

**Kansas Assessments  
in History and Government**

**2008**

**TECHNICAL MANUAL**

for the

**Kansas General Assessments**

*Prepared by:*

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

**Center for Educational Testing and Evaluation**  
The University of Kansas

**September 2008**

## Table of Contents

Purpose of the Technical Report.....	1
Introduction and Orientation.....	2
Test Development and Content Representation.....	4
Summary Statistics Spring 2008 Administration.....	7
Differential Item Functioning (DIF) Analyses .....	14
Test Equating .....	21
Equating Conversion Tables .....	35
Standard Setting.....	40
Reliability Analyses .....	53
Score Reliability.....	53
Classification Consistency .....	53
Conditional Standard Errors of Measurement .....	56
Validity .....	65
Correlations among Sub-domain Scores.....	67
Intercorrelations across Content Area Tests .....	69
Paper and Pencil versus Computer Administered Test Comparability.....	71
References.....	74
Appendix A Test Specifications .....	76

# **The Kansas Assessments in History and Government**

## **PURPOSE OF THE TECHNICAL REPORT**

The *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [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 History and Government Assessments for grades 6, 8, and high school. The information included in this report is intended for use by those who evaluate tests, interpret scores, or use test results in making educational decisions. It is assumed that the reader has some technical knowledge of test construction and measurement procedures.

Information is provided to address the technical quality of the assessments developed to measure history and government learning outcomes specific to the 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 equating analyses for history and government 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 following section presents evidence from a variety of validity studies, providing information on both internal and external sources of score validity. The final section outlines comparability between computerized testing and paper and pencil testing.

## Section 1

### INTRODUCTION AND ORIENTATION

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

- (1) provide aggregate state accountability and yearly progress information toward meeting the Kansas Curriculum Standards;
- (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 information, new Kansas History and Government Assessments were planned, 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 6, 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](http://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 history and government 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. To assure comparability of scores across the different forms of the history and government 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 30).

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 (NCLB) 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. While the History and Government tests are not part of federal accountability, they also have five performance categories to contribute coherently to the state's accountability system. 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 were offered under one of two modes: 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](http://www.kca.cete.us). Students who took the history and government tests online using KCA made up approximately 78% of tested sixth grade students, 76% of eighth grade students, and 85% of high school students. 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, and Poggio (2005) and Poggio, Glasnapp, Yang, Beauchamp, and 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 history and government in grades 6, 8, 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. A copy is also provided in Appendix A.

#### Item Type

The multiple choice item type is the only item type used on the Kansas General History and Government 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, a third party, to supply history and government 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 either accepted, edited, or rejected from its respective item pools.

## Item Delivery and Tryouts

All history and government 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.

In 2007, all Kansas schools were encouraged and invited to participate in pilot testing for the history and government assessments. Due to the large number of history and government items to be piloted, a fourth test section was added to state math and reading tests at all grade levels. For grades 6, 8, and high school, the fourth section consisted of history and government pilot items. All items in the history and government item pool were piloted. History and government pilot tests were administered using both the computer and P&P delivery modes. For the P&P pilot forms, items were randomly selected and assigned to forms. At each grade level, ten forms were created and administered via the P&P mode. The remaining items were also randomly selected and assigned to forms and were piloted via computer (KCA). At 6<sup>th</sup> grade, there were 18 forms; at 8<sup>th</sup> grade, 23 forms; and at high school, 24 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 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 500 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 the assessed grade levels, hundreds of items were piloted and subsequently evaluated by CETE test development staff using the classical item analysis procedures described above. To assist in the pilot item review process, a set of rules were adopted to assist in identifying items which function poorly (e.g., items that are too easy, too difficult, contain errors,

or have low or negative discrimination). The rules or criteria for identifying poorly functioning items were the following:

- $r_{pb} < 0.20$  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. Flagged items that aligned to an indicator with sufficient coverage for the construction of multiple test forms were rejected. 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 (i.e., 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 total items and items per indicator. In history and government, a sufficient number of items were available to build four operational forms at each of the three grade levels.

For the history and government 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 (descending p-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 History and Government Assessments in Spring 2008, analyses were conducted using classical and item response theory (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 and 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 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 History and Government Assessments were reported in September 2008.

### **Summary Statistics Spring 2008 Administration**

The summary statistics displayed in the Spring 2008 History and Government Administration tables were created using scores from students with an AYP status code equal to 1 and based on end-of-year data (AYP Status Codes for 2008 are defined below). The Differential Item Functioning Analyses, Test Equating, Standard Setting, Reliability Analyses, Validity Information and Comparability Study used all scores regardless of AYP status and were based on the end-of-testing window 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 the year. However, the scanned paper and pencil data typically require 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.

#### **2008 AYP Status Codes**

- 4 Incomplete test (fewer than one third of the items tried) so the student did not participate in testing.
- 3 Test does not represent a reasonable measure of the student's ability (a Kansas Assessment of Modified Measures [KAMM] or Alternate assessment was given to a student with no Special Education (SPED) code, the student was tested at a grade level different from their KIDS grade, the student did not try, cheating happened) so the 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 the 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 the student did not participate in testing.
- 0 The student was exempted from testing by a Special Circumstance code so the student was not included in AYP reporting.
- 1 The student was included in participation and % proficient calculations.
- 2 The student was represented by an opportunity-to-learn (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 either arrived after September 20th or 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 6 History and Government Spring Testing Results

	<b>Forms</b>				
	<b>637 (P&amp;P)</b>	<b>637 (KCA)</b>	<b>156</b>	<b>394</b>	<b>542</b>
<b>Number of Items</b>	48	48	48	48	48
<b>Sample Size (N)</b>	6934	7900	6468	6484	6470
<b>Mean Raw Score</b>	27.51	28.95	28.86	28.45	29.24
<b>SD Raw Score</b>	7.544	6.99	7.211	7.461	7.661
<b>Mean Scaled Score</b>	57.351	60.351	61.617	61.491	61.464
<b>SD Scaled Score</b>	15.717	14.559	14.63	14.463	14.436
<b>Reliability (alpha)</b>	0.836	0.813	0.822	0.836	0.845
<b>SEM Raw Score</b>	3.138	2.907	3.042	3.022	3.016
<b>SEM Scaled Score</b>	6.537	6.056	6.172	5.857	5.76
<b>Non-Mastery</b>	22.50%	14.80%	13.00%	12.50%	14.40%
<b>Mastery</b>	77.40%	85.10%	87.00%	87.50%	85.60%
<b>Academic Warning</b>	3.10%	1.30%	1.40%	1.10%	1.10%
<b>Approaches Standard</b>	19.40%	13.50%	11.60%	11.40%	13.30%
<b>Meets Standard</b>	42.00%	43.30%	44.50%	46.90%	45.00%
<b>Exceeds Standard</b>	28.00%	32.60%	30.10%	28.40%	29.10%
<b>Exemplary</b>	7.40%	9.20%	12.40%	12.20%	11.50%

### 2008 Grade 8 History and Government Spring Testing Results

	<b>Forms</b>				
	<b>849 (P&amp;P)</b>	<b>849 (KCA)</b>	<b>292</b>	<b>316</b>	<b>468</b>
<b>Number of Items</b>	60	60	60	60	60
<b>Sample Size (N)</b>	8363	7278	6416	6438	6442
<b>Mean Raw Score</b>	33.28	35.48	35.78	35.69	35.89
<b>SD Raw Score</b>	11.051	10.444	9.983	9.975	10.411
<b>Mean Scaled Score</b>	55.459	59.134	60.647	60.728	60.707
<b>SD Scaled Score</b>	18.424	17.41	17.135	17.117	17.119
<b>Reliability (alpha)</b>	0.903	0.894	0.822	0.884	0.894
<b>SEM Raw Score</b>	3.495	3.303	3.429	3.397	3.39
<b>SEM Scaled Score</b>	5.826	5.506	5.872	5.83	5.574
<b>Non-Mastery</b>	24.90%	17.80%	14.90%	15.30%	16.00%
<b>Mastery</b>	75.10%	82.20%	85.10%	84.70%	84.00%
<b>Academic Warning</b>	4.60%	2.20%	2.30%	2.30%	2.40%
<b>Approaches Standard</b>	20.30%	15.60%	12.60%	13.00%	13.60%
<b>Meets Standard</b>	43.10%	42.70%	42.60%	42.60%	43.40%
<b>Exceeds Standard</b>	20.40%	26.20%	27.20%	26.90%	26.20%
<b>Exemplary</b>	11.50%	13.20%	15.30%	15.10%	14.30%

**2008 High School U.S. History and Government Spring Testing Results**

	<b>Forms</b>				
	<b>946 (P&amp;P)</b>	<b>946 (KCA)</b>	<b>234</b>	<b>812</b>	<b>972</b>
<b>Number of Items</b>	29	29	30	29	30
<b>Sample Size (N)</b>	6638	6777	6404	6391	6391
<b>Grade 9</b>	1	12	9	10	12
<b>Grade 10</b>	9	117	118	108	110
<b>Grade 11</b>	6623	6644	6276	6273	6269
<b>Grade 12</b>	5	4	1	1	0
<b>Mean Raw Score</b>	18.38	18.22	19.76	18.27	19.21
<b>SD Raw Score</b>	5.224	5.127	4.946	4.956	5.041
<b>Mean Scaled Score</b>	63.396	62.84	61.574	63.614	61.589
<b>SD Scaled Score</b>	18.027	17.696	16.863	17.427	16.814
<b>Reliability (alpha)</b>	0.804	0.796	0.788	0.786	0.797
<b>SEM Raw Score</b>	2.336	2.293	2.277	2.172	2.271
<b>SEM Scaled Score</b>	7.981	7.993	7.764	8.062	7.576

**2008 High School World History and Government Spring Testing Results**

	<b>Forms</b>				
	<b>817 (P&amp;P)</b>	<b>817 (KCA)</b>	<b>296</b>	<b>319</b>	<b>588</b>
<b>Number of Items</b>	30	30	30	30	30
<b>Sample Size (N)</b>	6934	7900	7446	7469	7425
<b>Grade 9</b>	74	180	169	166	169
<b>Grade 10</b>	198	1093	999	1009	997
<b>Grade 11</b>	6659	6625	6277	6293	6259
<b>Grade 12</b>	3	2	1	1	0
<b>Mean Raw Score</b>	17.5	16.11	16.48	16.97	17.21
<b>SD Raw Score</b>	5.105	4.786	4.98	4.86	4.694
<b>Mean Scaled Score</b>	58.348	53.696	54.2909	54.226	54.314
<b>SD Scaled Score</b>	17.024	15.957	15.794	15.826	15.803
<b>Reliability (alpha)</b>	0.782	0.734	0.757	0.755	0.73
<b>SEM Raw Score</b>	2.384	2.468	2.455	2.406	2.469
<b>SEM Scaled Score</b>	7.949	8.23	7.786	7.833	8.212

## 2008 US and World History and Government Spring Testing Results

### All Form Combinations

<b>Number of Items</b>	60
<b>Sample Size (N)</b>	32,409
<b>Mean Scaled Score</b>	59.23
<b>SD Scaled Score</b>	15.596
<b>Reliability (stratified alpha)</b>	0.857
<b>Non-Mastery</b>	17.70%
<b>Mastery</b>	82.30%
<b>Academic Warning</b>	2.10%
<b>Approaches Standard</b>	15.60%
<b>Meets Standard</b>	47.20%
<b>Exceeds Standard</b>	26.90%
<b>Exemplary</b>	8.30%

### Section 3

## DIFFERENTIAL ITEM FUNCTIONING (DIF) ANALYSES

Examining differential item functioning (DIF) is an important step in test construction. DIF analysis 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 the test-takers have been matched on ability (or test total scores). DIF provides a necessary, but not sufficient, condition for item bias. Commonly, logical judgmental reviews of DIF items by panels representing impacted gender and ethnic groups need to be conducted before any judgments can be made about whether an item shows any bias, insensitivity, or offensiveness toward any gender or ethnic group.

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

Procedures for identifying DIF may be over-sensitive to different curriculum/instructional approaches that could influence performance given the content of an item. This effect is particularly important in Kansas where ethnic groups involved in the DIF analyses are largely congregated in a few districts, but where results would typically be compared to a random sample of 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 history and government 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 Caucasian students were drawn for each minority group.

In 2008, there were four history and government 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 history and government 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 history and government (Grades 6 and 8, and high school U.S. history and world history), the number of items on each test form at a given grade was 48, 60, 30, and 30, respectively. Thus, the total number of history and government items involved in the DIF analyses ranged from 192 at grade 6 to 240 at grades 8 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 Educational Testing Service (ETS) delta value larger than 1.5 and (2) the absolute ETS 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

Tables 3.1 through 3.4 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 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 the Hispanic/Caucasian and African American/Caucasian comparisons. In the third part (under the title *DIF Items*), ID numbers for items that are showing DIF are presented in the table. Specifically, the ID number for each item is a unique number in the CETE test system that makes it possible to track all changes and decisions made for the item. For each item ID number, a “+” or “-” sign indicates the direction of the DIF that the item shows. As mentioned earlier, items with “-” were seen to disadvantage the focal group while items with “+” advantaged this group in comparison to the reference group. The last part of each table gives the total counts of the number of flagged items for each test form (*Total*). Items that advantage or disadvantage the focal group were tallied separately.

## 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 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 analyses are 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 female 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 for students belonging to the respective ethnic or gender groupings. Each committee member was provided sets of flagged items from history and government 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 judgmental procedure. Panelists were directed to review each item flagged for DIF in the statistical analyses for students in their particular gender or ethnic group independently in terms of fairness, focusing specifically on content, language, offensiveness, or stereotypes that may have been present in the respective items. After the independent item review was completed by committee members, panelists engaged in a discussion regarding 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 was present, 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. Only one of the flagged items from the history and government assessments was judged to contain content or language that was biased against members of certain gender or ethnic groups. This item was deleted from the scoring of the forms on which it occurred and will not appear on future test forms.

Table 3.1  
*Summary of Differentially Functioning Items for Grade 6 History and Government Forms*

Form	Reference	DIF Group			DIF Items	Total
		N	Focal	N		
156	Male	3242	Female	3191	28820-	
	White	766	Hispanic	693	28835-	
		438	Black	339		
						<u>0 +, 2 -</u>
394	Male	3232	Female	3211	29009-	
	White	764	Hispanic	671		
		463	Black	359		
						<u>0 +, 1 -</u>
542	Male	3220	Female	3211		
	White	720	Hispanic	678	28837-	
		490	Black	341	28837-	
						<u>0 +, 2 -</u>
637	Male	3281	Female	3177		
	White	780	Hispanic	694		
		463	Black	363		
						<u>0 +, 0 -</u>
Total						<u>0 +, 5 -</u>

Table 3.2  
*Summary of Differentially Functioning Items for Grade 8 History and Government Forms*

Form	DIF Group				DIF Items	Total
	Reference	N	Focal	N		
292	Male	3263	Female	3111	29410+	
	White	711	Hispanic	636		
		504	Black	372	29207-	<u>1 +, 1 -</u>
316	Male	3286	Female	3110		
	White	793	Hispanic	639		
		473	Black	360		<u>0 +, 0 -</u>
468	Male	3289	Female	3107		
	White	804	Hispanic	673	29250-	
		468	Black	363		<u>0 +, 1 -</u>
849	Male	3295	Female	3101		
	White	764	Hispanic	655		
		478	Black	377	29243-	<u>0 +, 1 -</u>
Total						<u>1 +, 3 -</u>

Table 3.3  
*Summary of Differentially Functioning Items for Grade 11 U.S. History and Government Forms*

Form	DIF Group				DIF Items	Total
	Reference	N	Focal	N		
234	Male	3166	Female	3073		
	White	702	Hispanic	495	29708-	
		416	Black	320		<u>0 +, 1 -</u>
812	Male	3103	Female	3121		
	White	675	Hispanic	492	29578-	
		557	Black	375		<u>0 +, 1 -</u>
946	Male	3195	Female	3068	29875-	
	White	683	Hispanic	788	29578-	
		482	Black	371		<u>0 +, 2 -</u>
972	Male	3227	Female	3000		
	White	739	Hispanic	514		
		496	Black	372	29859+	<u>1 +, 0 -</u>
Total						<u>1 +, 4 -</u>



## 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 & Brennan, 1995; Lord, 1980). 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 History and Government 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. Test forms 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 of the equating design and methods, as well as issues in equating multiple forms of the Kansas Assessments.

### Procedures

#### Equating Design and Data Configuration

The Spring 2008 administration of the Kansas History and Government 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 history and government 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 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 group 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 that it may not sufficiently reflect the complete distribution of ability and performance of all students in the state. To the extent that the KCA score distribution differs from the complete state distribution, the score equating in one or both score distribution tails may contain greater error.

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

Table 4.1 below show the percentages of students across schools in Kansas who were administered the different KCA forms of the Kansas History and Government 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 4.1 suggest the equivalence of the KCA groups responding to each form, at all grades.

Table 4.1

*Number of KCA Kansas Schools and Percentages of Students Taking Different History and Government Test Forms*

Grade	N of Schools	Form	Gender		Race		Education		
			Total	Female	Male	White	Minority	Regular	Sped
6	513	156	25.0	49.6	50.4	77.1	22.9	92.8	7.2
	510	394	25.0	49.8	50.2	77.2	22.8	92.7	7.3
	511	542	25.0	49.9	50.1	77.7	22.3	92.6	7.4
	<b>508</b>	<b>637</b>	<b>25.1</b>	<b>49.2</b>	<b>50.8</b>	<b>77.1</b>	<b>22.9</b>	<b>92.8</b>	<b>7.2</b>
8	408	292	24.9	48.8	51.2	77.9	22.1	93.4	6.6
	410	316	25.0	48.6	51.4	78.4	21.6	93.6	6.4
	408	468	25.0	48.6	51.4	78.0	22.0	93.2	6.8
	<b>406</b>	<b>849</b>	<b>25.0</b>	<b>48.5</b>	<b>51.5</b>	<b>78.0</b>	<b>22.0</b>	<b>93.4</b>	<b>6.6</b>
11 US History	335	234	25.0	49.3	50.7	81.2	18.8	93.2	6.8
	336	812	24.9	50.1	49.9	80.2	19.8	93.3	6.7
	<b>333</b>	<b>946</b>	<b>25.1</b>	<b>49.0</b>	<b>51.0</b>	<b>80.8</b>	<b>19.2</b>	<b>93.0</b>	<b>7.0</b>
	333	972	25.0	48.2	51.8	80.6	19.4	93.2	6.8
11 World History	332	296	25.0	48.4	51.6	80.6	19.4	93.1	6.9
	334	319	25.1	49.3	50.7	80.8	19.2	93.6	6.4
	330	588	25.0	49.8	50.2	80.9	19.1	93.0	7.0
	<b>334</b>	<b>817</b>	<b>25.0</b>	<b>49.1</b>	<b>50.9</b>	<b>80.6</b>	<b>19.4</b>	<b>92.9</b>	<b>7.1</b>

### 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 history and government, 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 was made in the scores for either set of data.

### **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 4.1 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) equating book). 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 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 was chosen that created a smooth equating function, as long as the following conditions were met:

- the first four moments of the base form distribution were "preserved" (i.e., the equated scores' moments are very similar or the same to 2 to 4 decimal places).
- across the raw score distribution, the differences between the equated and raw scores are no larger than  $\pm 1$  standard error (SE) of the unsmoothed equipercentile equating solution. Note that all S values led to some conversions outside  $\pm 1$ 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 4.2 shows a descriptive summary of the equating samples obtained for history and government. The numbers for the base form at each grade level are shown in bold. All forms at a grade level were constructed with the same content and statistical specifications. Table 4.2 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. The table shows that all the forms across grade levels had sufficient reliability for equating purposes.

Table 4.2

*Descriptive Statistics for Equating Samples for History and Government by Test Form*

Grade	Form	N Items	N	Reliability ( $\alpha$ )	Mean Raw Score	SD Raw Score	Mean Percent Correct	SD Percent Correct
6	156	48	6433	0.82	28.91	7.18	60.22	14.96
	394	48	6443	0.84	28.50	7.43	59.38	15.49
	542	48	6431	0.84	29.29	7.62	61.02	15.88
	<b>637</b>	<b>48</b>	<b>6458</b>	<b>0.81</b>	<b>29.58</b>	<b>6.91</b>	<b>61.63</b>	<b>14.40</b>
8	292	60	6374	0.88	35.87	9.92	59.78	16.53
	316	60	6396	0.88	35.78	9.90	59.64	16.50
	468	60	6396	0.89	35.98	10.35	59.97	17.24
	<b>849</b>	<b>60</b>	<b>6396</b>	<b>0.89</b>	<b>36.51</b>	<b>10.21</b>	<b>60.85</b>	<b>17.01</b>
11	234	30	6239	0.79	19.82	4.93	66.05	16.42
US	812	29	6224	0.78	18.35	4.92	63.27	16.98
History	<b>946</b>	<b>29</b>	<b>6263</b>	<b>0.79</b>	<b>18.54</b>	<b>5.01</b>	<b>63.93</b>	<b>17.26</b>
	972	30	6227	0.80	19.27	5.03	64.24	16.75
11	296	30	6230	0.76	16.63	5.01	55.42	16.69
World	319	30	6249	0.76	17.15	4.85	57.15	16.16
History	588	30	6221	0.74	17.36	4.72	57.88	15.72
	<b>817</b>	<b>30</b>	<b>6227</b>	<b>0.73</b>	<b>16.43</b>	<b>4.75</b>	<b>54.75</b>	<b>15.84</b>

### Example Equating Output

In total, 12 separate equating analyses were conducted (3 forms equated to a base form for each of the four history and government assessments). Since these produce a large quantity of output, only one equating example is demonstrated here. Similar output is available for all other equating analyses. The following example is based on the grade 6 assessment. For this equating, Form 156 was equated to Form 637 (the base form) on the percent correct scale. Both of these forms contain 48 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 distribution were inspected to determine comparability. The histograms for Base Form 637 and equated Form 156 are contained in Figures 4.1 and 4.2, respectively. Visual inspection of these figures demonstrated that distributions of scores were generally very similar, indicating the appropriateness of equating these forms.

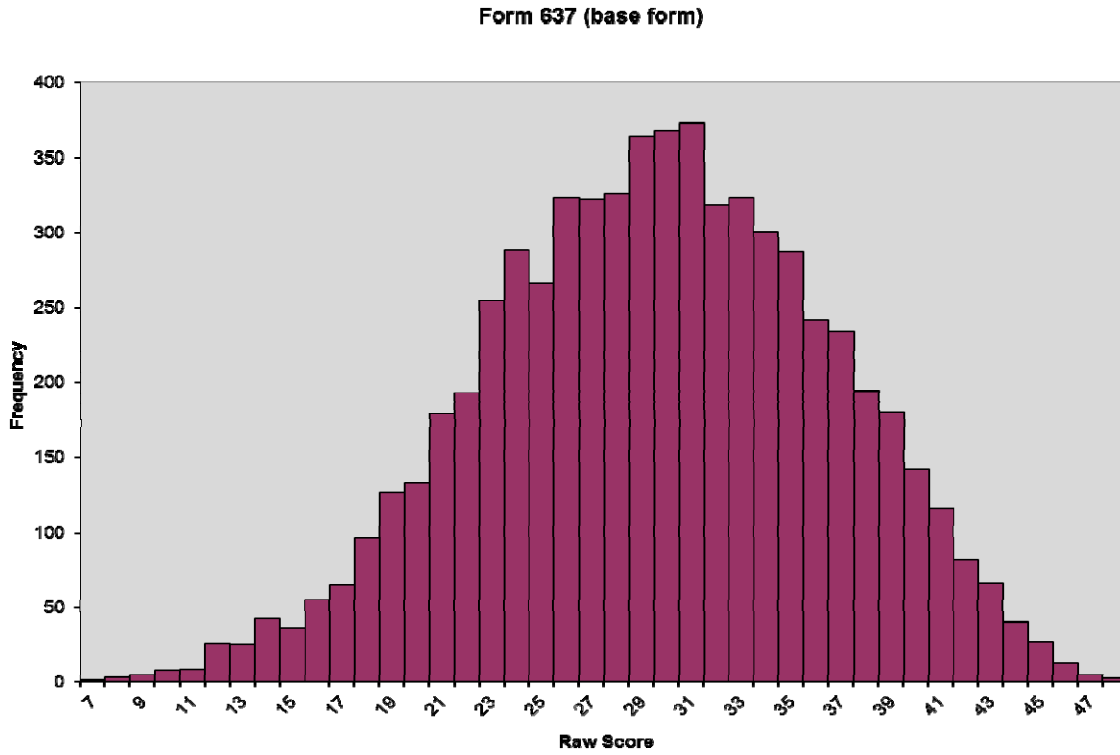


Figure 4.1. Raw score distribution for Base Form 637, Grade 6 History and Government.

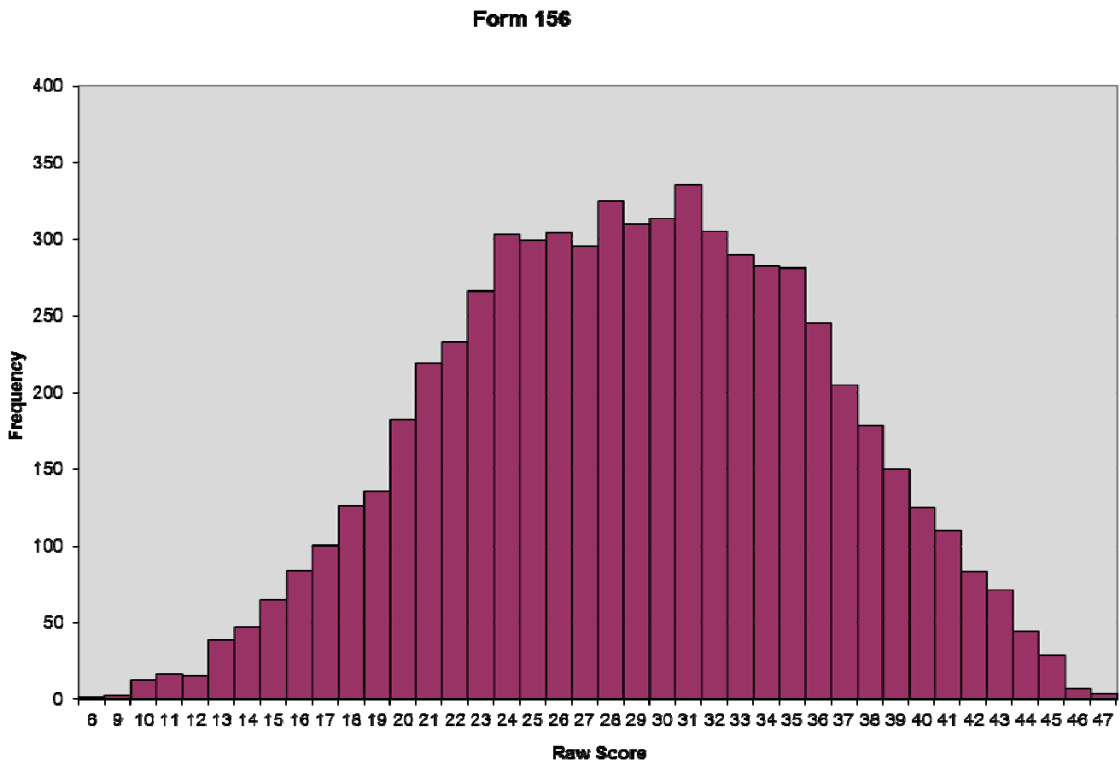


Figure 4.2. Raw score distribution for equated Form 156, Grade 6 History and Government.

Table 4.3 shows that the equipercntile 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 4.3  
*Moments of the Base Form 637 and Equated Form 156 by Equating Method*

<b>Test Form/Method</b>	<b>Mean</b>	<b>SD</b>	<b>Skewness</b>	<b>Kurtosis</b>
<b>Raw Scores</b>				
<b>Form 637</b>	29.5827	6.9123	-0.1409	2.6847
<b>Form 156</b>	28.9074	7.1821	-0.0625	2.4728
<b>Form 156 equated to Base Form 637</b>				
Unsmoothed	29.5814	6.9077	-0.1395	2.6781
S=0.01	29.5813	6.9058	-0.1416	2.6661
S=0.05	29.5805	6.9061	-0.1420	2.6637
<b>S=0.10</b>	<b>29.5802</b>	<b>6.9037</b>	<b>-0.1416</b>	<b>2.6568</b>
S=0.20	29.5802	6.9104	-0.1388	2.6321
S=0.30	29.5801	6.9151	-0.1353	2.6071
S=0.40	29.5797	6.9188	-0.1319	2.5876
S=0.50	29.5789	6.9219	-0.1283	2.5717
S=0.75	29.5718	6.9274	-0.1185	2.5408
S=1.00	29.5827	6.9306	-0.1067	2.5183

The mean difference between base Form 637 and Form 156 was 0.675. After equating, the mean difference between the forms was -0.0025. 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 equipercntile 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).

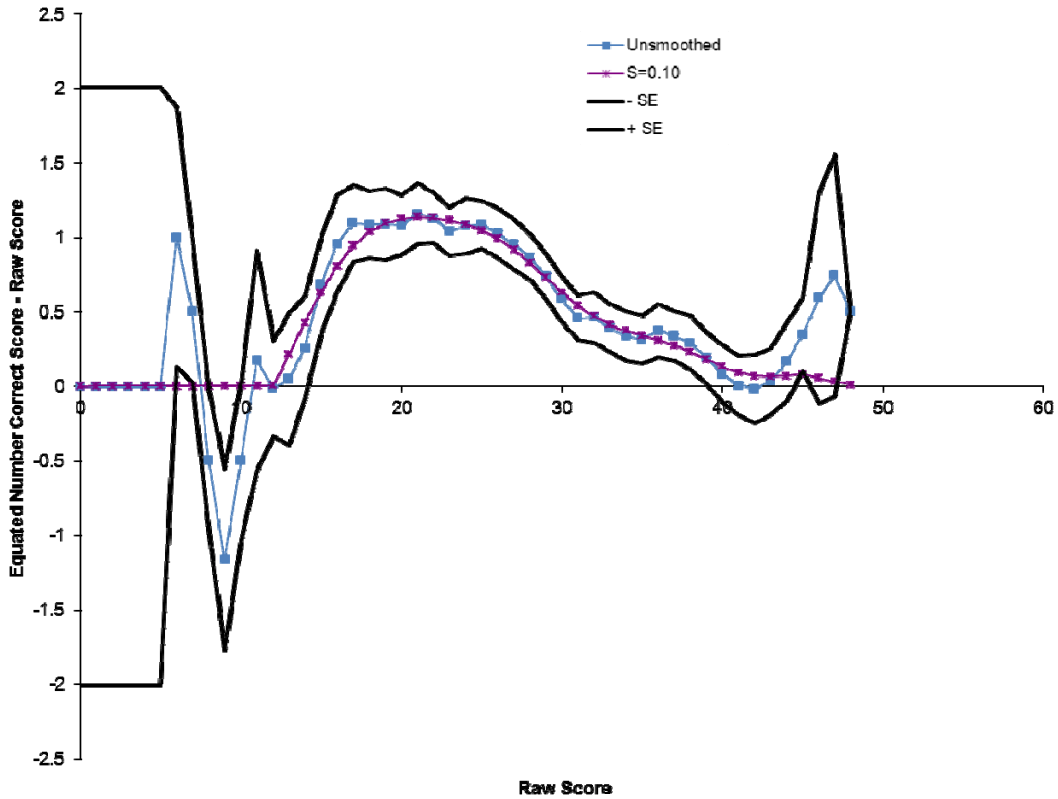


Figure 4.3. Grade 6 History and Government Form 156 equated to base Form 637 (number-correct scale).

Table 4.4 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 ( $S = 0.1$ ) chosen for this equating.

Table 4.4  
*Conversion Table for Various Methods*

Raw Score	Unsmoothed	S=0.01	S=0.05	<b>S=0.10</b>	S=0.20	S=0.30	S=0.40	S=0.50	S=0.75	S=1.00
0	0.00	0.00	0.00	<b>0.00</b>	0.01	0.01	0.02	0.02	0.03	0.04
1	1.00	0.99	0.99	<b>1.00</b>	1.03	1.04	1.06	1.07	1.09	1.11
2	2.00	1.98	1.98	<b>2.00</b>	2.04	2.07	2.09	2.11	2.15	2.18
3	3.00	2.98	2.98	<b>3.00</b>	3.06	3.10	3.13	3.16	3.21	3.25
4	4.00	3.97	3.97	<b>4.00</b>	4.07	4.13	4.17	4.20	4.27	4.32
5	5.00	4.96	4.96	<b>5.00</b>	5.09	5.16	5.20	5.24	5.32	5.39
6	7.00	5.96	5.96	<b>6.00</b>	6.11	6.19	6.24	6.29	6.38	6.47
7	7.50	6.95	6.95	<b>7.00</b>	7.12	7.21	7.28	7.33	7.44	7.54
8	7.50	7.94	7.95	<b>8.00</b>	8.14	8.24	8.32	8.38	8.50	8.61
9	7.84	8.94	8.94	<b>9.00</b>	9.16	9.27	9.35	9.42	9.56	9.68
10	9.50	9.93	9.93	<b>10.00</b>	10.17	10.30	10.39	10.46	10.62	10.75
11	11.18	10.92	10.93	<b>11.01</b>	11.19	11.33	11.43	11.51	11.68	11.82
12	11.99	11.92	11.92	<b>12.01</b>	12.22	12.37	12.48	12.56	12.74	12.88
13	13.05	13.09	13.15	<b>13.21</b>	13.37	13.48	13.57	13.63	13.78	13.90
14	14.26	14.33	14.40	<b>14.43</b>	14.51	14.59	14.65	14.71	14.82	14.91
15	15.68	15.64	15.64	<b>15.63</b>	15.66	15.70	15.74	15.77	15.85	15.93
16	16.96	16.92	16.85	<b>16.81</b>	16.79	16.80	16.82	16.84	16.89	16.94
17	18.10	18.06	17.99	<b>17.95</b>	17.90	17.89	17.89	17.90	17.92	17.95
18	19.09	19.10	19.08	<b>19.04</b>	18.98	18.96	18.95	18.94	18.95	18.96
19	20.09	20.10	20.12	<b>20.10</b>	20.05	20.01	20.00	19.98	19.97	19.96
20	21.09	21.11	21.13	<b>21.13</b>	21.09	21.05	21.03	21.01	20.98	20.96
21	22.16	22.13	22.13	<b>22.14</b>	22.11	22.08	22.05	22.02	21.98	21.96
22	23.13	23.11	23.12	<b>23.13</b>	23.12	23.08	23.05	23.03	22.98	22.95
23	24.04	24.08	24.10	<b>24.12</b>	24.10	24.07	24.04	24.02	23.97	23.93
24	25.08	25.08	25.09	<b>25.09</b>	25.07	25.04	25.02	25.00	24.95	24.91
25	26.09	26.08	26.06	<b>26.05</b>	26.03	26.01	25.98	25.97	25.92	25.88
26	27.03	27.03	27.01	<b>26.99</b>	26.97	26.96	26.94	26.92	26.89	26.85
27	27.96	27.96	27.94	<b>27.92</b>	27.90	27.89	27.88	27.87	27.85	27.81
28	28.87	28.86	28.84	<b>28.83</b>	28.83	28.83	28.82	28.82	28.80	28.77
29	29.74	29.73	29.73	<b>29.73</b>	29.74	29.75	29.76	29.76	29.75	29.73
30	30.59	30.60	30.62	<b>30.63</b>	30.66	30.68	30.69	30.69	30.69	30.68
31	31.46	31.49	31.52	<b>31.54</b>	31.58	31.60	31.62	31.63	31.64	31.63
32	32.47	32.44	32.45	<b>32.47</b>	32.51	32.54	32.55	32.57	32.58	32.58
33	33.39	33.39	33.40	<b>33.42</b>	33.45	33.47	33.49	33.50	33.53	33.53
34	34.34	34.35	34.37	<b>34.37</b>	34.39	34.41	34.43	34.45	34.47	34.48
35	35.32	35.34	35.35	<b>35.34</b>	35.34	35.36	35.38	35.39	35.42	35.43
36	36.38	36.36	36.34	<b>36.31</b>	36.30	36.31	36.33	36.34	36.36	36.38
37	37.34	37.35	37.31	<b>37.28</b>	37.26	37.27	37.28	37.29	37.31	37.33
38	38.30	38.29	38.25	<b>38.23</b>	38.22	38.22	38.23	38.24	38.26	38.28
39	39.19	39.20	39.18	<b>39.18</b>	39.18	39.18	39.19	39.19	39.21	39.23
40	40.08	40.09	40.11	<b>40.13</b>	40.14	40.15	40.15	40.15	40.16	40.18
41	41.01	41.01	41.06	<b>41.09</b>	41.11	41.11	41.11	41.11	41.11	41.13
42	41.99	41.99	42.04	<b>42.07</b>	42.08	42.08	42.07	42.07	42.07	42.08
43	43.03	43.04	43.06	<b>43.07</b>	43.06	43.05	43.04	43.03	43.02	43.03
44	44.17	44.16	44.11	<b>44.07</b>	44.04	44.02	44.00	43.99	43.97	43.98
45	45.35	45.20	45.13	<b>45.08</b>	45.04	45.01	44.99	44.98	44.96	44.97
46	46.60	46.14	46.09	<b>46.06</b>	46.03	46.01	45.99	45.98	45.97	45.98
47	47.75	47.09	47.05	<b>47.04</b>	47.02	47.01	47.00	46.99	46.98	46.99
48	48.50	48.03	48.02	<b>48.01</b>	48.01	48.00	48.00	48.00	47.99	48.00

Table 4.5 shows the same conversion table when transformed onto the percent correct metric for the same sample of score values. In this percent correct metric, differences in distributional smoothness throughout the scale are not immediately apparent. However, using these conversion tables in accord with the moments output from the various methods and the graphical representation of the conversions for each method is useful. Given a consideration of the multiple criteria discussed for making equating decisions, it appeared that the equipercentile method with a smoothing parameter applied ( $S = 0.10$ ) gave the most reasonable conversion.

Table 4.5

*Conversion Table for Various Methods Expressed in Percent Correct Metric*

Raw Score	Unsmoothed	S=0.01	S=0.05	<b>S=0.10</b>	S=0.20	S=0.30	S=0.40	S=0.50	S=0.75	S=1.00
0	0	0	0	<b>0</b>	0	0	0	0	0	0
1	2	2	2	<b>2</b>	2	2	2	2	2	2
2	4	4	4	<b>4</b>	4	4	4	4	4	5
3	6	6	6	<b>6</b>	6	6	7	7	7	7
4	8	8	8	<b>8</b>	8	9	9	9	9	9
5	10	10	10	<b>10</b>	11	11	11	11	11	11
6	15	12	12	<b>13</b>	13	13	13	13	13	13
7	16	14	14	<b>15</b>	15	15	15	15	16	16
8	16	17	17	<b>17</b>	17	17	17	17	18	18
9	16	19	19	<b>19</b>	19	19	19	20	20	20
10	20	21	21	<b>21</b>	21	21	22	22	22	22
11	23	23	23	<b>23</b>	23	24	24	24	24	25
12	25	25	25	<b>25</b>	25	26	26	26	27	27
13	27	27	27	<b>28</b>	28	28	28	28	29	29
14	30	30	30	<b>30</b>	30	30	31	31	31	31
15	33	33	33	<b>33</b>	33	33	33	33	33	33
16	35	35	35	<b>35</b>	35	35	35	35	35	35
17	38	38	37	<b>37</b>	37	37	37	37	37	37
18	40	40	40	<b>40</b>	40	40	39	39	39	40
19	42	42	42	<b>42</b>	42	42	42	42	42	42
20	44	44	44	<b>44</b>	44	44	44	44	44	44
21	46	46	46	<b>46</b>	46	46	46	46	46	46
22	48	48	48	<b>48</b>	48	48	48	48	48	48
23	50	50	50	<b>50</b>	50	50	50	50	50	50
24	52	52	52	<b>52</b>	52	52	52	52	52	52
25	54	54	54	<b>54</b>	54	54	54	54	54	54
26	56	56	56	<b>56</b>	56	56	56	56	56	56
27	58	58	58	<b>58</b>	58	58	58	58	58	58
28	60	60	60	<b>60</b>	60	60	60	60	60	60
29	62	62	62	<b>62</b>	62	62	62	62	62	62
30	64	64	64	<b>64</b>	64	64	64	64	64	64
31	66	66	66	<b>66</b>	66	66	66	66	66	66
32	68	68	68	<b>68</b>	68	68	68	68	68	68
33	70	70	70	<b>70</b>	70	70	70	70	70	70
34	72	72	72	<b>72</b>	72	72	72	72	72	72
35	74	74	74	<b>74</b>	74	74	74	74	74	74
36	76	76	76	<b>76</b>	76	76	76	76	76	76
37	78	78	78	<b>78</b>	78	78	78	78	78	78
38	80	80	80	<b>80</b>	80	80	80	80	80	80
39	82	82	82	<b>82</b>	82	82	82	82	82	82
40	84	84	84	<b>84</b>	84	84	84	84	84	84
41	85	85	86	<b>86</b>	86	86	86	86	86	86
42	87	87	88	<b>88</b>	88	88	88	88	88	88
43	90	90	90	<b>90</b>	90	90	90	90	90	90
44	92	92	92	<b>92</b>	92	92	92	92	92	92
45	94	94	94	<b>94</b>	94	94	94	94	94	94
46	97	96	96	<b>96</b>	96	96	96	96	96	96
47	99	98	98	<b>98</b>	98	98	98	98	98	98
48	101	100	100	<b>100</b>	100	100	100	100	100	100

## Equating Decisions

The equating methods selected for each form in history and government are summarized in this section of the report. A total of 12 equating analyses were performed, 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 history and government required the use of conversion tables. These conversion tables, similar to the samples detailed in Tables 4.4 and 4.5, are not provided in this report due to 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 for all 12 equating decisions. The smoothing parameter, *S*, selected for each decision, was 0.05 or 0.10. All forms were equated at the raw score level and subsequently expressed in the percent correct metric.

Table 4.6 details the equating decisions adopted for history and government. The smoothing parameters varied across forms, and are detailed in the table.

Table 4.6  
*Summary of Equating Decisions for History and Government*

<b>History and Government</b>		
<b>Grade</b>	<b>Form</b>	<b>S</b>
6 (Base Form 637)	156	0.05
	394	0.05
	542	0.05
8 (Base Form 849)	292	0.10
	316	0.10
	468	0.10
11 U.S. History (Base Form 946)	234	0.10
	812	0.10
	972	0.10
11 World History (Base Form 817)	296	0.10
	319	0.10
	588	0.10

## **Summary**

Test forms for the 2008 Kansas Assessments in History and Government were built to the same specifications, as articulated by KSDE. The reliability coefficients for all forms were acceptable for the purpose of equating. Furthermore, 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.

## 2008 Grade 6 History and Government Equating Conversion Table Results

Raw Score	Base Form 637		Form 156		Form 394		Form 542	
	Score	% Correct	Equated Score	Equated % Correct	Equated Score	Equated % Correct	Equated Score	Equated % Correct
0	0	0	0.00	0	0.02	0	0.03	0
1	1	2	0.99	2	1.07	2	1.07	2
2	2	4	1.98	4	2.12	4	2.12	4
3	3	6	2.98	6	3.16	7	3.17	7
4	4	8	3.97	8	4.21	9	4.22	9
5	5	10	4.96	10	5.26	11	5.27	11
6	6	13	5.96	12	6.30	13	6.32	13
7	7	15	6.95	14	7.35	15	7.37	15
8	8	17	7.95	17	8.40	17	8.42	18
9	9	19	8.94	19	9.44	20	9.47	20
10	10	21	9.93	21	10.49	22	10.52	22
11	11	23	10.93	23	11.54	24	11.57	24
12	12	25	11.92	25	12.62	26	12.68	26
13	13	27	13.15	27	13.85	29	13.99	29
14	14	29	14.40	30	15.25	32	15.29	32
15	15	31	15.64	33	16.72	35	16.55	34
16	16	33	16.85	35	17.98	37	17.73	37
17	17	35	17.99	37	19.04	40	18.81	39
18	18	38	19.08	40	20.01	42	19.81	41
19	19	40	20.12	42	20.96	44	20.74	43
20	20	42	21.13	44	21.89	46	21.62	45
21	21	44	22.13	46	22.81	48	22.47	47
22	22	46	23.12	48	23.72	49	23.29	49
23	23	48	24.10	50	24.62	51	24.11	50
24	24	50	25.09	52	25.55	53	24.93	52
25	25	52	26.06	54	26.49	55	25.75	54
26	26	54	27.01	56	27.45	57	26.59	55
27	27	56	27.94	58	28.36	59	27.44	57
28	28	58	28.84	60	29.23	61	28.29	59
29	29	60	29.73	62	30.08	63	29.16	61
30	30	63	30.62	64	30.91	64	30.02	63
31	31	65	31.52	66	31.79	66	30.90	64
32	32	67	32.45	68	32.72	68	31.79	66
33	33	69	33.40	70	33.67	70	32.71	68
34	34	71	34.37	72	34.63	72	33.64	70
35	35	73	35.35	74	35.59	74	34.60	72
36	36	75	36.34	76	36.53	76	35.58	74
37	37	77	37.31	78	37.43	78	36.57	76
38	38	79	38.25	80	38.33	80	37.56	78
39	39	81	39.18	82	39.24	82	38.53	80
40	40	83	40.11	84	40.17	84	39.47	82
41	41	85	41.06	86	41.13	86	40.40	84
42	42	88	42.04	88	42.07	88	41.34	86
43	43	90	43.06	90	42.93	89	42.28	88
44	44	92	44.11	92	43.77	91	43.23	90
45	45	94	45.13	94	44.62	93	44.19	92
46	46	96	46.09	96	45.70	95	45.33	94
47	47	98	47.05	98	46.82	98	46.60	97
48	48	100	48.02	100	47.94	100	47.87	100

**2008 Grade 8 History and Government Conversion Table Results**

Raw Score	Base Form 849		Form 292		Form 316		Form 468	
	Score	% Correct	S = .10 Equated Score	S = .10 Equated % Correct	S = .10 Equated Score	S = .10 Equated % Correct	S = .10 Equated Score	S = .10 Equated % Correct
0	0	0	0.02	0	0.01	0	0.02	0
1	1	2	1.06	2	1.04	2	1.05	2
2	2	3	2.10	3	2.07	3	2.08	3
3	3	5	3.14	5	3.09	5	3.11	5
4	4	7	4.17	7	4.12	7	4.14	7
5	5	8	5.21	9	5.15	9	5.17	9
6	6	10	6.25	10	6.18	10	6.20	10
7	7	12	7.29	12	7.20	12	7.23	12
8	8	13	8.33	14	8.23	14	8.26	14
9	9	15	9.37	16	9.26	15	9.29	15
10	10	17	10.41	17	10.28	17	10.32	17
11	11	18	11.45	19	11.31	19	11.35	19
12	12	20	12.49	21	12.34	21	12.38	21
13	13	22	13.50	23	13.36	22	13.42	22
14	14	23	14.44	24	14.32	24	14.44	24
15	15	25	15.38	26	15.28	25	15.46	26
16	16	27	16.33	27	16.24	27	16.48	27
17	17	28	17.27	29	17.21	29	17.50	29
18	18	30	18.22	30	18.18	30	18.51	31
19	19	32	19.17	32	19.15	32	19.52	33
20	20	33	20.12	34	20.12	34	20.53	34
21	21	35	21.07	35	21.11	35	21.54	36
22	22	37	22.04	37	22.10	37	22.55	38
23	23	38	23.01	38	23.10	39	23.56	39
24	24	40	23.99	40	24.11	40	24.57	41
25	25	42	25.00	42	25.13	42	25.59	43
26	26	43	26.02	43	26.16	44	26.61	44
27	27	45	27.06	45	27.20	45	27.63	46
28	28	47	28.12	47	28.26	47	28.65	48
29	29	48	29.20	49	29.33	49	29.68	49
30	30	50	30.30	50	30.41	51	30.71	51
31	31	52	31.41	52	31.51	53	31.74	53
32	32	53	32.52	54	32.61	54	32.76	55
33	33	55	33.63	56	33.71	56	33.78	56
34	34	57	34.73	58	34.81	58	34.80	58

Raw Score	Base Form 849		Form 292		Form 316		Form 468	
	Score	% Correct	S = .10 Equated Score	S = .10 Equated % Correct	S = .10 Equated Score	S = .10 Equated % Correct	S = .10 Equated Score	S = .10 Equated % Correct
35	35	58	35.82	60	35.90	60	35.80	60
36	36	60	36.90	61	36.99	62	36.80	61
37	37	62	37.95	63	38.06	63	37.78	63
38	38	63	38.99	65	39.11	65	38.76	65
39	39	65	40.01	67	40.15	67	39.72	66
40	40	67	41.02	68	41.16	69	40.68	68
41	41	68	42.01	70	42.16	70	41.63	69
42	42	70	43.00	72	43.15	72	42.58	71
43	43	72	43.98	73	44.13	74	43.52	73
44	44	73	44.95	75	45.10	75	44.46	74
45	45	75	45.94	77	46.07	77	45.40	76
46	46	77	46.92	78	47.03	78	46.34	77
47	47	78	47.91	80	47.99	80	47.28	79
48	48	80	48.89	81	48.96	82	48.21	80
49	49	82	49.87	83	49.91	83	49.15	82
50	50	83	50.85	85	50.87	85	50.08	83
51	51	85	51.81	86	51.82	86	51.01	85
52	52	87	52.76	88	52.77	88	51.93	87
53	53	88	53.70	89	53.71	90	52.85	88
54	54	90	54.63	91	54.65	91	53.77	90
55	55	92	55.56	93	55.60	93	54.68	91
56	56	93	56.47	94	56.50	94	55.60	93
57	57	95	57.37	96	57.39	96	56.66	94
58	58	97	58.26	97	58.28	97	57.76	96
59	59	98	59.16	99	59.17	99	58.86	98
60	60	100	60.05	100	60.06	100	59.95	100

**2008 High School U.S. History and Government Conversion Table Results**

Raw Score	Base Form 946		Form 234		Form 812		Form 946	
	Score	% Correct	S = .10 Equated Score	S = .10 Equated % Correct	S = .10 Equated Score	S = .10 Equated % Correct	S = .10 Equated Score	S = .10 Equated % Correct
0	0	0	-0.02	0	0.01	0	-0.01	0
1	1	3	0.94	3	1.04	4	0.96	3
2	2	7	1.90	6	2.07	7	1.94	6
3	3	10	2.86	10	3.09	11	2.91	10
4	4	14	3.82	13	4.12	14	3.89	13
5	5	17	4.78	16	5.15	18	4.86	16
6	6	21	5.74	19	6.16	21	5.84	19
7	7	24	6.55	22	7.11	25	6.71	22
8	8	28	7.32	24	8.06	28	7.57	25
9	9	31	8.11	27	9.02	31	8.44	28
10	10	34	8.91	30	9.99	34	9.32	31
11	11	38	9.74	32	10.97	38	10.23	34
12	12	41	10.59	35	11.98	41	11.17	37
13	13	45	11.47	38	13.01	45	12.15	40
14	14	48	12.38	41	14.06	48	13.16	44
15	15	52	13.34	44	15.12	52	14.20	47
16	16	55	14.34	48	16.18	56	15.24	51
17	17	59	15.38	51	17.22	59	16.26	54
18	18	62	16.45	55	18.24	63	17.26	58
19	19	66	17.54	58	19.25	66	18.26	61
20	20	69	18.65	62	20.25	70	19.26	64
21	21	72	19.77	66	21.24	73	20.27	68
22	22	76	20.88	70	22.25	77	21.29	71
23	23	79	21.96	73	23.25	80	22.29	74
24	24	83	22.97	77	24.26	84	23.27	78
25	25	86	23.94	80	25.27	87	24.25	81
26	26	90	24.88	83	26.31	91	25.23	84
27	27	93	25.83	86	27.36	94	26.24	87
28	28	97	26.83	89	28.26	97	27.26	91
29	29	100	27.84	93	29.09	100	28.49	95
30			29.60	99			29.83	99

**2008 High School World History and Government Conversion Table Results**

Raw Score	Base Form 817		Form 296 S = .10 S = .10		Form 319 S = .10 S = .10		Form 588 S = .10 S = .10	
	Score	% Correct	Equated Score	Equated % Correct	Equated Score	Equated % Correct	Equated Score	Equated % Correct
0	0	0	0.01	0	0.00	0	-0.06	0
1	1	3	1.02	3	0.99	3	0.81	3
2	2	7	2.04	7	1.98	7	1.68	6
3	3	10	3.06	10	2.98	10	2.55	9
4	4	13	4.07	14	3.97	13	3.42	11
5	5	17	5.09	17	4.96	17	4.30	14
6	6	20	6.10	20	5.96	20	5.17	17
7	7	23	7.10	24	6.85	23	6.10	20
8	8	27	8.09	27	7.75	26	7.08	24
9	9	30	9.09	30	8.64	29	8.06	27
10	10	33	10.08	34	9.54	32	9.04	30
11	11	37	11.07	37	10.44	35	10.02	33
12	12	40	12.05	40	11.36	38	11.01	37
13	13	43	13.02	43	12.29	41	12.01	40
14	14	47	13.99	47	13.24	44	13.01	43
15	15	50	14.96	50	14.21	47	14.02	47
16	16	53	15.91	53	15.20	51	15.04	50
17	17	57	16.85	56	16.20	54	16.05	54
18	18	60	17.79	59	17.21	57	17.06	57
19	19	63	18.71	62	18.21	61	18.07	60
20	20	67	19.63	65	19.22	64	19.07	64
21	21	70	20.55	69	20.22	67	20.07	67
22	22	73	21.48	72	21.24	71	21.08	70
23	23	77	22.41	75	22.25	74	22.10	74
24	24	80	23.34	78	23.25	78	23.12	77
25	25	83	24.28	81	24.25	81	24.15	81
26	26	87	25.21	84	25.24	84	25.19	84
27	27	90	26.15	87	26.22	87	26.23	87
28	28	93	27.11	90	27.24	91	27.39	91
29	29	97	28.46	95	28.54	95	28.63	95
30	30	100	29.82	99	29.85	99	29.88	100

## Section 5

### STANDARD SETTING

#### 2008 Kansas History and Government Performance Standards

Performance standards were set for the 2008 Kansas History and Government Assessments using a multistep process 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 or academic discipline. The last step is to present the State Board of Education with the information it needs to set the Kansas cut score policy.

#### 1. Development of Performance Level Names

The Kansas State Board of Education met on August 8, 2006, and 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 history and government at grades 6, 8, and high school.

### **3. Bookmark Procedure**

The Bookmark Procedure (Mitzel, Lewis, Patz, & Green, 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 history and government were conducted for grades 6, 8, and high school (U.S. history and world history). The four 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*

<b>Subject</b>	<b>Grade</b>	<b># Items (General)</b>
Social Studies	6	48
Social Studies	8	60
U.S. History	High School	29
World History	High School	30

Ordered item booklets were prepared according to guidelines prescribed by Mitzel et al. (2001). Ordering of items was accomplished by (1) fitting an Item Response Theory (IRT) model to the test data, (2) determining response probability (RP) -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  $\theta_j$  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. The only effect on the probability of success is the c-parameter. Items such as these, although they do not increase reliability or 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.

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  $\theta_p$  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 history and government assessments, and ordered item booklets were created on the basis of them.

Table 5.2  
*Item Ordering for the High School U.S. History Assessment*

<b>Ordered Booklet</b>	<b>Item Number</b>	<b>RP-67</b>
1	12	-2.2628
2	7	-1.5660
3	14	-1.1891
4	29	-1.1748
5	22	-1.0982
6	27	-0.7130
7	4	-0.4870
8	16	-0.4577
9	25	-0.3834
10	1	-0.1040
11	18	-0.0723
12	23	0.0418
13	13	0.1009
14	20	0.1072
15	26	0.1273
16	15	0.2053
17	2	0.3012
18	9	0.4682
19	11	0.6368
20	10	0.7298
21	24	0.9038
22	21	1.0463
23	28	1.0599
24	30	1.2002
25	5	1.2272
26	8	1.2756
27	3	1.2803
28	6	1.3804
29	19	2.6061

*Note.* Item # 17 was removed, but other item names are unchanged. ( e.g., there is still an item #30)

## Panel Participants

Invitation letters were sent to a random 50% 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 social studies. 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 both days 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 20<sup>th</sup>, then continue on Saturday, June 21<sup>st</sup> 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 in social studies content and grade (elementary, middle/jr. high, or high school), 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 history and government 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 history and government 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 the forty-six Kansas educators participated on one of seven 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*

<b>Grade and Subject</b>	<b>General</b>
Grade 6 Social Studies	8
Grade 8 Social Studies	8
HS U.S. History and Government	8
HS World History and Government	8

Of these participants, 29 were female and 22 were male; 48 were Caucasian and 3 were minorities; 5 had fewer than two years teaching experience, 9 had 3-5 years of experience, 16 had 6-10 years, 13 had 11-20 years, and 8 had more than 20 years; 1 came from an inner city school, 8 from other urban, 23 from suburban, and 19 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 history and government assessments.

For the high school history and government assessment, the test is divided into two parts: one for U.S. history and one for world history. 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  
*History and Government Bookmark Procedure Recommended Performance Level Score Ranges*

Performance Level	General		
	6	8	HS
Academic Warning	0-29	0-30	0-28
Approaches Standard	30-48	31-45	29-45
Meets Standard	49-65	46-67	46-68
Exceeds Standard	66-83	68-79	69-83
Exemplary	84-100	80-100	84-100

### Evaluation of Bookmark Procedure

At the end of the Bookmark Procedure meeting, participants were asked to evaluate the session. Key findings included: 100% of the participants in history and government found the training adequate or very adequate, and 79% of the history and government 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 history and government 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. 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 Figure 5.1 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.
- *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 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.

**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% 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  $(0.6 \times .5) + (0.8 \times .25) + (0.9 \times .25)$ , 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 History and Government 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		
	6	8	HS	6	8	HS
Academic Warning	29	30	28	30	27	28
Approaches Standard	48	45	45	47	44	44
Meets Standard	65	67	68	66	68	66
Exceeds Standard	83	79	83	81	81	80
Exemplary	100	100	100	100	100	100

Table 5.6  
*General History and Government Effect of Cross-Grade Consistency on Percent in Category*

Performance Level	Bookmark Recommended Percent in Category			Cross-Grade Consistent Percent in Category		
	6	8	HS	6	8	HS
Academic Warning	1%	5%	2%	2%	3%	2%
Approaches Standard	14%	17%	18%	14%	16%	15%
Meets Standard	45%	39%	49%	45%	41%	47%
Exceeds Standard	32%	25%	25%	29%	27%	27%
Exemplary	8%	15%	6%	10%	12%	8%

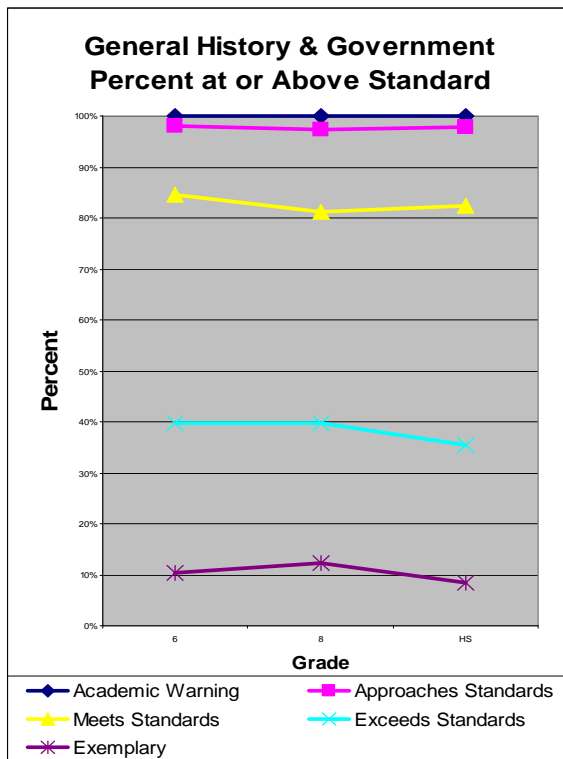


Figure 5.1. Examples of more consistent cut score graphs presented to participants.

**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 History and Government*

Assessment Type	Grade	Academic Warning	Approaches Standard	Meets Standard	Exceeds Standard	Exemplary
General	6	0-27	28-45	46-64	65-79	80-100
General	8	0-26	27-41	42-66	67-79	80-100
General	HS	0-27	28-43	44-66	67-80	81-100

**Evaluation of Standard Setting Policy Advisory Group Meeting**

At the end, participants evaluated the meeting, and 95% 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 History & Government	2	4	16	39	39

In all cases, more than 78% of the participants were confident or very confident in their decisions. 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%.

**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 4.2. 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 0.73 to a high of 0.90 across all the history and government forms. The overall general standard errors of measurement on the scale score range from 5.57 to 8.23.

Table 6.1  
*History and Government Form Reliabilities by Test Form*

Grade	Form	( $\alpha$ ) Reliability	SEM % Correct
6	156	0.82	6.17
	394	0.84	5.86
	542	0.85	5.76
	<b>637</b>	<b>0.83</b>	<b>6.33</b>
8	292	0.88	5.87
	316	0.88	5.83
	468	0.89	5.57
	<b>849</b>	<b>0.90</b>	<b>5.71</b>
11 US	234	0.79	7.76
	812	0.79	8.06
	<b>946</b>	<b>0.80</b>	<b>7.99</b>
	972	0.80	7.58
11 World	296	0.76	7.79
	319	0.76	7.83
	588	0.73	8.21
	<b>817</b>	<b>0.76</b>	<b>8.23</b>

#### 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 the probability of correctly categorizing a student, if their true category was known. 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* (AERA/APA/NCME, 1999): “When a test or combination of measures is used to make categorical decisions, estimates should be provided of the percentage of examinees who would be classified in the same way on two applications of the procedure, using the same form or alternate forms of the instruments” (p. 35).

## **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 using 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 using 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**

Analyses were conducted using the base form data from the 2008 administration for grades 6 and 8. For high school, performance levels were applied to a composite score obtained using the base forms for U.S. history and world history. Only students who took both base forms were included in the calculation. Since the total number of items varied between forms, the equated scale scores were used to calculate the mean, variance, and reliability. Table 6.2 presents descriptive statistics for the three analyses.

Table 6.2  
*Descriptive Statistics for Each Grade Analysis*

<b>Statistic</b>	<b>Grade 6</b>	<b>Grade 8</b>	<b>High School</b>
<i>n</i>	15,024	15,641	8546
Max possible	48	60	100
Mean	28.23	34.30	60.019
Variance	53.379	117.253	312.461
Reliability	0.827	0.900	0.880

### Results

Table 6.3 summarizes the consistency and accuracy results across all five performance level categories for the three grade levels.

Table 6.3  
*Summary of Consistency and Accuracy Results*

<b>Approach</b>		<b>Grade 6</b>	<b>Grade 8</b>	<b>High School</b>
Hanson & Brennan	Consistency	0.60	0.66	0.62
	Accuracy	0.71	0.75	0.71
Livingston & Lewis	Consistency	0.60	0.65	0.63
	Accuracy	0.71	0.75	0.73

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 6.4 presents the results for that binary classification decision.

Table 6.4  
*Summary of Consistency and Accuracy Results of Academic Warning or Approaches Standard versus Meets Standard, Exceeds Standard, or Exemplary*

<b>Approach</b>		<b>Grade 6</b>	<b>Grade 8</b>	<b>High School</b>
Hanson & Brennan	Consistency	0.87	0.90	0.91
	Accuracy	0.91	0.93	0.93
Livingston & Lewis	Consistency	0.87	0.89	0.91
	Accuracy	0.91	0.92	0.93

## 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 vary 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 History and Government Assessments in Spring 2008. In 2007, parallel test forms were constructed for grades 6, 8, and 11 (U.S. history and world history). There were four test forms per grade except for grade 11 which had eight test forms total: four each for U.S. history and world history. 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 ( $\hat{\sigma}_{E \cdot X_p}$ ) was estimated 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  $\mu_x$  is the total test mean, and  $\sigma_x^2$  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 History and Government 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 CSEMs 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 6.5 – 6.8. 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 6.5

Conditional Standard Errors of Measurement (CSEM) for Grade 6 History and Government by Test Form

Form 156				Form 394				Form 542				Form 637			
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	1.9	1	0.9	2	1.9	1	0.9	2	1.9	1	0.9	2	1.9
2	1.3	4	2.7	2	1.3	4	2.7	2	1.3	4	2.7	2	1.3	4	2.7
3	1.6	6	3.3	3	1.6	7	3.3	3	1.6	7	3.3	3	1.6	6	3.3
4	1.8	8	3.8	4	1.8	9	3.7	4	1.8	9	3.8	4	1.8	8	3.7
5	2.0	10	4.2	5	2.0	11	4.1	5	2.0	11	4.2	5	2.0	10	4.1
6	2.2	12	4.5	6	2.1	13	4.5	6	2.2	13	4.5	6	2.1	13	4.5
7	2.3	14	4.8	7	2.3	15	4.8	7	2.3	15	4.8	7	2.3	15	4.8
8	2.4	17	5.1	8	2.4	17	5.0	8	2.4	18	5.1	8	2.4	17	5.0
9	2.5	19	5.3	9	2.5	20	5.3	9	2.6	20	5.3	9	2.5	19	5.3
10	2.7	21	5.5	10	2.6	22	5.5	10	2.7	22	5.5	10	2.6	21	5.5
11	2.7	23	5.7	11	2.7	24	5.7	11	2.7	24	5.7	11	2.7	23	5.7
12	2.8	25	5.9	12	2.8	26	5.8	12	2.8	26	5.9	12	2.8	25	5.9
13	2.9	27	6.0	13	2.9	29	6.0	13	2.9	29	6.1	13	2.9	27	6.0
14	3.0	30	6.2	14	2.9	32	6.1	14	3.0	32	6.2	14	3.0	29	6.1
15	3.0	33	6.3	15	3.0	35	6.3	15	3.0	34	6.3	15	3.0	31	6.3
16	3.1	35	6.4	16	3.1	37	6.4	16	3.1	37	6.4	16	3.1	33	6.4
17	3.1	37	6.5	17	3.1	40	6.5	17	3.1	39	6.5	17	3.1	35	6.5
18	3.2	40	6.6	18	3.1	42	6.5	18	3.2	41	6.6	18	3.1	38	6.5
19	3.2	42	6.7	19	3.2	44	6.6	19	3.2	43	6.7	19	3.2	40	6.6
20	3.2	44	6.7	20	3.2	46	6.7	20	3.2	45	6.7	20	3.2	42	6.7
21	3.2	46	6.7	21	3.2	48	6.7	21	3.2	47	6.8	21	3.2	44	6.7
22	3.3	48	6.8	22	3.2	49	6.7	22	3.3	49	6.8	22	3.2	46	6.7
23	3.3	50	6.8	23	3.2	51	6.7	23	3.3	50	6.8	23	3.2	48	6.8
24	3.3	52	6.8	24	3.2	53	6.8	24	3.3	52	6.8	24	3.2	50	6.8
25	3.3	54	6.8	25	3.2	55	6.7	25	3.3	54	6.8	25	3.2	52	6.8
26	3.3	56	6.8	26	3.2	57	6.7	26	3.3	55	6.8	26	3.2	54	6.7
27	3.2	58	6.7	27	3.2	59	6.7	27	3.2	57	6.8	27	3.2	56	6.7
28	3.2	60	6.7	28	3.2	61	6.7	28	3.2	59	6.7	28	3.2	58	6.7
29	3.2	62	6.7	29	3.2	63	6.6	29	3.2	61	6.7	29	3.2	60	6.6
30	3.2	64	6.6	30	3.1	64	6.5	30	3.2	63	6.6	30	3.1	63	6.5
31	3.1	66	6.5	31	3.1	66	6.5	31	3.1	64	6.5	31	3.1	65	6.5
32	3.1	68	6.4	32	3.1	68	6.4	32	3.1	66	6.4	32	3.1	67	6.4
33	3.0	70	6.3	33	3.0	70	6.3	33	3.0	68	6.3	33	3.0	69	6.3
34	3.0	72	6.2	34	2.9	72	6.1	34	3.0	70	6.2	34	3.0	71	6.1
35	2.9	74	6.0	35	2.9	74	6.0	35	2.9	72	6.1	35	2.9	73	6.0
36	2.8	76	5.9	36	2.8	76	5.8	36	2.8	74	5.9	36	2.8	75	5.9
37	2.7	78	5.7	37	2.7	78	5.7	37	2.7	76	5.7	37	2.7	77	5.7
38	2.7	80	5.5	38	2.6	80	5.5	38	2.7	78	5.5	38	2.6	79	5.5
39	2.5	82	5.3	39	2.5	82	5.3	39	2.6	80	5.3	39	2.5	81	5.3
40	2.4	84	5.1	40	2.4	84	5.0	40	2.4	82	5.1	40	2.4	83	5.0
41	2.3	86	4.8	41	2.3	86	4.8	41	2.3	84	4.8	41	2.3	85	4.8
42	2.2	88	4.5	42	2.1	88	4.5	42	2.2	86	4.5	42	2.1	88	4.5
43	2.0	90	4.2	43	2.0	89	4.1	43	2.0	88	4.2	43	2.0	90	4.1
44	1.8	92	3.8	44	1.8	91	3.7	44	1.8	90	3.8	44	1.8	92	3.7
45	1.6	94	3.3	45	1.6	93	3.3	45	1.6	92	3.3	45	1.6	94	3.3
46	1.3	96	2.7	46	1.3	95	2.7	46	1.3	94	2.7	46	1.3	96	2.7
47	0.9	98	1.9	47	0.9	98	1.9	47	0.9	97	1.9	47	0.9	98	1.9
48	0.0	100	0.0	48	0.0	100	0.0	48	0.0	100	0.0	48	0.0	100	0.0

Table 6.6

*Conditional Standard Errors of Measurement (CSEM) for Grade 8 History and Government by Test Form*

Form 292				Form 316				Form 468				Form 849			
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	1.0	2	1.6	1	0.9	2	1.6	1	0.9	2	1.6	1	1.0	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	2.7	3	1.6	5	2.7	3	1.6	5	2.7	3	1.6	5	2.7
4	1.9	7	3.1	4	1.8	7	3.1	4	1.8	7	3.1	4	1.9	7	3.1
5	2.1	9	3.4	5	2.0	9	3.4	5	2.0	9	3.4	5	2.1	8	3.4
6	2.2	10	3.7	6	2.2	10	3.7	6	2.2	10	3.7	6	2.2	10	3.7
7	2.4	12	4.0	7	2.4	12	3.9	7	2.4	12	4.0	7	2.4	12	4.0
8	2.5	14	4.2	8	2.5	14	4.2	8	2.5	14	4.2	8	2.5	13	4.2
9	2.7	16	4.4	9	2.6	15	4.4	9	2.6	15	4.4	9	2.7	15	4.4
10	2.8	17	4.6	10	2.7	17	4.6	10	2.8	17	4.6	10	2.8	17	4.6
11	2.9	19	4.8	11	2.8	19	4.7	11	2.9	19	4.8	11	2.9	18	4.8
12	3.0	21	5.0	12	2.9	21	4.9	12	3.0	21	4.9	12	3.0	20	5.0
13	3.1	23	5.1	13	3.0	22	5.1	13	3.0	22	5.1	13	3.1	22	5.1
14	3.1	24	5.2	14	3.1	24	5.2	14	3.1	24	5.2	14	3.1	23	5.2
15	3.2	26	5.4	15	3.2	25	5.3	15	3.2	26	5.3	15	3.2	25	5.4
16	3.3	27	5.5	16	3.3	27	5.4	16	3.3	27	5.4	16	3.3	27	5.5
17	3.3	29	5.6	17	3.3	29	5.5	17	3.3	29	5.6	17	3.3	28	5.6
18	3.4	30	5.7	18	3.4	30	5.6	18	3.4	31	5.6	18	3.4	30	5.7
19	3.5	32	5.8	19	3.4	32	5.7	19	3.4	33	5.7	19	3.5	32	5.8
20	3.5	34	5.8	20	3.5	34	5.8	20	3.5	34	5.8	20	3.5	33	5.8
21	3.5	35	5.9	21	3.5	35	5.8	21	3.5	36	5.9	21	3.5	35	5.9
22	3.6	37	6.0	22	3.5	37	5.9	22	3.6	38	5.9	22	3.6	37	6.0
23	3.6	38	6.0	23	3.6	39	6.0	23	3.6	39	6.0	23	3.6	38	6.0
24	3.6	40	6.1	24	3.6	40	6.0	24	3.6	41	6.0	24	3.6	40	6.1
25	3.7	42	6.1	25	3.6	42	6.0	25	3.6	43	6.1	25	3.7	42	6.1
26	3.7	43	6.1	26	3.6	44	6.1	26	3.7	44	6.1	26	3.7	43	6.1
27	3.7	45	6.2	27	3.7	45	6.1	27	3.7	46	6.1	27	3.7	45	6.2
28	3.7	47	6.2	28	3.7	47	6.1	28	3.7	48	6.1	28	3.7	47	6.2
29	3.7	49	6.2	29	3.7	49	6.1	29	3.7	49	6.2	29	3.7	48	6.2
30	3.7	50	6.2	30	3.7	51	6.1	30	3.7	51	6.2	30	3.7	50	6.2
31	3.7	52	6.2	31	3.7	53	6.1	31	3.7	53	6.2	31	3.7	52	6.2
32	3.7	54	6.2	32	3.7	54	6.1	32	3.7	55	6.1	32	3.7	53	6.2
33	3.7	56	6.2	33	3.7	56	6.1	33	3.7	56	6.1	33	3.7	55	6.2
34	3.7	58	6.1	34	3.6	58	6.1	34	3.7	58	6.1	34	3.7	57	6.1
35	3.7	60	6.1	35	3.6	60	6.0	35	3.6	60	6.1	35	3.7	58	6.1

Table 6.6 Continued

*Conditional Standard Errors of Measurement (CSEM) for Grade 8 History and Government by Test Form (Continued)*

Form 292				Form 316				Form 468				Form 849			
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
36	3.6	61	6.1	36	3.6	62	6.0	36	3.6	61	6.0	36	3.6	60	6.1
37	3.6	63	6.0	37	3.6	63	6.0	37	3.6	63	6.0	37	3.6	62	6.0
38	3.6	65	6.0	38	3.5	65	5.9	38	3.6	65	5.9	38	3.6	63	6.0
39	3.5	67	5.9	39	3.5	67	5.8	39	3.5	66	5.9	39	3.5	65	5.9
40	3.5	68	5.8	40	3.5	69	5.8	40	3.5	68	5.8	40	3.5	67	5.8
41	3.5	70	5.8	41	3.4	70	5.7	41	3.4	69	5.7	41	3.5	68	5.8
42	3.4	72	5.7	42	3.4	72	5.6	42	3.4	71	5.6	42	3.4	70	5.7
43	3.3	73	5.6	43	3.3	74	5.5	43	3.3	73	5.6	43	3.3	72	5.6
44	3.3	75	5.5	44	3.3	75	5.4	44	3.3	74	5.4	44	3.3	73	5.5
45	3.2	77	5.4	45	3.2	77	5.3	45	3.2	76	5.3	45	3.2	75	5.4
46	3.1	78	5.2	46	3.1	78	5.2	46	3.1	77	5.2	46	3.1	77	5.2
47	3.1	80	5.1	47	3.0	80	5.1	47	3.0	79	5.1	47	3.1	78	5.1
48	3.0	81	5.0	48	2.9	82	4.9	48	3.0	80	4.9	48	3.0	80	5.0
49	2.9	83	4.8	49	2.8	83	4.7	49	2.9	82	4.8	49	2.9	82	4.8
50	2.8	85	4.6	50	2.7	85	4.6	50	2.8	83	4.6	50	2.8	83	4.6
51	2.7	86	4.4	51	2.6	86	4.4	51	2.6	85	4.4	51	2.7	85	4.4
52	2.5	88	4.2	52	2.5	88	4.2	52	2.5	87	4.2	52	2.5	87	4.2
53	2.4	89	4.0	53	2.4	90	3.9	53	2.4	88	4.0	53	2.4	88	4.0
54	2.2	91	3.7	54	2.2	91	3.7	54	2.2	90	3.7	54	2.2	90	3.7
55	2.1	93	3.4	55	2.0	93	3.4	55	2.0	91	3.4	55	2.1	92	3.4
56	1.9	94	3.1	56	1.8	94	3.1	56	1.8	93	3.1	56	1.9	93	3.1
57	1.6	96	2.7	57	1.6	96	2.7	57	1.6	94	2.7	57	1.6	95	2.7
58	1.3	97	2.2	58	1.3	97	2.2	58	1.3	96	2.2	58	1.3	97	2.2
59	1.0	99	1.6	59	0.9	99	1.6	59	0.9	98	1.6	59	1.0	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 6.7

*Conditional Standard Errors of Measurement (CSEM) for Grade 11 U.S. History by Test Form*

Form 234				Form 812				Form 946				Form 972			
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	4	3.2	1	0.9	3	3.2	1	0.9	3	3.0
2	1.3	6	4.3	2	1.3	7	4.4	2	1.3	7	4.5	2	1.3	6	4.2
3	1.5	10	5.1	3	1.5	11	5.3	3	1.6	10	5.4	3	1.5	10	5.1
4	1.7	13	5.8	4	1.8	14	6.0	4	1.8	14	6.1	4	1.7	13	5.7
5	1.9	16	6.4	5	1.9	18	6.6	5	2.0	17	6.7	5	1.9	16	6.3
6	2.0	19	6.8	6	2.1	21	7.1	6	2.1	21	7.2	6	2.0	19	6.7
7	2.2	22	7.2	7	2.2	25	7.5	7	2.2	24	7.6	7	2.1	22	7.1
8	2.3	24	7.6	8	2.3	28	7.8	8	2.3	28	8.0	8	2.2	25	7.4
9	2.3	27	7.8	9	2.3	31	8.1	9	2.4	31	8.2	9	2.3	28	7.7
10	2.4	30	8.0	10	2.4	34	8.3	10	2.5	34	8.5	10	2.4	31	7.9
11	2.5	32	8.2	11	2.5	38	8.5	11	2.5	38	8.6	11	2.4	34	8.1
12	2.5	35	8.4	12	2.5	41	8.6	12	2.5	41	8.8	12	2.5	37	8.3
13	2.5	38	8.5	13	2.5	45	8.7	13	2.6	45	8.9	13	2.5	40	8.3
14	2.6	41	8.5	14	2.5	48	8.7	14	2.6	48	8.9	14	2.5	44	8.4
15	2.6	44	8.5	15	2.5	52	8.7	15	2.6	52	8.9	15	2.5	47	8.4
16	2.6	48	8.5	16	2.5	56	8.7	16	2.6	55	8.9	16	2.5	51	8.4
17	2.5	51	8.5	17	2.5	59	8.6	17	2.5	59	8.8	17	2.5	54	8.3
18	2.5	55	8.4	18	2.5	63	8.5	18	2.5	62	8.6	18	2.5	58	8.3
19	2.5	58	8.2	19	2.4	66	8.3	19	2.5	66	8.5	19	2.4	61	8.1
20	2.4	62	8.0	20	2.3	70	8.1	20	2.4	69	8.2	20	2.4	64	7.9
21	2.3	66	7.8	21	2.3	73	7.8	21	2.3	72	8.0	21	2.3	68	7.7
22	2.3	70	7.6	22	2.2	77	7.5	22	2.2	76	7.6	22	2.2	71	7.4
23	2.2	73	7.2	23	2.1	80	7.1	23	2.1	79	7.2	23	2.1	74	7.1
24	2.0	77	6.8	24	1.9	84	6.6	24	2.0	83	6.7	24	2.0	78	6.7
25	1.9	80	6.4	25	1.8	87	6.0	25	1.8	86	6.1	25	1.9	81	6.3
26	1.7	83	5.8	26	1.5	91	5.3	26	1.6	90	5.4	26	1.7	84	5.7
27	1.5	86	5.1	27	1.3	94	4.4	27	1.3	93	4.5	27	1.5	87	5.1
28	1.3	89	4.3	28	0.9	97	3.2	28	0.9	97	3.2	28	1.3	91	4.2
29	0.9	93	3.1	29	0.0	100	0.0	29	0.0	100	0.0	29	0.9	95	3.0
30	0.0	99	0.0					30	0.0	99	0.0	30	0.0	99	0.0

Table 6.8

Conditional Standard Errors of Measurement (CSEM) for Grade 11 World History by Test Form

Form 296				Form 319				Form 588				Form 817			
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	3.1	1	0.9	3	3.2
2	1.3	7	4.4	2	1.3	7	4.3	2	1.3	6	4.3	2	1.3	7	4.4
3	1.6	10	5.2	3	1.5	10	5.1	3	1.6	9	5.2	3	1.6	10	5.3
4	1.8	14	5.9	4	1.7	13	5.8	4	1.8	11	5.9	4	1.8	13	6.0
5	2.0	17	6.5	5	1.9	17	6.4	5	1.9	14	6.5	5	2.0	17	6.6
6	2.1	20	7.0	6	2.1	20	6.8	6	2.1	17	6.9	6	2.1	20	7.0
7	2.2	24	7.4	7	2.2	23	7.2	7	2.2	20	7.3	7	2.2	23	7.4
8	2.3	27	7.7	8	2.3	26	7.6	8	2.3	24	7.7	8	2.3	27	7.8
9	2.4	30	8.0	9	2.4	29	7.8	9	2.4	27	7.9	9	2.4	30	8.1
10	2.5	34	8.2	10	2.4	32	8.1	10	2.5	30	8.2	10	2.5	33	8.3
11	2.5	37	8.4	11	2.5	35	8.2	11	2.5	33	8.4	11	2.5	37	8.5
12	2.6	40	8.5	12	2.5	38	8.4	12	2.5	37	8.5	12	2.6	40	8.6
13	2.6	43	8.6	13	2.5	41	8.5	13	2.6	40	8.6	13	2.6	43	8.7
14	2.6	47	8.7	14	2.6	44	8.5	14	2.6	43	8.6	14	2.6	47	8.8
15	2.6	50	8.7	15	2.6	47	8.6	15	2.6	47	8.7	15	2.6	50	8.8
16	2.6	53	8.7	16	2.6	51	8.5	16	2.6	50	8.6	16	2.6	53	8.8
17	2.6	56	8.6	17	2.5	54	8.5	17	2.6	54	8.6	17	2.6	57	8.7
18	2.6	59	8.5	18	2.5	57	8.4	18	2.5	57	8.5	18	2.6	60	8.6
19	2.5	62	8.4	19	2.5	61	8.2	19	2.5	60	8.4	19	2.5	63	8.5
20	2.5	65	8.2	20	2.4	64	8.1	20	2.5	64	8.2	20	2.5	67	8.3
21	2.4	69	8.0	21	2.4	67	7.8	21	2.4	67	7.9	21	2.4	70	8.1
22	2.3	72	7.7	22	2.3	71	7.6	22	2.3	70	7.7	22	2.3	73	7.8
23	2.2	75	7.4	23	2.2	74	7.2	23	2.2	74	7.3	23	2.2	77	7.4
24	2.1	78	7.0	24	2.1	78	6.8	24	2.1	77	6.9	24	2.1	80	7.0
25	2.0	81	6.5	25	1.9	81	6.4	25	1.9	81	6.5	25	2.0	83	6.6
26	1.8	84	5.9	26	1.7	84	5.8	26	1.8	84	5.9	26	1.8	87	6.0
27	1.6	87	5.2	27	1.5	87	5.1	27	1.6	87	5.2	27	1.6	90	5.3
28	1.3	90	4.4	28	1.3	91	4.3	28	1.3	91	4.3	28	1.3	93	4.4
29	0.9	95	3.1	29	0.9	95	3.1	29	0.9	95	3.1	29	0.9	97	3.2
30	0.0	99	0.0	30	0.0	99	0.0	30	0.0	100	0.0	30	0.0	100	0.0

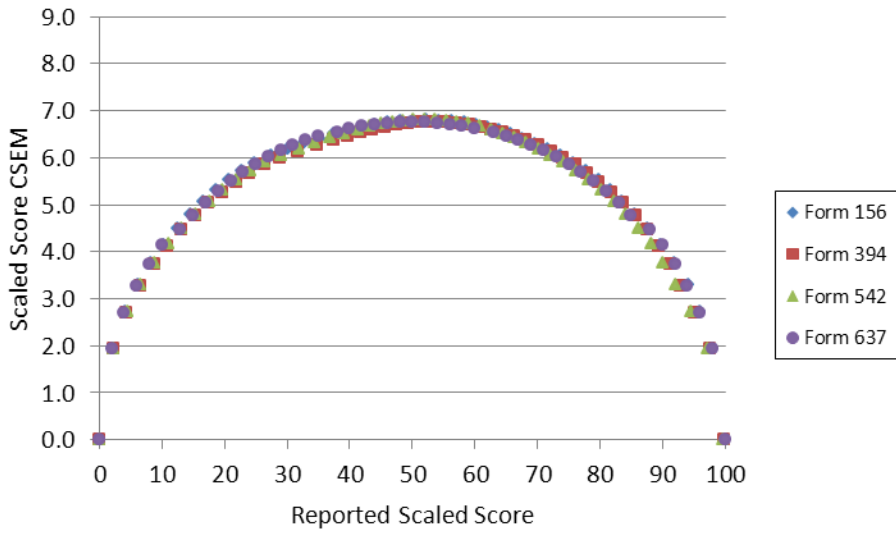


Figure 6.1. Conditional Standard Errors of Measurement (CSEM) for grade 6 history and government by test form.

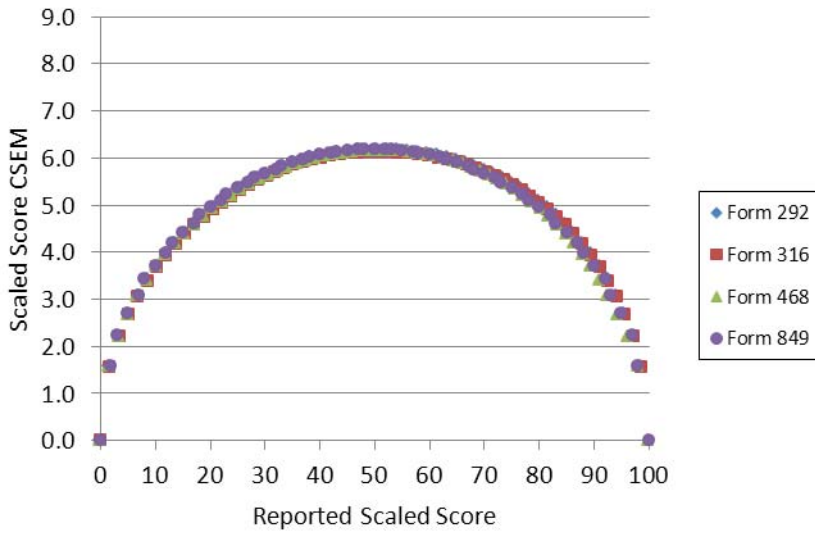


Figure 6.2. Conditional Standard Errors of Measurement (CSEM) for grade 8 history and government by test form.

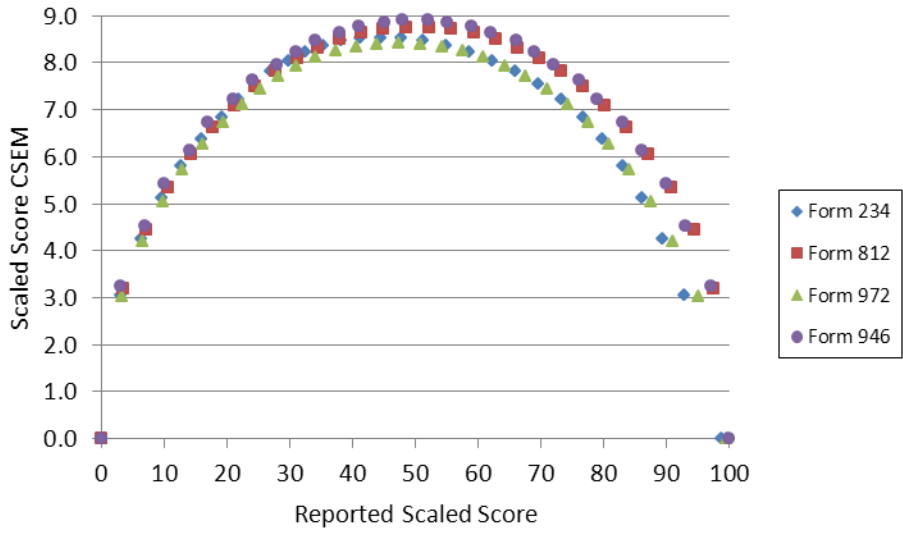


Figure 6.3. Conditional Standard Errors of Measurement (CSEM) for grade 11 U.S. history by test form.

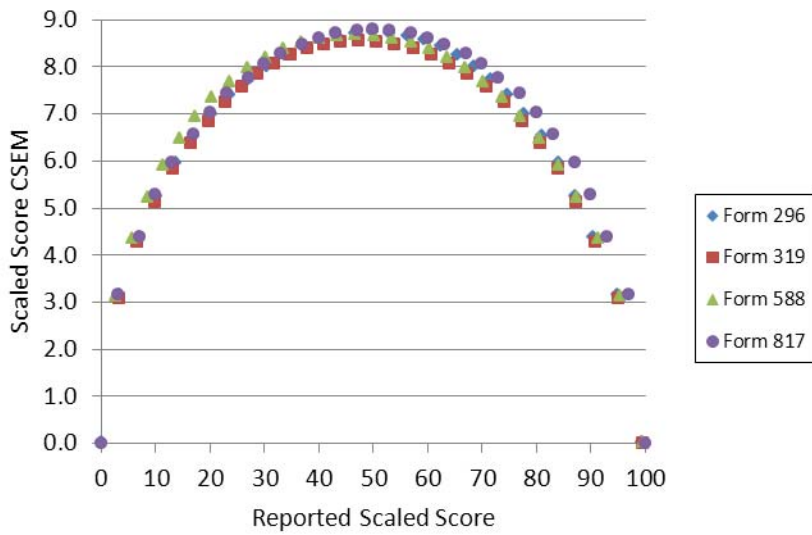


Figure 6.4. Conditional Standard Errors of Measurement (CSEM) for grade 11 world history by test form.

## 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 History and Government Sub-domain Scores at Benchmark Level**

To be an effective history and government test, one must be able to demonstrate that the test, in fact, measures a student’s knowledge of history and government 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 history and government.

### **Procedure**

#### **Sample**

The correlational study among sub-domain scores was based on samples of students who were administered the Kansas general assessments in history and government 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 U.S. history and world history. 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 History and Government*

<b>Grade</b>	<b>Form</b>	<b>Number of Items</b>	<b>Sample Size</b>
6	637	44	15,024
8	849	60	15,641
11	946 (U.S.)	29	8,546
11	817 (World)	30	

#### **Method**

In the Kansas History and Government Assessments, there were four standards on each test form per grade. Each standard tests a specific content area, which include civics-government, economics, geography, and history. These are standards 1, 2, 3, and 4, respectively. Table 7.2 below displays the number of items per standard 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 Standard by Grade Level*

<b>Grade</b>	<b>Civics-Government</b>	<b>Economics</b>	<b>Geography</b>	<b>History</b>
<b>6</b>	8	8	8	24
<b>8</b>	10	10	10	30
<b>11</b>	9	12	10	28

## 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 0.43 to 0.57 in grade 6, 0.52 to 0.68 in grade 8, and 0.57 to 0.66 in grade 11. The greatest correlation for all grades appears between standards 2 (economics) and 4 (history), grade 6 ( $r = 0.57$ ), 8 ( $r = 0.68$ ), and 11 ( $r = 0.65$ ). The smallest correlation for all grades appears between standards 1 (civics) and 3 (geography), with grade 6, ( $r = 0.43$ ), 8 ( $r = 0.52$ ), and 11 ( $r = 0.57$ ).

Table 7.3  
*Correlations among Sub-domain Scores by Standard for Grade 6 History and Government*  
*(n = 15,024)*

	Civics - Government	Economics	Geography	History
Civics - Government Pearson Correlation				
Economics Pearson Correlation	.440**			
Geography Pearson Correlation	.430**	.495**		
History Pearson Correlation	.502**	.569**	.545**	

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

Table 7.4  
*Correlations among Sub-domain Scores by Standard for Grade 8 History and Government*  
 (n = 15,641)

		Civics – Government	Economics	Geography	History
Civics - Government	Pearson Correlation				
Economics	Pearson Correlation	.541**			
Geography	Pearson Correlation	.524**	.645**		
History	Pearson Correlation	.586**	.684**	.676**	

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

Table 7.5  
*Correlations among Sub-domain Scores by Standard for Grade 11 History and Government*  
 (n = 8600)

		Civics - Government	Economics	Geography	History
Civics – Government	Pearson Correlation				
Economics	Pearson Correlation	.588**			
Geography	Pearson Correlation	.565**	.612**		
History	Pearson Correlation	.628**	.646**	.626**	

\*\* . 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. In the case of state assessments, intercorrelations assess if there are meaningful relationships among history and government, mathematics, and reading scores for grade 6. 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 history and government scores’ criterion validity includes assessing the relationships of total scores in the areas of history and government with mathematics and reading for grades 6 and 8. High school values were not included due to inadequate sample sizes in mathematics and reading at grade 11.

The evaluation of the strength of relationships was based on samples of students who were administered any form of the Kansas general assessments, including computer or paper-and-pencil versions. In 2008, parallel test forms of the Kansas general assessment were constructed at grades 6 and 8 for mathematics (N=34,350, N=34,829 respectively) reading (N=34,233, N=34,860, respectively), and history and government (N=35,306, N=35,758, respectively). For each subject, four test forms were given in grades 6 and 8.

Pearson product-moment correlations were calculated using the total score for history and government, mathematics, and reading for grades 6 and 8.

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 and 7.7, and detail the intercorrelations for grades 6 and 8, respectively, including the observed correlations and the correlations corrected for attenuation.

Intercorrelations for grade 6 ranged from 0.615 to 0.716 for observed and 0.704 to 0.781 for corrected. Grade 8 values ranged from 0.661 to 0.732 for observed and 0.724 to 0.806 for corrected.

Table 7.6  
*Intercorrelations for Grade 6*

Grade 6	Observed Correlation		Correlation After Correction for Attenuation	
	Mathematics	Reading	Mathematics	Reading
History and Government	0.615	0.666	0.704	0.774
Mathematics		0.716		0.781

Table 7.7  
*Intercorrelations for Grade 8*

Grade 8	Observed Correlation		Correlation After Correction for Attenuation	
	Mathematics	Reading	Mathematics	Reading
History and Government	0.661	0.729	0.724	0.806
Mathematics		0.732		0.789

## Section 8

### PAPER AND PENCIL VERSUS COMPUTER ADMINISTERED

#### TEST COMPARABILITY STUDIES

##### 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 dual system. Districts have the option of using paper and pencil (P&P) or computerized administration program. In the spring of 2008, an overwhelming majority of the students were tested online: approximately 78% in reading, 80% in mathematics, 82% in science, and 81% in history and 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 optional online testing in Kansas (Glasnapp, Poggio, Yang, & Poggio, 2004; Poggio, Glasnapp, Yang, & Poggio, 2005; Poggio, Glasnapp, Yang, Beauchamp, & Dunham, 2005; Yang, Glasnapp, & Poggio, 2006). In these studies, the conditions and constraints of the testing program 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 for online assessment implementation. Similarly, schools that volunteer students for online testing would have to test selected students with 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 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 history and government. In Kansas, grades 6, 8, and high school are the tested grade levels, and 2008 was the initial year for history and government testing. Thus, longitudinal data do not exist whereby prior year history and government 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 history and government. At grades 6 and 8, 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 (U.S. history and world history) 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 6, one at grade 8, and two at the high school level (one for the U.S. history part test form and one for the world history part test form). Findings were comparable to those obtained in reading, mathematics, and science.

## **Evaluating Construct Equivalency Across Test Mode**

Construct equivalency across test mode is the foremost concern in score comparability studies. For history and government, equivalency was assessed using multi-group confirmatory factor analysis and differential item functioning analysis. Results were similar to those found in reading, mathematics, and science.

## **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 and 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.

Based on the results of the analyses conducted, there does not appear to be sufficient evidence to suggest that the history and government 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% 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

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, D.C.: Author.
- Berk, R. A. (Ed.). (1982). *Handbook of methods for detecting test bias*. Baltimore, MD: The Johns Hopkins University Press.
- 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](http://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* (5<sup>th</sup> 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.
- 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.
- 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.

APPENDIX A  
Tested Indicators/Test Specifications  
Grades 6, 8, U.S. History, and World History

**Sixth Grade Tested Indicators/Test Specifications**

**CIVICS-GOVERNMENT:**

<b>Indicator</b>	<b>Text of Indicator</b>	<b>Number of Items</b>
SS.5.1.2.4K	identifies important founding fathers and their contributions (e.g., ▲George Mason, ▲Thomas Jefferson, ▲James Madison, ▲George Washington, ▲Benjamin Franklin, Thomas Paine, Samuel Adams, John Adams)	2
SS.5.1.3.4A	explains the functions of the three branches of federal government (e.g., legislative-makes laws, executive-enforces laws, judicial-interprets laws)	2
SS.6.1.4.1A	compares and contrasts the rights of people living in ancient Greece (Sparta and Athens) and classical Rome with the modern United States	2
SS.6.1.5.1K	identifies the basic features of systems of government (e.g., republic, democracy, monarchy, dictatorship, oligarchy, theocracy)	2

**ECONOMICS:**

SS.5.2.2.2K	identifies factors that change supply or demand for a product (e.g., supply: technology changes; demand: invention of new and substitute goods; supply or demand: climate and weather)	2
SS.5.2.5.1A	( <b>\$</b> ) determines the costs and benefits of a spending, saving, or borrowing decision	2
SS.6.2.1.1K	explains how scarcity of resources requires communities and nations to make choices about goods and services (e.g., what foods to eat, where to settle, how to use land)	2
SS.6.2.3.2K	identifies barriers to trade among nations (e.g., treaties, war, transportation, geography)	2

**GEOGRAPHY:**

SS.5.3.1.2K	locates major physical and political features of Earth from memory (e.g., ▲Boston, ▲Philadelphia, ▲England, ▲France, ▲Italy, ▲Spain, ▲North America, ▲Atlantic Ocean, ▲Pacific Ocean, Yucatan Peninsula, Germany, Aleutian Islands, Bering Strait, Chesapeake Bay, Hudson Bay, Mexico City, Montreal, Netherlands, Norway, Ohio River, Portugal, Quebec City, St. Lawrence River)	2
SS.6.3.1.1A	explains and uses map titles, symbols, cardinal and intermediate directions, legends, latitude and longitude	2
SS.6.3.2.3K	identifies and describes the location, landscape, climate, and resources of early world civilizations (e.g., ▲Mesopotamia, ▲Egypt, ▲India, ▲China, ▲Greece, ▲Rome, ▲Middle/South America, Western Europe, West Africa, Japan)	2
SS.6.3.4.2K	describes the forces and processes of conflict and cooperation that divide or unite people (e.g., ▲uneven distribution of resources, ▲water use in ancient Mesopotamia, ▲building projects in ancient Egypt and ▲Middle/South America, ▲the Greek city-states, empire building, movements for independence or rights)	2

Indicator	Text of Indicator	Number of Items
<b>HISTORY:</b>		
SS.5.4.1.1K	explains how various American Indians adapted to their environment in relationship to shelter and food (e.g., Plains, Woodland, Northwest Coast, Southeast and Pueblo cultures in the period from 1700-1820)	2
SS.5.4.1.3A	compares the motives and technology that encouraged European exploration of the Americas (e.g., motives: trade, expansion, wealth, discovery; technology: improved ship building, sextant, cartography)	2
SS.5.4.2.3K	describes political and economic structures in the New England, Middle, and Southern Colonies (e.g., political: House of Burgesses, town meetings, colonial forms of representation; economics: agriculture, trade)	2
SS.5.4.3.1K	describes the causes of the American Revolution (e.g., Proclamation of 1763, Intolerable Acts, Stamp Act, taxation without representation)	2
SS.5.4.3.6K	describes how the Constitutional Convention led to the creation of the United States Constitution (e.g., Great Compromise, Three-Fifths Compromise)	2
SS.5.4.4.1A	uses historical timelines to trace the cause and effect relationships between events in different places during the same time period (e.g., colonial America and England)	2
SS.6.4.1.2A	compares the origin and accomplishments of early river valley civilizations (e.g., Tigris and Euphrates (Mesopotamia): city-states, Hammurabi's code; Nile Valley (Egypt): Pharaoh, centralized government; Indus Valley (India): Mohenjo Daro; Huang He (China): Shang Dynasty)	2
SS.6.4.2.1K	compares and contrasts characteristics of classic Greek government (e.g., city-states, slavery, rule by aristocrats and tyrants, Athens: development of democracy, Sparta: city's needs come first)	2
SS.6.4.2.4K	describes key characteristics of classical Roman government (e.g., Roman Republic: senate, consuls, veto, written law; Roman Empire: emperors, expansion)	2
SS.6.4.2.6A	examines the central beliefs of Christianity, Hinduism, Buddhism, Judaism, and Islam	2
SS.6.4.2.9K	describes key accomplishments of ancient China (e.g., Great Wall of China, Shi Huangdi, dynastic cycle, Mandate of Heaven, Taoism, Confucianism, civil service, Silk Road)	2
SS.6.4.4.1A	examines a topic in World history to analyze changes over time and makes logical inferences concerning cause and effect (e.g., spread of ideas and innovation, rise and fall of empires)	2

**Eighth Grade Tested Indicators/Test Specifications**

Indicator	Text of Indicator	Number of Items
<b>CIVICS-GOVERNMENT:</b>		
SS.7.1.1.2A	compares how juveniles and adults are treated differently under law (e.g., due process, trial, age restrictions, punishment, rehabilitation, <i>diversion</i> )	2
SS.7.1.2.1K	defines the <i>rights</i> guaranteed, granted, and protected by the Kansas <i>Constitution</i> and its amendments	2
SS.7.1.5.3K	identifies the <i>goods</i> and <i>services</i> provided by local government in the <i>community</i> (e.g., education, health agency, fire department, police, care for local community property, parks and recreation)	2
SS.8.1.3.3K	explains how the United States Constitution can be changed through amendments	2
SS.8.1.3.4A	analyzes the Declaration of Independence and the United States Constitution to identify essential ideas of American constitutional government	2
<b>ECONOMICS:</b>		
SS.7.2.3.1A	describes examples of factors that might influence <i>international trade</i> (e.g., United States economic sanctions, weather, exchange rates, war, boycotts, embargos)	2
SS.7.2.5.1A	(\$ ) compares the <i>benefits</i> and <i>costs</i> of <i>spending</i> , <i>saving</i> , or <i>borrowing</i> decisions based on information about products and <i>services</i>	2
SS.8.2.1.1A	analyzes the effect of <i>scarcity</i> on the <i>price</i> , <i>production</i> , <i>consumption</i> and <i>distribution</i> of <i>goods</i> and <i>services</i> (e.g., price goes up and production goes down, consumption goes down and distribution is limited)	2
SS.8.2.2.1K	explains how relative <i>price</i> , people’s economic decisions, and innovations influence the <i>market</i> system (e.g., cotton gin led to increased <i>productivity</i> , more cotton produced, higher <i>profits</i> , and lower prices; steamboat led to increased <i>distribution of goods</i> , which brought down prices of goods and allowed goods to be more affordable to people across the United States; development of railroad led to transportation of cattle to eastern markets, price was decreased and profit was increased, timely access to beef)	2
SS.8.2.2.4K	(\$ ) describes the positive and negative <i>incentives</i> to which employees respond (e.g., wage levels, <i>benefits</i> , work hours, working conditions)	2
<b>GEOGRAPHY:</b>		
SS.7.3.2.4K	identifies the various physical and human criteria that can be used to define a region (e.g., physical: mountain, coastal, climate; human: religion, ethnicity, language, economic, government)	2
SS.7.3.4.3K	identifies the geographic factors that influence world <i>trade</i> and <i>interdependence</i> (e.g., <i>location</i> advantage, <i>resource distribution</i> , labor <i>cost</i> , <i>technology</i> , trade networks and organizations)	2
SS.7.3.5.1K	identifies ways in which technologies have modified the physical environment of various world cultures (e.g., dams, levees, aqueducts, irrigation, roads, bridges, plow)	2
SS.8.3.4.1A	evaluates demographic data to analyze population characteristics in the United States over time (e.g., birth/death rates, population growth rates, migration patterns: rural, urban)	2
SS.8.3.4.2A	analyzes push-pull factors including economic, political, and social factors that contribute to human migration and settlement in United States (e.g., economic: availability of natural resources, job opportunities created by technology; political: Jim Crow laws, free-staters; social factors: religious, ethnic discrimination)	2

Indicator	Text of Indicator	Number of Items
<b>HISTORY:</b>		
SS.7.4.1.4A	analyzes the impact of the Indian Removal Act of 1830 on the way of life for emigrant Indian tribes relocated to Kansas (e.g., loss of land and customary resources, disease and starvation, assimilation, inter-tribal conflict)	2
SS.7.4.2.2K	describes how the dispute over slavery shaped life in Kansas Territory (e.g., border ruffians, bushwhackers, jayhawkers, the Underground Railroad, free-staters, abolitionists)	2
SS.7.4.3.1K	describes the reasons for tension between the American Indians and the United States <i>government</i> over land in Kansas (e.g., encroachment on Indian lands, <i>depletion</i> of the buffalo and other <i>natural resources</i> , the Sand Creek massacre, broken promises)	2
SS.7.4.3.5K	describes the reasons for the Exoduster movement from the South to Kansas (e.g., relatively free land, symbol of Kansas as a free state, the rise of Jim Crow laws in the South, promotions of Benjamin “Pap” Singleton)	2
SS.7.4.4.2K	describes the development of Populism in Kansas (e.g., disillusionment with big Eastern business, railroads, government corruption, high debts and low prices for farmers)	2
SS.7.4.5.1A	compares agricultural practices before and after the dust storms of the 1930s (e.g., rotation of crops, shelter belts, irrigation, terracing, stubble mulch)	2
SS.7.4.7.2A	examines different types of <i>primary sources</i> in Kansas history and analyzes them in terms of credibility, purpose, and point of view (e.g., census records, diaries, photographs, letters, <i>government</i> documents)	2
SS.8.4.1.4A	explains the impact of constitutional interpretation during the <i>era</i> (e.g., Alien and Sedition Act, Louisiana Purchase, Marshall Court - <i>Marbury v. Madison</i> , <i>McCulloch v. Maryland</i> (1819))	2
SS.8.4.1.5A	analyzes how territorial expansion of the United States affected relations with external powers and American Indians (e.g., Louisiana Purchase, concept of Manifest Destiny, previous land policies-Northwest Ordinance, Mexican-American War, Gold Rush)	2
SS.8.4.1.6A	explains how the Industrial Revolution and technological developments impacted different parts of American <i>society</i> (e.g., interchangeable parts, cotton gin, railroads, steamboats, canals)	2
SS.8.4.2.3K	retraces events that led to sectionalism and secession prior to the Civil War (e.g., Missouri Compromise, Compromise of 1850, Kansas-Nebraska Act-Popular <i>Sovereignty</i> , <i>Uncle Tom’s Cabin</i> )	2
SS.8.4.2.5K	describes the turning points of the Civil War (e.g., Antietam, Gettysburg, Emancipation Proclamation, and Sherman’s March to the Sea)	2
SS.8.4.2.9A	analyzes the impact of the end of slavery on African Americans (e.g., Black Codes; sharecropping; Jim Crow; Amendments 13, 14, and 15; Frederick Douglass; Ku Klux Klan; Exodusters)	2
SS.8.4.3.2K	explains the impact of the railroad on the settlement and development of the West (e.g., transcontinental railroad, cattle towns, Fred Harvey, town speculation, railroad land, <i>immigrant</i> agents)	2
SS.8.4.4.4A	compares contrasting descriptions of the same event in United States history to understand how people differ in their interpretations of historical events	2

<b>High School Tested Indicators/Test Specifications – United States Assessment</b>		
<b>Indicator</b>	<b>Text of Indicator</b>	<b>Number of Items</b>
<b>CIVICS-GOVERNMENT:</b>		
SS.HS.1.1.2A United States	analyzes how the rule of law can be used to protect the rights of individuals and to promote the common good (e.g., eminent domain, martial law during disasters, health and safety issues)	2
SS.HS.1.2.2K United States	understands core civic values inherent in the United States Constitution, Bill of Rights, and Declaration of Independence that have been the foundation for unity in American society (e.g., right to freedom of speech, religion, press, assembly; equality; human dignity; civic responsibility; sovereignty of the people)	2
SS.HS.1.3.2K United States	explains Constitutional powers (e.g., ▲expressed/enumerated, ▲implied, inherent, ▲reserved, concurrent)	2
SS.HS.1.4.1A United States	examines the role of political parties in channeling public opinion, allowing people to act jointly, nominating candidates, conducting campaigns, and training future leaders	2
SS.HS.1.5.3A United States	examines the purpose and functions of multi-national organizations (e.g., United Nations, NATO, International Red Cross)	2
<b>ECONOMICS:</b>		
SS.HS.2.2.4K United States	explains the factors that could change supply of or demand for a product (e.g., societal values; prohibition of alcohol; scarcity of resources: war; technology: assembly line production)	2
SS.HS.2.4.4A United States	evaluate the costs and benefits of governmental economic and social policies on society (e.g., minimum wage laws, anti-trust laws, EPA Regulations, Social Security, farm subsidies, international sanctions on agriculture, Medicare, unemployment insurance, corporate tax credits, public works projects)	2
SS.HS.2.5.6A United States	( <b>\$</b> ) analyzes the costs and benefits of investment alternatives (e.g., stock market, bonds, real estate)	2
<b>HISTORY:</b>		
SS.HS.4a.2.1A United States	uses primary source materials to explore individual experiences in the Dust Bowl in Kansas (e.g., diaries, oral histories, letters)	2
SS.HS.4b.1.4A United States	examines the emergence of the United States in international affairs at the turn of the 20th century (e.g., debate over imperialism, Spanish-American War, Philippine Insurrection, Panama Canal, Open Door policy, Roosevelt Corollary, Dollar Diplomacy)	2
SS.HS.4b.2.2A United States	analyzes the costs and benefits of New Deal programs. (e.g., budget deficits vs. creating employment, expanding government: CCC, WPA, Social Security, TVA, community infrastructure improved; dependence on subsidies)	2
SS.HS.4b.2.6K United States	discusses how World War II influenced the home front (e.g., women in the work place, rationing, role of the radio in communicating news from the war front, victory gardens, conscientious objectors)	2
SS.HS.4b.3.2A United States	analyzes the origins of the Cold War (e.g., establishment of the Soviet Bloc, Mao’s victory in China, Marshall Plan, Berlin Blockade, Iron Curtain)	2
SS.HS.4b.3.7K United States	examines the struggle for racial and gender equality and for the extension of civil rights (e.g., Brown v. Topeka Board of Education, Little Rock Nine, Martin Luther King, Jr., Montgomery Bus Boycott, Voting Rights Act of 1965, Betty Friedan, NOW, ERA, Title IX)	2
SS.HS.4b.5.3A United States	uses primary and secondary sources about an event in U.S. history to develop a credible interpretation of the event, evaluating on its meaning (e.g., uses provided primary and secondary sources to interpret a historical-based conclusion)	2

**High School Tested Indicators/Test Specifications – World Assessment**

Indicator	Text of Indicator	Number of Items
<b>ECONOMICS:</b>		
SS.HS.2.1.2K World	explains how economic choices made by societies have intended and unintended consequences. (e.g., mercantilism, “planned economy” under Soviet Union, Adam Smith-Invisible hand/Laissez Faire)	2
SS.HS.2.3.2A World	compares characteristics of traditional, command, market, and mixed economies on the basis of property rights, factors of production and locus of economic decision making (e.g., what, how, for whom)	2
SS.HS.2.5.3A World	(\$ ) explains how the demand for and supply of labor are influenced by productivity, education, skills, retraining, and wage rates (e.g., spinning mills and the beginning of the modern factory system, the increased use of machinery throughout the Industrial Revolution, assembly lines)	2
<b>GEOGRAPHY:</b>		
SS.HS.3.1.1K World	locates major political and physical features of Earth from memory and compares the relative locations of those features. Locations will be included in indicator at each grade level (e.g., ▲Beijing, ▲English Channel, ▲India, ▲Iraq, ▲Moscow, ▲Sahara Desert, ▲South Africa, ▲Venezuela, Balkan Peninsula, Berlin, Black Sea, Bosphorus Strait, Euphrates River, Geneva, Hong Kong, Israel, Libya, North Korea, Pakistan, Saudi Arabia, Singapore, South Korea, Suez Canal, Tigris River, Tokyo, Yangtze River)	2
SS.HS.3.2.2A World	analyzes the factors that contribute to human changes in regions (e.g., technology alters use of place, migration, changes in cultural characteristics, political factors)	2
SS.HS.3.4.5K World	gives examples of how cultural cooperation and conflict are involved in shaping the distribution of and connections between cultural, political, and economic spaces on Earth (e.g., cultural: Hindu vs. Muslims in India, political: International Court of Justice and Hong Kong, economic: World Trade Organization)	2
SS.HS.3.5.1A World	examines the impact that technology has on human modification of the physical environment (e.g., over-fishing, logging and mining, construction on flood plains, internal combustion engine, toxic waste)	2
SS.HS.3.5.2A World	examines alternative strategies to respond to constraints placed on human systems by the physical environment (e.g., irrigation, terracing, sustainable agriculture, water diversion, natural disaster-resistant construction)	2
<b>HISTORY:</b>		
SS.HS.4c.1.1A World	analyzes the changes in European thought and culture resulting from the Renaissance (e.g., more secular worldview; Machiavelli, Shakespeare; humanism; innovations in art: Michelangelo, DaVinci, architecture: St. Peters Dome)	2
SS.HS.4c.1.7K World	describes why East Asia withdrew into isolationism during a time of European expansion (e.g., Tokugawa Shogunate, end of Great Ming Naval Expeditions)	2
SS.HS.4c.2.2K World	explains essential concepts from the Enlightenment that represented a turning point in intellectual history (e.g., ideas of Thomas Hobbes, John Locke, Voltaire, Montesquieu, Mary Wollstonecraft, Jean Jacques Rousseau, Enlightened despotism, salons)	2
SS.HS.4c.2.5A World	compares and contrasts German unification with the Meiji restoration (e.g., nationalism, militarism, modernization, industrialization)	2
SS.HS.4c.2.8A World	examines causes of anti-colonial movements in Latin America, Asia, and Africa (e.g., ▲Haitian Revolution; Bolivar; San Martin; Hidalgo and Morelos; Taiping Rebellion; ▲Boxer Rebellion; ▲Sepoy Rebellion; ▲Zulu Wars)	2
SS.HS.4c.3.3A World	examines the nature of totalitarianism in fascist Germany and communist Soviet Union (e.g., one party rule; systematic violation of human rights; secret police; state supremacy over individual rights; role of private property; class structure)	2
SS.HS.4c.4.3K World	describes the emergence of the Middle East as an influential region in world politics (e.g., creation of the state of Israel; emerging Middle Eastern post WWII nationalism: Suez Crisis; petroleum based interdependence)	2

