode and the other administered the test form via the P&P mode, that served as the basis for configuring the matched groups data sets used in the test mode comparability analyses. For each grade level test form, test mode groups were matched proportionally on racial background and information on free/reduced lunch through random deletion of cases within the larger of the race by lunch group for the two mode groups. Table 8.1 provides a summary for each grade level of the number of students taking the test form in each mode (CBT or P&P) prior to and after the matched groups sampling.

Table 8.1
Number of CBT and P&P Student Test Takers Prior to and After the Matched Groups Sampling

Grade Level	Prior to Sampling		After Sampling	
	P&P	CBT	P&P	CBT
Grade 4	6461	6868	5536	5536
Grade 7	7115	6742	5714	5714
H.S. Life	5481	6553	5026	5026
H.S. Physical	5442	6555	4967	4967

Evaluating Construct Equivalency across Test Mode

Construct equivalency across test mode is the foremost concern in score comparability studies. If a test, when administered in different testing modes, measures different constructs, it is futile to conduct any further kind of score comparability analysis. Therefore, it is important to investigate that the testing mode (i.e., P&P vs. CBT) does not change the underlying construct to be measured.

Two data analysis approaches were taken to investigate construct equivalency. One approach used a multi-group Confirmatory Factor Analysis (CFA) in the context of structural equation modeling - SEM (Byrne, 2006) to examine the equivalency of the structure of the underlying construct (e.g., structural relations among distinctive components of the construct, etc.) measured by the tests across different test delivery modes, P&P versus CBT. The second approach examined differential performance at the item level using the Mantel-Haenszel Differential Item Functioning (DIF) procedure.

Data Sets Examined

Single matched groups P&P and CBT test taker data files containing 2008 scored item response level data were configured to investigate construct equivalency between P&P and CBT test takers. Group matching was done on 2008 racial and free/reduced lunch information as described previously.

Multi-Group Confirmatory Factor Analysis (CFA)

The first data analysis approach taken to investigate construct equivalency used a multi-group Confirmatory Factor Analysis (CFA) in the context of structural equation modeling - SEM (Byrne, 2006) to examine the equivalency of the structure of the underlying construct (e.g., structural relations among distinctive components of the construct, etc.) measured by the tests across different test delivery modes, P&P versus CBT. For the analyses, the items were grouped in facet representative parcels. The use of item parcels has advantages over the use of individual items as indicators for a variety of reasons, including keeping the ratio for manifest indicators to latent constructs manageable and reducing the number of free parameters in the model to decrease sample size requirements (Hall, Snell, & Singer, 1999). The parceling was accomplished by grouping items that corresponded to similar indicators within curriculum benchmark and standards and then computing an average for each parcel.

Within the domain of science, the tests target indicators across seven content standard areas. However, within each test, there were not a sufficient number of items to create at least two parcel sets of items within each content standard area. Thus, depending on grade level and test form, parcels were formed consisting of 4 to 6 items measuring similar indicators or benchmarks and grouped under factors to provide the structure for the analysis. At grade 4, 10 parcels were created and grouped under three factors. At grade 7, 12 parcels were formed and grouped under four factors. For the high school life science test form, six parcels were formed, all grouped under a single factor. For the high school physical science test form, seven parcels were formed and grouped under three factors. Since the constructs defined by the parcel sets under each factor are expected to be highly correlated, a second order multiple group CFA was considered appropriate for those instances when more than one factor was in the model. Table 8.2 displays the number of parcels and the number of items per parcel in science for each form of the test.

For each latent construct (Factor 1, for example), the loading on the first of the corresponding parcels was fixed to 1 in order to set a scale. The remaining loadings were estimated and then constrained to be equal across the two groups (P&P vs. CBT) in order to test the model equivalence.

The multiple group CFA analysis was conducted in EQS (Bentler, 1995). The Mardia's coefficient and its normalization were used to judge the normality of the data. Accordingly, in all cases, the maximum likelihood solution (normal distribution theory) is reported. Model fit indices were then evaluated in terms of the cutoff values proposed by Hu and Bentler (1999). While the overall model chi-square statistic was statistically significant in all cases when comparing the two groups, the chi-square statistic is known to be extremely sensitive to sample size in these analyses and thus, statistical significance testing is not typically used as a criterion for drawing conclusions about group differences. Rather, the Comparative Fit Index (CFI) is the criterion typically examined with values of 0.95 or above indicative of equivalence (Hu and Bentler, 1999). For all analyses conducted, the CFI values demonstrate that the constraints hold across groups as all are greater than the standard 0.95, thus supporting the structural similarity of test forms for every grade level test form across the P&P and CBT administration samples. Additionally, the root mean-square errors of approximation (RMSEA) were all less than 0.05, which also indicates that the constraints hold. Furthermore, the RMSEA 90% confidence interval also supports the equality of loadings across the P&P and CBT groups as all intervals are within the standard acceptable range (0.00, 0.08).

Table 8.2
Number of Parcels and Items per Parcel for each Science Grade Level Test Form

Parcels	**Grade 4**	**Grade 7**	**H.S. Life**	**H.S. Physical**
Factor 1				
A1	4	6	6	4
A2	4	4	6	4
A3	6	6	4	4
A4	4		4	
A5			6	
A6			4	
Factor 2				
B1	4	6		4
B2	4	6		6
B3	4	6		
B4	4			
Factor 3				
C1	4	4		4
C2	6	4		4
C3		4		
C4		4		
Factor 4				
D1		4		
D2		6		

Table 8.3 displays the model fit indices in science for each form of the tests. Based on the multi-group CFA fixed factor loadings across groups and results, there is strong evidence supporting the construct equivalence of these versions of the tests.

Table 8.3
CFA Model Fit Indices for Each Grade Level Test Form

Grade	Chi-Square	CFI	RMSEA	90% RMSEA CI	N in each group
4	347.952**	0.99	0.026	(0.024, 0.029)	5536
7	448.736**	0.992	0.023	(0.021, 0.025)	5714
H. S. Life	79.959 **	0.995	0.022	(0.017, 0.028)	5026
H. S. Physical	79.363**	0.996	0.02	(0.015, 0.025)	4967

** Significant at 0.001, normal (Maximum Likelihood) indices are reported.

Differential Item Functioning Analyses (DIF)

The procedure used to explore differential item functioning across modes was the Mantel-Haenszel (MH) technique. The criteria used in these analyses were (1) the absolute delta value larger than 1.5 and (2) the absolute delta value statistically significantly larger than 1.0. Using a significant level of 0.01, the second criterion is equivalent to a MH chi-squared value of 12.7866. In the analyses, the CBT group served as the focal group. Items with negative delta values created a disadvantage for the CBT group while items with positive values created an advantage for the CBT group in comparison to the P&P group.

In the DIF analyses for science items, 164 items across the grade level test forms were evaluated with **none** of the items (0.0%) flagged for potential DIF. For the science items, there appeared to be no differential item functioning across a wide variety of item types and formats, thus adding support to the construct equivalency of the science test forms across the two modes of administration.

Score Distribution Equivalency

Table 8.4 provides information on the score distributional characteristics for the grade level samples of CBT and P&P test takers. Presented are the mean percent correct scores, standard deviations and percent classified as Proficient (above the state's cut-score that defines "Meets Standard") for the two matched CBT and P&P groups. For the grade 4 test form, the mean percent correct score difference between the two groups is 1.72 percent correct units with the CBT group scoring higher. With 44 items on the test, this difference is equivalent to the CBT group receiving a score that is, on average, 0.76 of 1 item correct more than the P&P group. The standard deviations are approximately equal. While the mean and standard deviation provide information on the similarities of the location and spread of the score distribution, another important index in the NCLB mandate is the proportion or percentage of students judged to be proficient based on cut scores established by the state. These proficiency percentages are affected not only by the location of the distribution on the score scale continuum (mean) and the spread of

the scores (standard deviation), but also are impacted by the shape of the distribution. For grade 4 scores, the percentage of students classified as meeting the state's definition of being Proficient differed by 1.5 percentage points with 90.0 percent of the P&P test takers judged as Proficient and 91.5 percent of the CBT test takers judged as Proficient.

For the grade 7 test form, the mean percent correct score difference between the two groups is 1.30 percent correct units with the CBT group scoring higher. With 60 items on the test, this difference is equivalent to the CBT group receiving a score that is, on average, 0.78 of 1 item correct more than the P&P group. The standard deviations are approximately equal. For grade 7 scores, the percentage of students classified as meeting the state's definition of being Proficient differed by 1.7 percentage points with 81.8 percent of the P&P test takers judged as Proficient and 83.5 percent of the CBT test takers judged as Proficient.

On the high school test forms, the results were reversed with P&P test takers scoring slightly higher than CBT test takers on average, 0.91 percent correct units on the life science form and 1.36 percent correct units on the physical science test forms. With 30 items on each test, these differences are equivalent to the P&P group receiving a score that is, on average, 0.21 of 1 item correct more for life science and 0.41 of 1 item correct more for physical science than the CBT group. Applying an arbitrary cut-score of 40 percent correct or higher (the cut-score for the combined two parts) to each part, the percent Proficient differences are 0.4 of a percent for life science and 1.3 percent for physical science.

Table 8.4
Grade Level Means, Standard Deviations, and Percentage Proficient for Matched Groups of Students Taking Tests in the P&P or CTB Mode

Grade Test Form	Mode of Testing	Sample Size	Mean	Standard Deviation	Percent Proficient
Grade 4	P&P	5536	72.33	15.19	90.0
	KCA	5536	74.05	15.00	91.5
Grade 7	P&P	5714	62.35	16.77	81.8
	KCA	5714	63.65	16.33	83.5
H.S. Life Science	P&P	5026	58.33	17.56	85.2
	KCA	5026	57.42	17.12	84.8
H.S. Physical Science	P&P	4967	56.06	17.66	82.1
	KCA	4967	54.70	17.07	80.8

Based on the evidence examining the differences in score distributional characteristics for the matched groups P&P and CBT test taker samples, the differences are judged to be slight. The data would not appear to provide sufficient evidence of differences of a magnitude that would warrant differential treatment of the data.

Summary Conclusions and Recommendations

The current set of analyses adds to the information base on CBT versus P&P comparability of test forms as delivered online by the KCA software and system as implemented in Kansas. The evidence is consistent, supporting the equivalency of the two modes in delivering items in formats that do not differentially impact the underlying construct being measured. This support for construct equivalency was found across two data analysis approaches, one using a multi-group Confirmatory Factor Analysis (CFA) in the context of structural equation modeling - SEM (Byrne, 2006) to examine the equivalency of the structure of the underlying construct (e.g., structural relations among distinctive components of the construct, etc.) measured by the tests, and the other examined differential performance at the item level using the Mantel-Haenszel Differential Item Functioning (DIF) procedure. No differences were found in any of the analyses conducted when examining results against industry standard statistical criteria for detecting differences.

When examining the CBT and P&P samples for evidence of score distributional equivalence, differences were found as one would expect when comparing any two sample sets of data, but these differences are judged to be slight. In terms of the consequences of treating the test forms as equivalent, very slight differences were observed in the percent of students classified as being Proficient, the index of primary importance for the state in making adequate yearly progress under the NCLB mandate. These latter differences were 1.50 and 1.30 percentage points for the grade 4 and grade 7 students, respectively, with CBT students performing better. Similar results were found for the two high school tests in terms of the magnitude of the difference (0.4 of a percent for life science and 1.3 percent for physical science), but in the latter cases, the P&P samples performed at the higher level.

Based on the results of the analyses conducted, there does not appear to be sufficient evidence to suggest that the science tests as administered in Kansas are measuring meaningfully different constructs or result in score distributional differences that are practically meaningful such that scores from one or the other administration mode are in need of adjustment. The easiest way to lay the CBT and P&P comparability issue to rest is to mandate that the vast majority of test takers move to the online administration of the test. Kansas is already approaching that mandate under its voluntary approach with districts choosing to administer the tests online for approximately 80 percent of the students in the state. It is recommended that the state either mandate that all students be tested online except for students needing specific P&P accommodated forms of the test or that they systematically encourage and provide assistance to facilitate the transition to online testing by those districts and buildings that use P&P as the predominant mode of testing.

REFERENCES

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

Bentler, P. M. (1995). *EQS: Structural equations program manual.* Encino, CA: Multivariate Software.

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

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

Byrne, B. M. (2006). *Structural equation modeling with EQS: Basic concepts, applications, and programming.* Mahwah, NJ: Lawrence Erlbaum.

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

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

Glasnapp, D. R., Poggio, J. P., Yang, X., & Poggio, A. (2004). *Student attitudes and perceptions regarding computerized testing as a method for formal assessment*. Paper presented at the NCME annual meeting, San Diego, April.

Hal, R., Snell, F., & Singer, M. (1999). Item parceling strategies in SEM: Investigating the subtle effects of unmodeled secondary constructs. *Organizational Research Methods*, *3*, 233-256.

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

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

Hu, L. & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1-55.

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

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

Kolen, M. J. & Brennan, R. L. (2004). *Test Equating, Scaling, and Linking: Methods and Practices* (2nd ed.). New York: Springer-Verlag.

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

Lord, F. M. (1955). Estimating test reliability. *Educational and Psychological Measurement, 15,* 325-336.

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

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

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

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

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

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

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

Yang, X., Glasnapp, D. R., & Poggio, J. (2006). *Score Comparability between Computerized and Paper-and-Pencil Linear Tests*. Draft Report submitted to the Kansas State Department of Education, December.

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