South Carolina
High School Assessment Program

English Language Arts and Mathematics
2004 Operational Test Technical Report

Prepared by

Bokhee Yoon, Kyunghee Suh, and Karen Thornton

American Institutes for Research

Edited and Issued by the

South Carolina Department of Education

Inez M. Tenenbaum
State Superintendent of Education

January 2006

Contents

Lists of Tables and Figures

TABLES

FIGURES

Chapter 1

HISTORY AND OVERVIEW

The South Carolina Education Accountability Act (EAA) of 1998 mandates that all public school students pass an exit examination as one requirement for earning a high school diploma. The federal No Child Left Behind Act (NCLBA) of 2001 mandates that states assess public high school students' academic achievement in reading, language arts, and mathematics. The High School Assessment Program (HSAP) tests were developed to meet both statutory purposes by serving as a criterion for eligibility to receive a South Carolina high school diploma and as a primary source for reporting the federally mandated data required by the NCLBA.

1.1 Preliminary Identification of Assessment Standards and Measurement Issues

In August 2002, an English language arts group and a mathematics group were convened. Each group consisted of State Department of Education (SDE) content, curriculum, and assessment staff members; regional curriculum specialists from South Carolina school districts; and AIR and Insite, Inc., content specialists met to review the curriculum standards and identify the content appropriate for testing. These two groups also identified issues related to assessing the standards with multiple-choice, constructed-response, and extended-response items and discussed a range of assessment issues.

The assessment standards were drafted on the basis of the South Carolina curriculum standards for grades nine through twelve and the foundation skills found in the curriculum standards for grades seven and eight.

Mathematics

In the selection of an appropriate subset of the South Carolina mathematics curriculum standards for grades nine through twelve, the following issues were considered:

- The content must be appropriate for grades ten through twelve (i.e., "high school").
- The test must support at least three levels of achievement—two of which must be *proficient* and above.
- The students being assessed must have had the opportunity by grade ten to learn the content assessed on the test.
- All students who are in their second year of high school after their initial enrollment in the ninth grade will be assessed. Therefore, the test must contain items that measure the range of achievement that can be demonstrated by South Carolina students.
- Students being assessed are currently enrolled in a variety of mathematics classes (Algebra 1, Mathematics for the Technologies 2, Geometry, Algebra 2, etc.).

The August 2002 mathematics group reviewed the South Carolina mathematics curriculum standards for grades nine through twelve and identified the content for the new high school examination that South Carolina students would have an opportunity to learn by the tenth grade. The curriculum standards for grades seven and eight were reviewed for the purpose of

identifying foundation skills implied by the standards selected for inclusion on the new high school examination.

Once test content had been identified, work began on the measurement guidelines. The measurement guidelines contained the item specifications that were used to guide item writing and review. In addition to the item characteristics, the group agreed on the following general characteristics for the test instrument:

- The test must contain multiple-choice items.
- The test must contain integrated-response items that focus on mathematical processes.
- The test must contain real-world applications whenever possible.
- The geometry items must reflect middle school foundation skills to ensure that students have received the opportunity to learn the tested material.
- The test items for each content domain must be written at all levels of difficulty.

English Language Arts

The August 2002 English language arts group analyzed the South Carolina English language arts curriculum standards for grades seven through twelve. Standards were identified as appropriate either for measurement on the test or for measurement in the classroom.

Some content domains were deemed inappropriate for a high-stakes test. For example, "Listening" was seen as an area that could not be fairly assessed on such a test.

The participants agreed that multiple-choice and constructed-response items are appropriate for measuring English language arts content knowledge. The group agreed on the following general features of the operational test:

- An operational test form should contain approximately 60 multiple-choice items.
- Constructed-response reading items must also be included on the test.
- Informational and literary texts representing varying levels of complexity must be included on the test.
- One extended-response writing prompt must be included on the test.
- The same 15-point scoring rubric used on the Palmetto Achievement Challenge Tests (PACT) must be used for the extended-response prompt.

1.2 DEVELOPMENT OF PRELIMINARY TEST AND ITEM SPECIFICATIONS

Test and item specifications were developed to ensure alignment of the test items with the identified curriculum standards. In late September and early October 2002, two content review committees (CRCs) comprising South Carolina educators were convened to review the proposed assessment standards, resolve measurement issues, develop and refine the item specifications, and make preliminary decisions about allocations of the items to the testing domains and assessment standards.

The meeting dates were as follows:

- mathematics CRC, September 24 and 25, 2002
- English language arts CRC, October 1 and 2, 2002

To assist with the identification of standards to be assessed, CRC members were provided with the curriculum standards, proposed measurement guidelines and proposed test/item specifications, sample test items coded to these guidelines and specifications, and a taxonomy of process levels. The two CRCs were led through these preliminary guidelines and specifications, standard by standard. The purpose of the sample item set was to provide committee members with an idea of how the standard might be assessed. General measurement issues and standard-specific issues were reviewed by the CRCs using the set of sample test items aligned with the standards. The CRCs provided feedback to the SDE regarding appropriate and inappropriate item types and eligible process levels for AIR to use in item writing.

During this activity, CRC discussion was documented for use in preparing preliminary measurement guidelines and test/item specifications. Each CRC's comments were recorded and included in separate annotated versions of the measurement guidelines and specifications.

Finally, the committee members made recommendations on the allocation of items to each outcome to be measured and then recorded their judgments on the item allocation form. These recommendations were used to guide the SDE in making the final determination on item allocation.

Following this process, the measurement guidelines and specifications were updated and were finalized during the December 2002 CRC item-review meetings.

Stakeholder Meetings

As part of the process of identifying standards appropriate for assessment by the new high school test, the SDE held regional stakeholder meetings during October and November 2002 for the following two purposes:

- to inform the public and the education community about the state and federal requirements that served as the basis for conceptualizing a new examination for high school students and

- to give the public and the education community the opportunity to provide the SDE with feedback on those English language arts and mathematics skills that were considered essential requirements for a student to receive a state high school diploma.

Because the new exit examination is a legislative requirement, the meetings were not used as forums to debate whether the HSAP should be implemented.

Insite, a subcontractor to AIR, partnered with the SDE and AIR to coordinate the logistics at the meeting sites, coordinate public notification of the meetings, invite selected stakeholder groups, determine the presentation format, and develop the review forms of proposed test content.

Seven regional meetings were initially scheduled:

October 7	Richland Northeast High School Auditorium, 7500 Brookfield Road, Columbia
October 10	Charleston Convention Center, 5001 Coliseum Drive, Charleston
October 14	Horry County School District Office, 1605 Horry Street, Conway
October 15	Aiken High School Gymnasium, 449 Rutland Drive, Aiken
October 16	Lexington High School Auditorium/Cafeteria, 2463 Augusta Highway, Lexington
October 17	Rock Hill School District Office, Training Room, 660 North Anderson Road, Rock Hill
October 24	Golden Strip Career Center, Greenville County School District, 1120 East Butler Road, Greenville

Subsequently, based upon district requests, three additional stakeholder meetings were added by the SDE in the following locations: Beaufort, Florence, and Greenwood.

One consistent comment from these meetings was that the test should include "real life" skills. The major stakeholder comments were incorporated in the measurement guidelines and item writing.

1.3 DEVELOPMENT OF THE FIELD-TEST ITEM POOL

Writing original items and augmenting them with appropriate items from the SDE's existing item pool was considered the most efficient and effective approach for producing the required number of items.

Development of ELA Selections

As an initial step in ELA item development, AIR used the SDE's guidelines to create reading and writing selections that were reviewed by a sensitivity review committee (SRC) prior to the development of the actual test items. A meeting of the SRC was held in October to identify any selections that might contain language or subject matter that was potentially offensive to a particular subgroup of students and/or that could give an advantage to a particular subgroup.

Committee members were provided with AIR's guidelines for bias, sensitivity, and language simplification; AIR staff then trained the committee members in the appropriate use of those guidelines. Committee members next evaluated each selection for its appropriateness for South Carolina's high school population. Accepted selections were made available for item development.

Development of ELA and Mathematics Test Items

Using the preliminary test and item specifications, AIR staff trained qualified item writers, each of whom had prior item-writing experience: either they had been previously trained at AIR item-writing workshops or they had been trained elsewhere in multiple-choice, constructed-response, and extended-response item writing. A content-area assessment specialist worked with the item writers to explain the purpose of the assessment, to review measurement practices in item writing, and to interpret the meaning of the assessment standards and the measurement guidelines. Sample items that had been used during the CRC meetings served as models for the

writers to use in creating items to match the standards. To ensure that the items tapped a range of difficulty and taxonomic levels as required by the SDE, item writers used a method based on Bloom's taxonomy to develop item types incorporating a variety of cognitive processing levels—from comprehension to evaluation.

Item writers were also trained using item review criteria as a guide. These criteria were developed into a checklist to be used throughout the writing and review process. Writers followed the procedure of drafting items, receiving feedback from the assessment specialist, and then revising and submitting final drafts.

Approximately 1,500 items were created for each content area. After the items were written, AIR and Insite content and assessment specialists reviewed them. Insite has extensive experience in item development in South Carolina. Items were reviewed independently for alignment with the assessment standards and the item review criteria. AIR and Insite content and assessment specialists discussed issues and revised items as needed.

Pilot Test

Insite, in partnership with AIR, conducted a pilot test in November 2002 to "try out" item types and administration procedures.

The pilot test was administered to eleventh-grade students during one class period throughout the week of November 18–22, 2002. Eleventh-graders were chosen in order to avoid exposing the items to tenth-grade students who would be participating in the spring 2003 field test.

Four secondary schools—three of which used block scheduling—were chosen for the pilot test administration. These particular schools were selected on the basis of their geographic location and the diversity of their student population and, more specifically, on the basis of their Basic Skills Assessment Program (BSAP) high school scores and their location characteristics (i.e., urban, rural, and suburban). Three classes at each of the four secondary schools were selected to participate in the testing, with each class taking either an English language arts or a mathematics pilot form. Each test form was administered to approximately 25 students, for a total of 400 students.

Four mathematics pilot forms and eight English language arts pilot forms were spiraled across classrooms and administered to students to ensure that a range of students responded to each form. Each test form was administered to approximately 25 students, for a total of 400 students.

For English language arts, a combination of Tech Prep and College Prep students were administered each test form. For mathematics, one Tech Prep class, one College Prep class, and two classes comprising Tech Prep and College Prep students participated in the pilot. Of the eight English language arts, two forms contained constructed-response reading items, and six forms included one extended-response writing item. Two of the eight English language arts forms were spiraled and administered in each classroom.

Two versions of each of the four mathematics forms were developed using a counterbalanced design so that potential presentation differences between the multiple-choice and gridded-response items could be determined. Therefore, the two counterbalanced versions of the same form were spiraled within a classroom.

On each counterbalanced version of a mathematics form, all students within each classroom were given the same stem, stimulus, or both, but in a different response format. For example, multiple-choice items in version A were in a gridded-response format in version B, while the gridded-response items in version B were in multiple-choice format in version A. Each counterbalanced version of a form also contained the same two constructed-response items.

The results of this pilot test were summarized and sent to the SDE. As a result of the pilot test, the SDE decided not to use gridded-response items.

Item Review

Various groups examined the content validity and potential bias issues of the items in the item pool. Once the newly developed items were reviewed and approved by AIR, they were submitted to the SDE for review. AIR incorporated the SDE's revisions to the items. Once the field-test item pool was reviewed and approved by SDE measurement and curriculum staff, the items were reviewed both by the CRCs and by the SRC. These committees convened in December 2002.

The CRCs consisted of expert representatives, including SDE mathematics and English language arts curriculum and assessment staff, district-level curriculum specialists, and other South Carolina educators. The SRC consisted of individuals from a variety of organizations including the South Carolina Commission for Minority Affairs, women's study programs, and the South Carolina Association for Rural Education as well as school counselors and individuals from groups representing persons with disabilities. Several of the members had served on previous SDE bias and sensitivity review committees. The CRCs and the SRC played an integral role in ensuring that the alignment of the test items with the assessment standards as well as ensuring the appropriateness of the test content.

After a general introductory session, each of the two CRCs conducted a three-day review of the test items for its content area. Every CRC member was given a secure spiral-bound volume containing a representative sample of the field-test items to be reviewed. In each CRC, the content leader discussed the items in sets, grouped by domain, using the measurement guidelines and test/item specifications they had discussed previously. The CRCs used the related South Carolina curriculum standards to review the content that each item measured. Participants applied the item review criteria and voted individually either to keep, to revise, or to reject each item. Once all votes were registered, group leaders led discussion on those items for which consensus was not 100 percent; recommendations for revising items as well as recommendations for revisions to the measurement guidelines and test/item specifications were recorded for further review by the SDE.

Following the CRC meetings in December, the SRC met to review the mathematics and English language arts test items. The SRC members reviewed all the items using AIR's guidelines for bias, sensitivity, and language simplification. AIR leaders outlined the purpose of the meeting, discussed the guidelines, and worked through a few of the items aloud with the group before asking members to review items on their own. Members were reminded to concentrate on sensitivity issues rather than subject-matter content. After the committee members had completed their individual reviews, the group convened to discuss any items they identified as potentially problematic. The leaders used the documentation from the CRC meetings to inform the SRC members of any items that had been revised or deleted by the CRCs. SRC members reached consensus on revisions related to bias, sensitivity, and language simplification (e.g.,

change of context, simplification of sentence structure for clarity), and the leaders recorded comments and recommendations throughout the meeting. Items were identified for changes in wording, often to reflect regional usage or to improve the plausibility of a problem's context.

Following the CRC and SRC meetings, AIR content specialists worked with SDE staff to revise items and update the measurement guidelines and the field-test blueprint.

Development of Field-Test Forms

At the conclusion of the item-review process, field-test forms for both mathematics and English language arts were constructed from the pool of items that were approved during the internal and external review processes. These items measured the specified assessment standards that had been approved by the SDE.

There were no item statistics to guide the partitioning of items across forms. The field-test forms were constructed on the basis of the test blueprints in tables 1.1 and 1.2, and each field-test form contained anchor items.

In mathematics, each field-test form contained 77 multiple-choice items and 3 constructed-response items. A set of multiple-choice anchor items was embedded on each form. The anchor items, which were placed in the same position on all forms, were selected to ensure appropriate content representation. In English language arts, each field-test form contained 80 multiple-choice items, 2 constructed-response items, and 1 extended-response writing prompt.

TABLE 1.1

Mathematics Field-Test Blueprint

Domain	Assessment Standard	Average Number of Items per Form
Number and Operations	N 1	9
	N 2	10
	Anchor Items	3
Algebra	A 1	10
	A 2	14
	Anchor Items	5
Measurement and Geometry	MG 1	10
	MG 2	14
	Anchor Items	5
Data Analysis and Probability	DP 1	10
	Anchor Items	2
Integrated Responses		3

TABLE 1.2

English Language Arts Field-Test Blueprint

Domain	Assessment Standard	Average Number of Items per Form
Reading Comprehension	R1	31
	Anchor Items	5
Analysis of Text	R2	22
	Anchor Items	5
Word Study and Analysis	R3	10
	Anchor Items	2
Writing	W4	13
	Anchor Items	2
Research	RS	7
	Anchor Items	1

Once the test forms were constructed, they were sent to the SDE for revisions and approval. SDE staff reviewed field-test forms at the black-line and blue-line stages of production in February and March 2003, respectively. After AIR had incorporated the revisions to the field-test booklets as requested by the SDE, the test received final SDE approval. The field-test forms were administered to all eligible South Carolina students from April 29 to May 7, 2003, including makeup days. The English language arts test was given over two days. The mathematics test was administered on one day.

Eligible South Carolina students for the field testing were tenth-grade students who were enrolled in their second year of high school after their initial enrollment in the ninth grade. In order to meet NCLBA regulations, which required reporting group scores in 2003 and using the scores to determine AYP (adequate yearly progress), the 2003 field test was designed as a census field test. Administration and scoring procedures were designed to be the same as the administration procedures for the operational test (refer to chapters 3 and 4). The first HSAP operational test to be used as a graduation requirement was administered in spring 2004.

The spring 2003 field-test administration was designed to produce a sufficient number of items to build pre-equated operational test forms for both mathematics and English language arts (ELA). This goal was achieved for mathematics, and its first operational form was administered in spring 2004; however, for ELA, the spring 2003 field test did not result in enough items to produce the required number of pre-equated operational test forms. To fill the "gaps," additional item development was undertaken, and additional items were field tested with the ELA operational form administered in spring 2004.

Two field-test designs were adopted for the 2004 ELA administration. The first design added 10 field-test items to the operational test form. Eight sets of ten items were appended to the base form, resulting in the administration of a total of eight test forms and 80 field-test items. These added items did not count toward student scores. The second design was an embedded "operational" field-test design.

In order for the spring 2004 ELA assessment itself to meet the test blueprint requirements, it was necessary to conduct an embedded field test. Therefore, field-test items were included as part of the base (operational) test form. More items were field tested than were required to meet the test blueprint, and following analysis, the best set of embedded items was included in determining student scores if warranted. Seven field-test items were embedded in the operational form, 5 of which were used in calculating student scores.

Each ELA form comprised 56 operational items, 7 embedded field-test items, and 10 additional field-test items. The 56 operational and 7 embedded field-test items were common to all forms; however, the 10 additional field-test items were unique to each of the eight forms.

As stipulated in the test blueprint, 57 items were used to generate student ELA scores. This number was determined by starting with the 56 operational items, subtracting 4 items that were not scored, and adding the 5 embedded items that were scored.

Using spring 2004 HSAP test administration data, the preliminary cut scores that were recommended as a result of the standard-setting workshops in July 2003 were reviewed and finalized at the Technical Advisory Committee (TAC) meeting June 30, 2004.

This technical report summarizes the results of statistical and psychometric analyses performed on the spring 2004 operational data for mathematics and English language arts. In this report, all data are based on students in the regular schools only; students in adult education and district-approved home schools were excluded. For clarity, adult education and home school students were not included in statewide aggregate reports; they were, however, included in district and school aggregate reports and in all data files, and they received individual score reports.

Chapter 2

STUDENT DEMOGRAPHICS

2.1 STUDENT PARTICIPATION

For the spring 2004 HSAP administration, all students who were enrolled in their second year of high school after their initial enrollment in the ninth grade were required to take the HSAP mathematics and English language arts tests. Demographic data were collected for each student. These data included the categories of gender, race/ethnicity, grade, language fluency (i.e., LEP—limited English proficiency), lunch program participation, disability status, and migrant status. Table 2.1 presents the student participation in the spring 2004 HSAP administration by the demographic variables.

TABLE 2.1

Summary of Student Demographics in the Sample

Demographics	Mathematics		English Language Arts	
	N	%	N	%
All Students	52,913		53,222	
Gender				
Female	25,956	49.05	26,073	48.99
Male	26,242	49.59	26,408	49.62
Invalid	715	1.35	741	1.39
Ethnicity				
African American	21,450	40.54	21,603	40.59
African American/American Indian	79	0.15	80	0.15
American Indian	95	0.18	95	0.18
Asian	476	0.90	476	0.89
Hawaiian-Pacific Islander	50	0.09	50	0.09
Hispanic	1,144	2.16	1,150	2.16
White	28,663	54.17	28,781	54.08
White/African American	77	0.15	78	0.15
White/American Indian	64	0.12	65	0.12
White/Asian	60	0.11	60	0.11
Other	87	0.16	89	0.17
Invalid	668	1.26	695	1.31
Grade				
09	9,028	17.06	9,176	17.24
10	43,627	82.45	43,777	82.25
11	182	0.34	192	0.36
12	11	0.02	11	0.02
Invalid	65	0.12	66	0.13
Language				
English Speaker	52,132	98.52	52,440	98.53
Full LEP	436	0.82	428	0.80
LEP mainstream	108	0.20	114	0.21
Waiver	31	0.06	31	0.06
Exited	205	0.39	208	0.39
Unknown	1	0.01	1	0.00

TABLE 2.1
Summary of Student Demographics in the Sample

Demographics	Mathematics		English Language Arts	
	N	%	N	%
Lunch Program				
No free/reduced lunch	31,515	59.56	31,638	59.45
Free lunch	18,042	34.10	18,205	34.21
Reduced lunch	3,355	6.34	3,378	6.35
Unknown	1	0.00	1	0.00
IEP				
No	46,218	87.35	46,416	87.21
Yes	6,641	12.55	6,749	12.68
Unknown	54	0.10	57	0.11
Migrant				
No	52,874	99.93	53,183	99.92
Yes	38	0.07	38	0.07
Unknown	1	0.0	1	0.00

2.2 ACCOMMODATIONS AND MODIFICATIONS

Supplemental information regarding the administration of the HSAP to students with disabilities is provided in the *HSAP Test Administration Manual* (*TAM*) (SDE 2004a). The *TAM* provides guidelines for individualized education program (IEP) teams in making decisions about testing students with disabilities, and it outlines specific information regarding testing accommodations, testing modifications, test forms and materials, and administration procedures. A student with a documented disability is one who has been evaluated and found to meet the eligibility criteria for enrollment in special education as defined by the Individuals with Disabilities Education Act of 1997 and South Carolina State Board of Education Regulation 43-243.1 or is one who has a disability covered under Section 504 of the Rehabilitation Act of 1973.

The IEP or 504 Plan team determines how a student with disabilities participates in the HSAP assessments. Decisions about accommodations, modifications, and alternate assessment must be made on an individual student basis and not on the basis of the category of disability.

Accommodations

Accommodation is defined as a change in the testing environment, procedures, or presentation that does not alter what the test measures or the comparability of scores. The purpose of accommodations is to enable students to participate in an assessment in a way that allows knowledge and skills, rather than disabilities, to be assessed.

Examples of the accommodations that were allowed during the 2004 HSAP administration are changes in the test setting, timing, and scheduling: students were allowed to take the test in a different setting such as in a small group or individually as opposed to taking it with their class, students were allowed extended amounts of time to complete the test, and students were allowed to take the test over several days or periods during the day with frequent breaks. These are all general types of accommodations, and they can vary widely from child to child, according to what is specified in the IEP. Other accommodations allowed were the use of a poor speller's

dictionary (e.g., *The Misspeller's Dictionary*) for the ELA test, oral and signed administrations of the mathematics test, and the use of customized test materials (see section 3.4 below for more detail) such as loose-leaf test booklets, large-print test booklets, and braille and a regular-print Form C test booklets for both tests.

Modifications

Modification is defined as a change in the testing environment, procedures, or presentation that changes the meaning of the test scores. Modifications compromise the test validity and may alter the meaning and comparability of test scores.

The 2004 administration of the ELA test incorporated all of the State Department of Education–approved modifications, such as oral administration, signed administration, alternative scoring for extended-writing responses, and extended-writing options. The alternative scoring rubric was slightly different from the regular scoring rubric. If an alternative scoring accommodation was marked on a student's answer document, the extended-writing response (ER) was supposed to be scored using the alternative scoring rubric. During the spring 2004 hand scoring, it was decided that the regular scoring rubric would be used for all ER papers and that revised score reports would be issued after the ER papers had been rescored on the basis of the alternative scoring rubric. The data presented in this technical report include only the extended-response papers that were scored on the basis of the regular scoring rubric.

If a student received a test modification, the modification was noted on both the roster reports provided to the schools and districts and on the individual score reports. The summary results include scores for students who used modifications. Tables 2.2 and 2.3 present summaries of accommodations and modifications.

TABLE 2.2

Accommodations

Accommodation	Mathematics		English Language Arts	
	Form 30A (N = 50,128)	Customized Form (N = 2,785)	Form 30A (N = 50,251)	Customized Form (N = 2,971)
Setting	2.8*	63.6*	3.0*	64.2*
Presentation	0.0	18.9	0.1	21.0
Timing	0.3	9.7	0.4	10.1
Schedule	0.1	4.8	0.1	5.6
Response options	0.1	1.7	0.1	4.8
Loose leaf	0.0	2.7	0.0	3.4
Large print	0.0	0.9	0.0	1.2
Spelling	—	—	0.1	10.3
Oral administration	0.2	67.0	—	—
Signed administration	0.0	0.9	—	—
Braille	—	0.1	—	0.1
Form C	0.0	39.6	0.0	42.6
Other	0.1	0.7	0.1	0.8

* Percentages of total responses in column may exceed 100 percent because some students received accommodations of more than one category.

TABLE 2.3

Modifications for English Language Arts

Modifications	Form 30A (N = 50,251)	Customized Form (N = 2,971)
Alternative scoring	0.8*	34.1*
Extended writing options	0.1	4.8
Oral administration	0.2	72.9
Signed administration	0.0	1.0
Other	0.0	0.5

* Percentages of total responses in column may exceed 100 percent because some students received modifications of more than one category.

2.3 TEST ADMINISTRATION TIME

In addition to their demographic information, students were asked to record the times they started and finished the tests. In ELA, students recorded the times for sessions 1 and 2. These times were scanned, and the total testing time was calculated.

Approximately 5 percent of the students in mathematics, 4 percent of the students in ELA session 1, and 5 percent in ELA session 2 either left one or both time records blank or recorded invalid values. Consequently, it was not possible to calculate a total testing time for these students. Approximately 92 percent of the students took two hours and thirty minutes or less to finish the mathematics test. Approximately 90 percent of the students in session 1 and 88 percent of the students in session 2 finished the ELA test within two hours. Table 2.4 exhibits the results of this calculation.

TABLE 2.4

Time Taken

Time Taken	Mathematics (N = 52,913)	English Language Arts (N = 53,222)	
		Session 1	Session 2
0:15	0.3*	0.5*	0.4*
0:30	0.8	3.2	1.3
0:45	3.8	11.9	6.9
1:00	13.8	22.4	19.7
1:15	19.8	21.4	23.2
1:30	19.7	15.5	17.9
1:45	14.7	9.4	11.4
2:00	10.4	5.5	7.0
2:15	5.5	2.8	3.5
2:30	3.1	1.4	1.7
2:45	1.5	0.6	0.8
3 hours or more	1.9	1.1	1.2
Invalid**	4.5	4.3	4.9

* percentage of total responses in column

** The term "invalid" includes responses with no mark or double marks on start and stop time fields. Therefore, it was not possible to compute the difference between start and stop times.

2.4 STUDENT QUESTIONNAIRES

After the test administration, students were instructed to answer 19 questions for mathematics and 11 questions for ELA on the HSAP student questionnaire. The questionnaire topics encompassed test difficulty, class length, classroom activities, and calculator use (mathematics only).

Chapter 3

TEST ADMINISTRATION

3.1 TEST ADMINISTRATION WINDOW

The English language arts operational test was conducted in two sessions over two days, April 20 and 21, 2004. The mathematics test was conducted April 22, 2004. The HSAP makeup test window was April 23–30, 2004.

District test coordinators (DTCs) were instructed to administer makeup tests to all eligible students. The administration of one test per day was recommended; however, DTCs were advised that students could take both tests on one day if necessary.

3.2 TIMING OF THE TEST

The HSAP tests were not timed; however, each session had to be completed during a single day (unless a student's IEP or 504 Plan specifically stated that the student required administration over several days). The following time *estimates* were provided to districts and schools for scheduling purposes only:

English language arts, session 1.....2 hours

English language arts, session 2.....2 hours

Mathematics.....3 hours

Procedures were provided in the administration manuals for accommodating students who needed additional time to finish the operational test. Test administrators (TAs) were instructed to give these students as much time as they needed to finish the examination, provided school staff and space were available.

3.3 ADMINISTRATION MANUALS

Working with SDE staff, American Institutes for Research (AIR) staff drafted the administration manuals for the test. SDE staff reviewed and revised the manuals, and AIR finalized and printed them. Two types of manuals were produced for the HSAP tests: the *HSAP Test Administration Manual* (*TAM*) (SDE 2004a) and the *HSAP District Test Coordinator's Supplement* (SDE 2004b). The supplement included only the information that district test coordinators (DTCs) needed for the administration of the HSAP tests. The *TAM* contained the information that school test coordinators (STCs), TAs, and monitors needed in order to administer the tests to students in their schools.

For this administration, the *TAM* included additional graphics, as was suggested on the comment forms returned from the previous administration. Also, Appendix C in the *TAM* was revised to include a more detailed description of customized materials available. Graphics were newly added to clarify issues such as how to complete student demographic information and how to returning scorable and nonscorable test materials. There were also new tables showing the types of customized materials that are available for testing and the specific types of materials that students with disabilities require in order to be able to take the test.

3.4 CUSTOMIZED MATERIALS

Customized versions of the tests were available for both ELA and mathematics. Six different customized formats of the HSAP tests were available for this administration.

- Loose-leaf test booklets—printed single-sided and bound in three-ring binders—allowed individuals to remove the pages so that they could write or type answers to the constructed-response and extended-response items.
- Large-print booklets could be used for students who have difficulty reading text in a standard-size font. The large-print version was printed in a 9 x 12-inch spiral-bound booklet in an 18-point sans serif font.
- Braille booklets, produced for students who typically read classroom materials in braille, were printed as 11½ x 11-inch interpoint braille pages and bound in three-ring binders.
- A regular-print Form C test booklet was provided in test packets for students or TAs to use with other customized formats such as the oral script/audiotape; braille, large-print, and loose-leaf versions; and sign language videotapes. These booklets were saddle-stitched and printed in a 12-point font, just as the regular, noncustomized test booklets were.
- Oral administration scripts and audiotapes were provided for students whose 504 and IEP plans were written to require oral administration of tests. Scripts provided the directions to TAs regarding the appropriate way to read test questions, passages, and some answer choices to the students. Audiotapes were used for students testing individually or in small-group settings.
- Sign language videotapes were also produced and included the signed test directions, test questions, and some answer choices. The videotapes were produced in three languages: American Sign Language, Pidgin Signed English, and Signed Exact English.

3.5 PRETEST WORKSHOPS AND TRAINING

Pretest workshops were held February 23–24, 2004, in Columbia, South Carolina, to train the DTCs and some STCs. DTCs could bring up to three additional representatives to the workshop. SDE staff and AIR staff trained the district staff in attendance.

AIR was allotted approximately an hour and a half to review the HSAP manuals, security procedures, and any other pertinent information, including an in-depth review of the newly revised instructions for administering tests to students with disabilities.

The DTC supplements and the administration manuals were handed out to the coordinators during the workshop. The DTCs also received printed copies of the PowerPoint Presentation. In addition, the presentation was posted to the SDE Web site.

DTCs were instructed to train all STCs by April 13, 2004, and STCs were instructed to train all TAs and monitors by April 16, 2004.

3.6 MATERIALS SHIPPING AND RETURN

Test materials were shipped to district offices to arrive by April 2, 2004—twenty days before the first day of testing. Materials were sent to district offices to distribute to schools by April 13, 2004. Each school's shipment was boxed individually and labeled with the number of boxes shipped for that school.

The district office received a shipment of overage materials that included a 10 percent overage of all test materials, with the exception of customized formats, which were sent only in the quantities ordered. Overage materials were to be used by the DTCs to fulfill any additional materials requests from the STCs.

TAs were instructed to return test materials to their respective STC immediately after test administration. STCs redistributed test materials to the TAs who administered makeup tests. Those TAs were instructed to return the makeup materials at the end of the makeup session. STCs were instructed to return all materials—scorable and nonscorable—to their DTCs within one business day after makeup testing, but no later than April 30, 2004.

The DTCs were given three dates for returning materials to Pearson Education Measurement (PEM). The DTCs returned the first shipment to PEM by April 26, 2004. The first shipment contained the scorable ELA and mathematics test booklets, school header sheets, and class sheets for tests given on April 20, 21, and 22, 2004. The second shipment was due by May 3, 2004 and contained the scorable ELA and mathematics test booklets, school header sheets, and class sheets for makeup tests given on April 23–30, 2004. The last shipment contained all nonscorable secure test materials from the schools and the district.

With the overage materials, DTCs were sent "district coordinator kits," which included step-by-step directions on how to return scorable and nonscorable materials. These directions listed toll-free phone numbers to call to schedule pickup for returned materials.

3.7 TEST SECURITY

Test security was critical prior to, during, and following test administration. The specific procedures that were followed during the test administration and used in the handling of documentation were outlined in the *TAM*. The manual included a reprinted excerpt of S.C. Code Ann. § 59-1-445 (1990). In addition, the following administrative guidelines were included in the *HSAP TAM*:

- The STC should observe test administration activities and monitor adherence to test security. Examinees should be made aware that monitoring might occur.
- All secure test materials must be kept in a secure, locked location when not in use.
- Before testing, access to secure materials is restricted to supervised sessions conducted by the STC. Supervised sessions for coding answer-document demographic information may be held the week before testing. Review of test administration directions in oral and signed administration scripts is restricted to supervised sessions held after school on the day before each test.

- After testing, access to secure materials is restricted to makeup testing sessions and supervised sessions for completing or editing demographic codes on student answer documents.
- TAs are encouraged to walk around the room during testing to check that students are marking their answers in the correct sections of the answer documents. It is permissible for TAs to alert students that their answers are being marked in the wrong sections of the answer documents. However, it is not permissible for TAs to stop and read test items or students' responses in students' test booklets.

Following the test administration and the return of materials, PEM sent missing materials letters to districts identifying the number of unreturned secure materials and the barcode numbers of each missing document. The districts had two weeks to respond to the letter before PEM and AIR attempted to contact the DTCs via telephone. Subsequently, the districts either located and returned the materials or sent explanations as to why materials were not found. A toll-free telephone number was provided to answer DTCs' questions regarding the missing materials; in addition, follow-up procedures were employed until all materials were accounted for.

Secure Materials

It was explained to districts and schools that secure materials included regular-print test booklets and all customized test materials. In addition, reference sheets, scratch paper, and separate pages containing student writing were also considered to be secure materials and had to be returned with the nonscorable materials after administration of the tests. DTCs and STCs were instructed to keep secure materials in locked storage at all times when the materials were not in use. These materials were not to be left unattended at any time. Additional security policies requiring secure storage, limited access to items, and secure disposal of documents were explained in the manuals and at the pretest workshops.

Agreements to maintain test security and confidentiality were provided in both manuals, and extras were included in the district and school shipments. DTCs were instructed to have all persons with access to test materials sign security agreements if such agreements were not already on file at the district office for the current school year. This policy was stressed repeatedly in the manuals and during the pretest workshops.

Chapter 4

SCORING

For the spring 2004 HSAP mathematics and English language arts tests, the criteria used to score items were based on the item type. Multiple-choice items were scored using item keys indicating each correct option; constructed-response and extended-response items were scored on the basis of scoring rubrics. For extended-response items, a set of scoring rules was applied in creating final scores. This chapter describes the types of items used on the HSAP assessment, the scoring rules that were applied, and reader reliabilities.

4.1 TYPES OF ITEMS

The spring 2004 HSAP tests included three types of items: multiple choice, constructed response, and extended response.

Multiple Choice

For multiple-choice items, students selected an option from four alternatives: A, B, C, or D. Each multiple-choice item was scored as 1 for the correct response and 0 for an incorrect response. Missing responses (i.e., items that a student did not answer at all) and multiple responses were scored as incorrect.

Constructed Response

Constructed-response items were scored using a generic rubric of a 0–3 scale. Condition codes of B ("blank") and UR ("unreadable," or illegible) were used for nonscorable responses and appeared as 0 on the data file. For the purpose of calculating the total score, the condition codes were recoded as 0.

For the purpose of monitoring rater quality, 15 percent of the responses to each constructed-response item were double-read without resolution. The score assigned by the primary reader was taken as the final score for each constructed-response item. A detailed scoring rubric providing descriptions of the various score points was used in the scoring process.

Extended Response

An extended-response writing item was administered at the beginning of session 1 of the ELA test and was scored under four domains: content/development, organization, voice, and conventions. Score ranges for these domains are 1–4 for content/development, 1–4 for organization, 1–3 for voice, and 1–4 for conventions, for a total possible score of 15 points. Each extended-response item was independently read by two raters, for a total possible composite of 30 points. In addition to the double scoring, about 8 percent of the papers were back-read by chief readers.

For the nonscorable responses, condition codes of B ("blank"), OT ("off topic"), IS ("insufficient" response), and UR ("unreadable," or illegible response) were assigned. For scoring purposes, the condition codes were recoded as 0. The algorithm for scoring extended-

writing responses is presented in table 4.1 for scorable responses (e.g., 1–4 or 1–3 for domain scores). When a paper received a condition code, the paper was pulled and scored by supervisors. The scoring rules for these papers are presented in table 4.2. As with the constructed-response items, the extended-response items were also scored with a detailed rubric that was generic across all extended-response items.

TABLE 4.1

Extended-Response Writing Scoring Algorithm for Papers with Scorable Responses

Rule	First Score (R1)	Second Score (R2)	Action	Back Reading (BR)	Resolution Score (RS) [Third Score]	Final Score (F)
1	R1 = 1–4	R2 = R1	None	NA		F = R1 + R2
2	R1 = 1–4	R2 = 1–4 *and* is adjacent to R1	None	NA		F = R1 + R2
3	R1 = 1–4	R2 = 1–4 *and* is nonadjacent to R1	Resolution required	NA	RS = R1	F = RS + R1
4	R1 = 1–4	R2=1–4 *and* is nonadjacent to R1	Resolution required	NA	RS = R2	F = RS + R2
5	R1 = 1–4	R2 = 1–4 *and* is nonadjacent to R1	Resolution required	NA	RS is adjacent to R1 and R2	F = RS + RS
6	R1 = 1–4	R2 = 1–4 *and* is nonadjacent to R1	Resolution required	NA	RS is adjacent to R1 or R2 but not both	F = RS + R1 if R1 is closer to RS than R2 F = RS + R2 if R2 is closer to RS than R1
7	R1 = 1–4	R2 = R1		BR = R1 = R2		F = BR + R1
8	R1 = 1–4	R2 = R1		BR is adjacent to R1 and R2		F = BR + R1
9	R1 = 1–4	R2 = R1		BR is nonadjacent to R1 and R2		F = BR + BR
10	R1 = 1–4	R2 = 1–4 and R2 is adjacent to R1		BR = R1 and adjacent to R2		F = BR + R1
11	R1 = 1–4	R2 = 1–4 and R2 is adjacent to R1		BR = R2 and adjacent to R1		F = BR + R2
12	R1 = 1–4	R2 = 1–4 and R2 is adjacent to R1		BR is adjacent to R1 and discrepant to R2		F = BR + R1
13	R1 = 1–4	R2 = 1–4 and R2 is adjacent to R1		BR is adjacent to R2 and discrepant to R1		F = BR + R2
14	R1 = 1–4	R2 = 1–4 and R2 is adjacent to R1		BR is nonadjacent to R1 and R2		F = BR + BR

TABLE 4.2

Extended-Response Writing Scoring Algorithm for Papers with Condition Codes

Rule	Supervisor First Score (S1)	Supervisor Second Score (S2)	Action	BR	Supervisor Resolution Score (S3)	Final Score (F)
1	S1 = condition code	S2 = S1	None	NA		F = S1
2	S1 = 1–4	S2 = condition code	Resolution required	NA	S3 = 1–4	F = S3 + S1
3	S1 = condition code	S2 = 1–4	Resolution required	NA	S3 = 1–4	F = S3 + S2
4	S1 = 1–4	S2 = condition code	Resolution required	NA	S3 = condition code	F = S3
5	S1 = condition code	S2 = condition code but not equal to S1	Resolution required	NA	S3 = condition code	F = S3
6	S1 = condition code	S2 = condition code but not equal to S1	Resolution required	NA	S3 = 1–4	F = S3 + S3

4.2 TEST SPECIFICATIONS

The 2004 test specifications for mathematics and English language arts are shown in table 4.3. As noted previously, the 2004 HSAP assessments included multiple-choice, constructed-response, and extended-response items.

TABLE 4.3

Spring 2004 HSAP Distribution of Score Point Values by Reporting Category

Mathematics	Algebra	Data Analysis and Probability	Measurement and Geometry	Number and Operations	Integrated Responses
Percentage	27%	11%	27%	23%	13%
Multiple-choice points	19	8	19	16	—
Constructed-response points	—	—	—	—	9

English Language Arts	Reading Process and Comprehension	Analysis of Texts	Word Study and Analysis	Research	Writing
Percentage	23%	19%	9%	9%	41%
Multiple-choice points	18	15	8	8	8
Constructed-response points	3	3	—	—	—
Extended-response points	—	—	—	—	30

4.3 SCORING PROCESS

AIR's subcontractor, Pearson Educational Measurement (PEM), scored all items. Multiple-choice items were scored by PEM's electronic scanning system. Constructed-response (CR) and extended-response (ER) items were scored by trained scorers using the ePEN system (Electronic Performance Evaluation Network) at two scoring sites: ELA was scored in Mesa, Arizona, and mathematics was scored in Lansing, Michigan.

Prior to actual scoring of the constructed-response and extended-response items, range-finding meetings were held in Columbia, South Carolina, from March 29 through April 2, 2004. The purposes of the range-finding meetings were twofold: to identify sets of papers that were representative of the various performance levels defined by the rubric and to arrive at consensus scores on large sets of papers for use in training raters. Three range-finding committees—one each for reading, writing, and mathematics—were convened. The committees were composed of educators from South Carolina and were selected by the SDE. Each committee reviewed several items. That is, each committee reviewed multiple papers (students' responses written to a specific item) for multiple items.

AIR and SDE staff were on-site during the first week of rater training (scorers received on-line training via the ePEN system) and live scoring and monitored the scoring process until scoring was complete. Throughout the scoring process, PEM staff posted the performance of each reader (reader reliability statistics) once a day on the PEM's SchoolHouse Web site for AIR and SDE staff to review.

Throughout scoring, readers' performances were monitored through the use of validity papers, which are prescored responses distributed to readers throughout scoring to ensure that the readers, as well as scoring supervisors, do not drift from the scoring rubric. "True scores" for these papers were assigned by scoring leaders and then stored in the ePEN system. Reader agreement was checked on a regular basis—every twenty papers for the extended-response item and every sixty papers for CR items. This quality check was "blind" in that readers did not know they were scoring a validity paper.

4.4 READER RELIABILITY

In the scoring of constructed-response and extended-response items, 15 percent of the papers for CR items and 100 percent of the papers for ER items were independently scored by two readers. The reader consistency of the papers that were double-scored is reported in table 4.4.

The reported reader-reliability indexes are Spearman's rank-order correlations, rates of perfect agreement, rates of adjacent agreement, and proportions of discrepant scores. The term "perfect agreement" indicates that the two readers assigned the same score for the same written response. The term "adjacent agreement" indicates that the two readers differ by 1 point when evaluating the same response. "Discrepant" scores are those where the readers assigned scores that were 2 or more points apart on the same response. For the ER item, discrepant scores were resolved by a third reader. The resolved scores were used in computing the final score for the score reports.

TABLE 4.4

Reader Reliabilities for Scoring Constructed-Response and Extended-Response Items

Item	N	Percentage of Perfect Agreement	Percentage of Perfect and Adjacent Agreement
CR 1	7,938	87.3%	97.8%
CR 2	7,943	87.4%	99.5%
CR 3	7,938	86.0%	98.7%
CR 1	7,824	73.3%	99.4%
CR 2	7,892	69.9%	99.1%
ER content and development	52,341	75.3%	98.8%
ER organization	52,341	72.1%	98.3%
ER voice	52,341	75.9%	99.4%
ER convention	52,341	74.5%	97.5%

Note: The total number of ER papers (52,341) excludes 881 papers that were illegible.

4.5 TESTED/NOT TESTED FLAG

A student was considered "tested" in mathematics if he or she answered at least one question. The question could have been a multiple-choice or constructed-response item. A student was considered "tested" in ELA if he or she answered at least one question on either of the two days of testing. The one question could have been a multiple-choice item, constructed-response item, or extended-response writing item.

Chapter 5

TECHNICAL CHARACTERISTICS OF ITEMS

This section reports the results of item analyses based on classical test theory (CTT) using a proprietary program designed by AIR. Item difficulty (*p*) is the proportion (or percentage) of examinees correctly answering a dichotomously scored item.

The term "item discrimination" is defined as a correlation between the item score and the total score. For the discrimination index, point-biserial correlations were produced. In computing the point-biserial correlation, the item was excluded in the total score.

For the item discrimination index, AIR produced biserial correlations (i.e., product-moment correlations between a normally distributed latent variable underlying the right-wrong dichotomy and the total score) rather than point-biserial correlations (product-moment correlation between the dichotomous item score and the total score) (Millman and Greene 1989).

A "not-reached" (NR) item was defined as any item to which a student did not respond after the last item that he or she attempted in a session. The percentage of students who did not respond to an item and all the items thereafter was computed as NR. An "omit" was defined as any nonresponse item appearing between items with responses.

In recoding missing data for item analysis, all omitted and not-reached items were recoded as incorrect, with a zero score. After holding discussions, SDE and AIR staff decided to exclude from the CTT item analyses and item calibrations those students who had used customized materials and those students who had received the alternative scoring rubric modification.

5.1 ITEM NONRESPONSE RATES

Although the HSAP tests were not timed, students were required to finish each test session during one school day, unless they had an IEP that allowed for accommodations in administration. TAs were instructed that the expected finishing times for each session would be about two hours for ELA and approximately three hours for mathematics.

Table 5.1 presents the percentage of students who responded to the last two items in a given form, averaging across forms for ELA. The percentages listed in the "Last Item" column of table 5.1 represent the percentage of students who responded to the last item (CR item 3 for mathematics; a multiple-choice (MC) item in both session 1 and session 2 for ELA); the percentages in the adjacent column include students who omitted the last item but answered the next-to-last item (CR item 2 for mathematics; item 19 in session 1 and item 72 in session 2 for ELA). In ELA, item nonresponse rates were computed for each session separately. As can be expected, students tend to leave CR items blank more often than they do MC items, especially when the CR items appear at the end of the test.

TABLE 5.1

Percentage of Students Responding to Last and Second-to-Last Items

Subject	Responding to Last Item	Responding to Second-to-Last Item
Mathematics	92.5% (CR)	94.4% (CR)
ELA session 1	99.6% (MC)	99.5% (MC)
ELA session 2	99.3% (MC)	99.3% (MC)

5.2 CLASSICAL ITEM STATISTICS

Table 5.2 provides a summary of item *p*-values and item discriminations by item types and content areas for the mathematics operational items. Table 5.3 provides a summary of item *p*-values and item discriminations by item types and content areas for the English language arts operational and embedded field-test items. For CR and ER items, the *p*-value was computed as the ratio of the item mean to the item's maximum possible score (MPS). For the discrimination index, point-biserial correlations were computed between the item and the total raw score as the criterion. In computing the point-biserial correlation, the item was excluded in the total raw score.

TABLE 5.2

Summary of Classical Item Statistics for Mathematics

Item Type/Content Area	Number of Items	*p*-value	Point-Biserial Correlation
Multiple-choice items	62	0.69	0.39
Constructed-response items	3	0.61	0.69
Number and Operations	16	0.76	0.43
Algebra	19	0.70	0.38
Measurement and Geometry	19	0.62	0.37
Data Analysis and Probability	8	0.69	0.37

TABLE 5.3

Summary of Classical Item Statistics for English Language Arts

Item Type/Content Area	Number of Items	*p*-value	Point-Biserial Correlation
Multiple-choice items	57	0.69	0.35
Constructed-response items	2	0.63	0.49
Extended-response item	1	0.91	0.63
Reading Process and Comprehension	19	0.70	0.33
Analysis of Texts	16	0.69	0.40
Word Study and Analysis	8	0.76	0.41
Research	8	0.65	0.32
Writing	9	0.77	0.47
Field-test items	80	0.68	0.36

Chapter 6

ITEM CALIBRATION AND SCALING

6.1 METHODOLOGY AND SOFTWARE

The Rasch model was used in the item calibrations of the HSAP items. The one-parameter Rasch model (Rasch 1980; Wright and Stone 1979) was used to calibrate multiple-choice items. Constructed-response and extended-response items were calibrated with the Rasch partial credit model (Masters 1982). Calibrating mixed item types from different assessment modes (i.e., dichotomously and polytomously scored items) requires the use of a polytomous model, which allows the number of score categories (typically score points on a scoring rubric) to vary across assessment modes. The Rasch partial credit model (Wright and Masters 1982) can accommodate the mixing of dichotomous and polytomous items.

The Rasch partial credit model is widely used for high school graduation exams, particularly those with high stakes for students and educators. AIR used a one-to-one translation from the number of correct responses to the scale score in the Rasch model. Maintaining a correspondence between the raw number correct score and the scale score, while simultaneously equating multiple test forms, posed a challenge that was best met by using the one-parameter Rasch dichotomous model and the Rasch partial credit model (Wright and Masters 1982).

The WINSTEPS software program (Linacre and Wright 2003) was used in the item calibration. WINSTEPS employs a joint maximum likelihood approach to estimation (JMLE), which estimates the item and person parameters simultaneously. This estimation method is subject to small statistical biases, which increase as the length of the scale decreases. This estimation bias was corrected through the use of the WINSTEPS feature STBIAS=Y.

6.2 PRE-EQUATING

AIR staff conducted a field test with a sufficient number of items in spring 2003 to create a precalibrated item pool and to construct the equated operational test forms for each subject area.

Constructing a precalibrated item pool from a large field test accomplishes the following:

- helps balance the content and difficulty levels across forms up front,
- prevents item exposure by not having to use embedded field-test items in the operational administrations,
- expedites the operational scoring process by having the scoring conversion tables ready prior to the test administrations,
- allows evaluation of the decision-consistency levels across forms,
- provides more time for the quality control of score reports,
- reduces the cost of administering multiple forms in each operational administration with embedded field-test items,
- reduces the burden on TAs and students when contrasted with independent field testing, and

- simplifies the operational processes by administering one form in each administration.

Rasch-ability-to-scale-score conversion tables were produced before each test administration based on the item parameters in the precalibrated item pool.

6.3 ITEM CALIBRATION

For mathematics, the equated operational test forms were constructed from the precalibrated item pool based on the spring 2003 census field-test items; therefore, the raw-score-to-scale-score conversion table for the spring 2004 operational form was created before the test was administered.

For English language arts, although the spring 2003 field-test forms covered all standards specified in the ELA test specification, a few standards needed to be augmented with additional items. In order to replenish the ELA precalibrated item pool for these standards, the SDE and AIR decided to embed field-test items in spring 2004 and 2005 HSAP operational administrations. In the spring 2004 HSAP administration, 67 items (63 MC, 2 CR, and 1 ER) were common on all ELA forms. The 63 multiple-choice items included 56 operational items and 7 embedded field-test items for scoring. In each form, 10 unique field-test items (5 items at the end of each session) were added, resulting in a total of 80 unique field-test items for future use.

In English language arts, the field-test items (including all embedded and added field-test items) were placed on the item bank scale. The operational item parameters were anchored at the bank difficulty values; therefore, the field-test item difficulties were mapped onto the bank metric in the concurrent calibration. The pre-equated item parameters were used in scoring the spring 2004 administration; however, the post-equated item parameters and the student performance based on pre-equated and post-equated item parameters were subsequently reviewed with the SDE and the Technical Advisory Committee on June 30, 2004, and it was confirmed that the pre-equated item parameters should be used.

6.4 COMPOSITION OF THE CALIBRATION SAMPLE

A subset of the field-test items was expected to be used as operational items in order to fulfill test blueprint requirements. Early return samples were identified so that parameter estimation for the field-test items could begin as soon as possible after test administration and not jeopardize the score reporting schedule.

The calibration samples were preselected based on spring 2003 HSAP results (i.e., percentage of students at or above *proficient*). The goal was to have the calibration sample represent the academic performance and demographic characteristics of the total student population as closely as possible. School districts were selected as the sampling unit due to convenience because the test materials were returned by each district and not by individual schools. Two calibration samples were selected—the first sample was selected for the item calibration, and the second was selected to supplement the first sample. The rationale for drawing a second sample was the possibility that some of the districts in the first sample might not return their materials in a timely fashion.

In actuality, since there was enough time to score all of the student papers in both samples prior to the calibration analyses, all of the students in the two samples were included in the item calibration. In total, approximately 25,000 students from 110 schools and 41 districts (about half of the total grade 10 student population) were included in the calibration analyses. Students who took customized materials and students who received an alternative scoring modification were excluded from the analyses.

6.5 SCALING

Based on the precalibrated item pool, Rasch-ability-score-to-scale-score conversion tables were generated for each subject. These scores took into account any differences in the difficulty of the forms due to pre-equating; that is, all items shared a common metric so that the scale scores developed for each form were automatically adjusted for differences in item difficulty.

For the transformation of Rasch-ability-score-to-scale-score, the following steps were taken in generating scale scores:

Step 1: Linear transformation of Rasch-ability-score-to-scale-score, fixing the passing scale score (Level 2) at 200 with a standard deviation of 25,

$$SS = SS_C + B\left[\frac{\hat{\theta} - \theta_C}{\sigma_{\hat{\theta}}}\right] = 200 + 25\left[\frac{\hat{\theta} - \theta_C}{\sigma_{\hat{\theta}}}\right],$$

where the passing ability scores (θ_C) are -0.224 for mathematics and 0.015 for ELA and the standard deviations of *theta* ($\sigma_{\hat{\theta}}$) are 1.102 for mathematics and 1.046 for ELA.

Step 2: The decimals in the scale score were truncated to avoid the same scale score for two different raw scores.

Step 3: Scale scores less than 100 and greater than 320 were fixed with 100 and 320, respectively.

6.6 DEFINITION OF SCOREABILITY

A student was considered "tested" if the student answered at least one question in the test booklet. All tested students' item responses were scored. All omits and not-reached items were recoded as incorrect and given a zero score.

6.7 REPORTING OF ZERO AND PERFECT SCORE

In item response theory (IRT) maximum-likelihood ability estimation methods, zero and perfect scores are assigned the ability of minus and plus infinity. AIR used the WINSTEPS default setting in estimating the extreme values. That is, a fractional score point value was subtracted from perfect scores, and was added to zero scores.

6.8 POLICY DEFINITION OF ACHIEVEMENT LEVELS

After the spring 2003 HSAP census field test, AIR, in collaboration with its partner Insite conducted the standard-setting workshops for the HSAP mathematics and ELA examinations on July 21–25, 2003. In each subject, the workshop participants recommended three achievement-level cut scores: Level 2, Level 3, and Level 4. Level 2 was the cut required for student graduation purposes, and Levels 3 and 4 described students for AYP (adequate yearly progress) purposes. Achievement-level descriptions are provided below in tables 6.1 and 6.2. AIR outlined the details of the standard-setting process in its 2004 report to the SDE, "South Carolina High School Assessment Program English Language Arts and Mathematics Standard Setting Technical Report."

TABLE 6.1

Description of Achievement Levels for the HSAP Mathematics Test

Level	Description
4	The Level 4 student • has demonstrated an exceptional command of skills and knowledge required of high school students in South Carolina • analyzes, evaluates, and/or synthesizes mathematical concepts and procedures and solves problems using advanced arithmetic, algebraic, and measurement/geometric concepts and relationships • analyzes data representations and applies probability concepts • supports answers with mathematical work and/or explanations that thoroughly communicate mathematical reasoning • has met the exit examination requirement for a South Carolina high school diploma
3	The Level 3 student • has demonstrated proficiency in skills and knowledge required of high school students in South Carolina • applies mathematical concepts and procedures and solves problems using arithmetic, algebraic, and measurement/geometric concepts and relationships • interprets data representations and demonstrates a knowledge of probability concepts • supports answers with mathematical work and/or explanations that clearly communicate mathematical reasoning • has met the exit examination requirement for a South Carolina high school diploma
2	The Level 2 student • has demonstrated competence in skills and knowledge required of high school students in South Carolina • demonstrates an acceptable knowledge of fundamental mathematical concepts and procedures and solves problems using essential arithmetic, algebraic, and measurement/geometric concepts and relationships • demonstrates a knowledge of basic data representations and probability concepts • supports answers with mathematical work and/or explanations that adequately communicate mathematical reasoning • has met the exit examination requirement for a South Carolina high school diploma

TABLE 6.1

Description of Achievement Levels for the HSAP Mathematics Test

Level	Description
1	The Level 1 student • has not demonstrated competence in the skills and knowledge required of high school students in South Carolina • demonstrates a limited understanding of mathematical concepts • is able to use arithmetic, algebraic, and measurement/geometric concepts and relationships • demonstrates a knowledge of simple data representations and probability concepts • supports answers with mathematical work and/or explanations that minimally communicate mathematical reasoning • has not met the exit examination requirement for a South Carolina high school diploma

TABLE 6.2

Description of Achievement Levels for the HSAP English Language Arts Test

Level	Description
4	The Level 4 student • has demonstrated an exceptional command of skills and knowledge required of high school students in South Carolina • demonstrates comprehension of complex ideas and connects those ideas within a text, across texts, and beyond the text • displays exceptional writing skills by engaging the reader, effectively developing and organizing ideas, and using relevant supporting details, vivid language, and Standard American English • has met the exit examination requirement for a South Carolina high school diploma
3	The Level 3 student • has demonstrated proficiency in skills and knowledge required of high school students in South Carolina • demonstrates comprehension of complex ideas and connects those ideas within a text and across texts • displays effective writing skills by sustaining the reader's interest, clearly developing and organizing ideas, and using relevant supporting details and Standard American English • has met the exit examination requirement for a South Carolina high school diploma

TABLE 6.2

Description of Achievement Levels for the HSAP English Language Arts Test

Level	Description
2	The Level 2 student • has demonstrated competence in skills and knowledge required of high school students in South Carolina • demonstrates comprehension of essential ideas and shows some logical connections of those ideas within a text • displays acceptable writing skills by showing some awareness of audience, developing and organizing ideas, and using relevant supporting details and Standard American English • has met the exit examination requirement for a South Carolina high school diploma
1	The Level 1 student • has not demonstrated competence in skills and knowledge required of high school students in South Carolina • demonstrates limited comprehension of ideas and tenuous connections of those ideas within a text • displays limited writing skills, which may include little awareness of audience and purpose, partial development and organization of ideas, and deviations from Standard American English • has not met the exit examination requirement for a South Carolina high school diploma

6.9 CUT SCORE FOR ACHIEVEMENT LEVELS

The cut scores recommended by the panelists at the 2003 standard-setting workshops were considered preliminary because the SDE intended to reexamine the achievement levels set using the 2003 field-test data by adding the data from the spring 2004 operational HSAP administration. AIR performed a set of confirmatory standard-setting analyses using the 2003 and 2004 data. The results of these analyses were presented to the Technical Advisory Committee (TAC) and the SDE. After reviewing the results from the spring 2004 data, the TAC made recommendations about the locations of Level 3 and Level 4 cut scores. (See chapter 8, below, for the confirmatory process.) The cut scores for total scores for the spring 2004 operational HSAP test forms are presented in table 6.3.

TABLE 6.3

Cut Scores in Rasch Ability Scale and Scale Score for Total Score

	Level 2	Level 3	Level 4
Mathematics			
Rasch Ability	-0.190	0.664	1.613
Scale Score	200	220	241
English Language Arts			
Rasch Ability	0.064	0.994	1.758
Scale Score	200	223	241

6.10 CONTENT-AREA INFORMATION

In addition to total scores, information was reported for four content areas in mathematics and five content areas in ELA. For each content area, the following steps were taken:

Step 1: A raw-score-to-Rasch-ability-score conversion table was generated for each content area. The Level 2 passing ability score of a total score was located on the scale.

Step 2: A 68 percent confidence interval of the passing ability score (θ_c) was computed as: passing ability score (θ_c) $\pm$ 1 SE(θ_c). The ability scores were categorized into three classifications as following:

Adequate: if $\theta \geq \theta_c + 1SE$
May need improvement: if $\theta_c - 1SE \leq \theta < \theta_c + 1SE$
Needs improvement: if $\theta < \theta_c - 1SE$

The Rasch-ability-score-to-content-area cut scores used for the classifications for each content area are provided in table 6.4.

TABLE 6.4

Cut Scores on the Rasch Ability Scale, Associated Standard Errors, and Confidence Intervals for Content-Area Classifications

Content Area	Rasch Ability (θ)	SE(θ)	68% Confidence Interval	
			θ - 1SE	θ + 1SE
Mathematics				
Number and Operations	0.031	0.567	-0.536	0.598
Algebra	-0.004	0.477	-0.481	0.473
Measurement and Geometry	-0.072	0.478	-0.550	0.406
Data Analysis and Probability	-0.177	0.736	-0.913	0.559
English Language Arts				
Reading Process and Comprehension	0.190	0.479	-0.289	0.669
Analysis of Texts	0.026	0.503	-0.477	0.529
Word Study and Analysis	0.098	0.763	-0.665	0.861
Research	0.289	0.742	-0.453	1.031
Writing	0.083	0.399	-0.316	0.482

6.11 PERCENTAGE OF STUDENTS IN EACH ACHIEVEMENT LEVEL

Tables 6.5 and 6.6 present student performance on the 2004 HSAP operational test for mathematics and English language arts. Percentages of students in the four achievement levels are reported for overall and subgroups. Subgroups include the reporting categories of gender, ethnicity, language fluency (i.e., LEP—limited English proficiency), lunch program participation, migrant status, and disability. The summary includes all students who were tested but excludes students in adult education and district-approved home schools. Tables 6.7 and 6.8 provide the information for content areas. The information is summarized for Level 1 and at or above Level 2 for all students by gender and by ethnic group. Of those students who took both the mathematics and English language arts tests, 73.5 percent passed both tests.

TABLE 6.5

Spring 2004 HSAP Mathematics Operational Test: Percentage of Students in Achievement Levels Overall and by Subgroups

Subgroup	Achievement Levels				L2+*	L3+**	N
	Level 1	Level 2	Level 3	Level 4			
Overall	22.8	29.0	27.9	20.3	77.2	48.2	52,913
Gender							
Female	19.9	30.6	29.9	19.6	80.1	49.5	25,956
Male	25.0	27.3	26.4	21.3	75.0	47.7	26,242
Invalid	46.9	32.2	14.8	6.2	53.1	21.0	715
Ethnicity							
African American	36.6	35.9	21.0	6.4	63.4	27.5	21,450
Asian/Pacific Islander	6.5	15.4	29.1	49.0	93.5	78.1	526
Hispanic	27.3	35.1	25.7	11.9	72.7	37.6	1,144
American Indian	25.3	27.4	28.4	18.9	74.7	47.4	95
White	11.9	23.9	33.4	30.8	88.1	64.2	28,663
Other	27.0	24.8	28.9	19.3	73.0	48.2	367
Unknown	46.3	30.7	16.5	6.6	53.7	23.1	668
Language							
English speaker	22.7	28.9	28.0	20.4	77.3	48.4	52,132
Full LEP	38.1	34.9	20.2	6.9	61.9	27.1	436
LEP mainstream	16.7	38.9	28.7	15.7	83.3	44.4	108
Waiver	22.6	38.7	35.5	3.2	77.4	38.7	31
Exited	14.1	32.7	30.7	22.4	85.9	53.2	205
Unknown	0.0	100.0	0.0	0.0	100.0	0.0	1
Lunch Program							
No free/reduced lunch	14.7	24.9	31.6	28.8	85.3	60.4	31,515
Free lunch	36.7	35.2	21.3	6.8	63.3	28.1	18,042
Reduced lunch	23.6	34.1	29.4	12.9	76.4	42.3	3,355
Unknown	100.0	0.0	0.0	0.0	0.0	0.0	1
IEP							
Yes	66.4	22.8	8.7	2.1	33.6	10.8	6,641
No	16.5	29.9	30.7	22.9	83.5	53.6	46,218
Unknown	53.7	33.3	9.3	3.7	46.3	13.0	54
Migrant							
Yes	34.2	50.0	13.2	2.6	65.8	15.8	38
No	22.8	29.0	28.0	20.3	77.2	48.3	52,874
Unknown	0.0	100.0	0.0	0.0	100.0	0.0	1

***** indicates the percentage of students at or above Level 2
** indicates the percentage of students at or above Level 3

TABLE 6.6

Spring 2004 HSAP English Language Arts Operational Test: Percentage of Students in Achievement Levels Overall and by Subgroups

Subgroup	Achievement Levels				L2+*	L3+**	N
	Level 1	Level 2	Level 3	Level 4			
Overall	17.4	28.0	30.3	24.4	82.6	54.6	53,222
Gender							
Female	12.6	27.6	32.0	27.9	87.4	59.9	26,073
Male	21.7	28.2	28.8	21.3	78.3	50.2	26,408
Invalid	39.0	33.1	20.4	7.6	61.0	27.9	741
Ethnicity							
African American	27.7	37.2	25.4	9.7	72.3	35.1	21,603
Asian/Pacific Islander	11.6	19.8	28.1	40.5	88.4	68.6	526
Hispanic	32.1	30.3	23.3	14.3	67.9	37.7	1,150
American Indian	20.0	30.5	26.3	23.2	80.0	49.5	95
White	8.7	20.9	34.4	35.9	91.3	70.3	28,781
Other	18.8	26.1	30.9	24.2	81.2	55.1	372
Unknown	38.8	33.2	20.4	7.5	61.2	27.9	695
Language							
English Speaker	17.0	27.9	30.4	24.6	83.0	55.0	52,440
Full LEP	61.4	29.2	8.6	0.7	38.6	9.3	428
LEP mainstream	27.2	36.0	29.8	7.0	72.8	36.8	114
Waiver	25.8	32.3	25.8	16.1	74.2	41.9	31
Exited	18.3	30.3	35.6	15.9	81.7	51.4	208
Unknown	100.0	0.0	0.0	0.0	0.0	0.0	1
Lunch Program							
No Free/reduced Lunch	10.5	22.2	33.2	34.1	89.5	67.3	31,638
Free Lunch	29.4	37.1	24.7	8.9	70.6	33.6	18,205
Reduced Lunch	17.9	32.9	32.8	16.3	82.1	49.1	3,378
Unknown	0.0	100.0	0.0	0.0	100.0	0.0	1
IEP							
Yes	59.5	27.9	10.0	2.6	40.5	12.6	6,749
No	11.3	27.9	33.2	27.5	88.7	60.8	46,416
Unknown	47.4	36.8	15.8	0.0	52.6	15.8	57
Migrant							
Yes	52.6	26.3	13.2	7.9	47.4	21.1	38
No	17.4	28.0	30.3	24.4	82.6	54.6	53,183
Unknown	0.0	100.0	0.0	0.0	100.0	0.0	1

***** indicates the percentage of students at or above Level 2
** indicates the percentage of students at or above Level 3

TABLE 6.7

Spring 2004 HSAP Mathematics Operational Test: Content-Area Information

Subgroup	Level 1				Level 2 and Above			
	Needs Improvement	*May need improvement*	*Adequate*	N1*	*Needs Improvement*	*May need improvement*	*Adequate*	N2**
NUMBER AND OPERATIONS								
All students	81.2 %	17.7%	1.1%	12,045	5.7%	26.8%	67.5%	40,868
Females	80.8%	18.4%	0.8%	5,159	6.4%	28.9%	64.7%	20,797
Males	81.2%	17.5%	1.4%	6,551	5.0%	24.4%	70.6%	19,691
African Americans	81.8%	17.2%	1.0%	7,860	8.8%	38.1%	53.2%	13,590
Whites	79.2%	19.4%	1.4%	3,407	4.0%	20.6%	75.4%	25,256
ALGEBRA								
All students	59.4%	38.1%	2.5%	12,045	2.1%	23.6%	74.3%	40,868
Females	54.3%	42.7%	3.0%	5,159	1.9%	23.2%	74.9%	20,797
Males	63.2%	34.6%	2.1%	6,551	2.3%	23.7%	74.0%	19,691
African Americans	58.8%	38.5%	2.7%	7,860	2.8%	31.3%	65.9%	13,590
Whites	61.0%	37.1%	1.9%	3,407	1.7%	19.4%	78.9%	25,256
MEASUREMENT AND GEOMETRY								
All students	45.0%	53.4%	1.6	12,045	1.3%	34.2%	64.4%	40,868
Females	43.1%	55.1%	1.8	5,159	1.4%	36.2%	62.4%	20,797
Males	46.2%	52.3%	1.5	6,551	1.3%	31.8%	67.0%	19,691
African Americans	45.6%	53.0%	1.4	7,860	2.4%	50.7%	46.9%	13,590
Whites	43.4%	54.4%	2.1	3,407	0.8%	25.1%	74.1%	25,256
DATA ANALYSIS AND PROBABILITY								
All students	31.3%	63.1%	5.5%	12,045	1.2%	33.4%	65.4%	40,868
Females	30.0%	64.3%	5.7%	5,159	1.4%	33.3%	65.3%	20,797
Males	32.2%	62.4%	5.4%	6,551	1.0%	33.2%	65.8%	19,691
African Americans	32.8%	62.2%	5.1%	7,860	2.2%	46.0%	51.8%	13,590
Whites	27.4%	65.7%	6.9%	3,407	0.6%	26.4%	73.0%	25,256

* total number students in Level 1

** total number students in Levels 2, 3, and 4

TABLE 6.8

Spring 2004 HSAP English Language Arts Operational Test: Content-Area Information

Subgroup	Level 1				Level 2 and Above			
	Needs Improvement	*May need improvement*	*Adequate*	N1*	*Needs Improvement*	*May need improvement*	*Adequate*	N2**
READING PROCESS AND COMPREHENSION								
All students	46.9%	49.5%	3.7%	9,282	0.5%	23.0%	76.5%	43,940
Females	42.4%	54.3%	3.3%	3,273	0.5%	22.9%	76.6%	22,800
Males	49.2%	47.0%	3.8%	5,720	0.5%	22.9%	76.6%	20,688
African Americans	46.3%	50.3%	3.4%	5,978	0.8%	33.1%	66.1%	15,625
Whites	47.6%	48.3%	4.2%	2,515	0.3%	16.9%	82.8%	26,266
ANALYSIS OF TEXTS								
All students	73.8%	21.1%	5.2%	9,282	3.4%	14.1%	82.5%	43,940
Females	69.3%	24.7%	6.0%	3,273	2.6%	12.8%	84.6%	22,800
Males	76.2%	19.1%	4.7%	5,720	4.3%	15.4%	80.3%	20,688
African Americans	75.3%	20.7%	4.1%	5,978	6.0%	22.6%	71.4%	15,625
Whites	70.3%	22.0%	7.7%	2,515	1.8%	8.9%	89.3%	26,266
WORD STUDY AND ANALYSIS								
All students	57.3%	39.7%	3.0%	9,282	4.0%	41.5%	54.5%	43,940
Female	57.7%	40.4%	2.0%	3,273	4.9%	44.9%	50.2%	22,800
Male	56.9%	39.4%	3.7%	5,720	3.0%	37.5%	59.5%	20,688
African American	59.4%	38.5%	2.1%	5,978	7.4%	55.4%	37.3%	15,625
White	52.4%	42.4%	5.2%	2,515	1.9%	33.0%	65.1%	26,266
WRITING								
All students	61.1%	32.2%	6.8%	9,282	1.8%	14.8%	83.4%	43,940
Females	51.7%	39.2%	9.1%	3,273	1.0%	12.2%	86.7%	22,800
Males	66.4%	28.1%	5.5%	5,720	2.6%	17.5%	79.9%	20,688
African Americans	60.0%	33.1%	6.9%	5,978	2.5%	21.2%	76.3%	15,625
Whites	62.0%	30.7%	7.2%	2,515	1.3%	10.8%	87.9%	26,266
RESEARCH								
All students	44.4%	52.4%	3.3	9,282	3.3%	44.7%	52.1%	43,940
Females	41.9%	54.7%	3.5	3,273	3.0%	43.2%	53.8%	22,800
Males	45.7%	51.1%	3.2	5,720	3.5%	46.1%	50.4%	20,688
African Americans	44.2%	52.9%	2.9	5,978	4.8%	53.7%	41.5%	15,625
Whites	45.2%	50.9%	3.9	2,515	2.3%	39.2%	58.5%	26,266

* total number students in Level 1

** total number students in Levels 2, 3, and 4

Chapter 7

DESCRIPTIVE STATISTICS

Descriptive statistics of scale score distributions are presented in table 7.1. The scale score distributions are compared among all students, gender, and ethnic group categories in figures 1 and 2.

TABLE 7.1

Summary Statistics Overall and by Subgroups

Subgroup	N	Scale Score	
		Mean	SD
Mathematics			
All students	52,913	219.3	26.8
Males	26,242	218.9	28.5
Females	25,956	220.1	25.0
African Americans	21,450	207.0	21.3
Whites	28,663	228.7	26.5
English Language Arts			
All students	53,222	222.8	24.4
Males	26,408	219.5	25.3
Females	26,073	226.5	22.9
African Americans	21,603	212.4	22.2
Whites	28,781	231.2	22.5

FIGURE 1

Scale Score Distribution for Mathematics

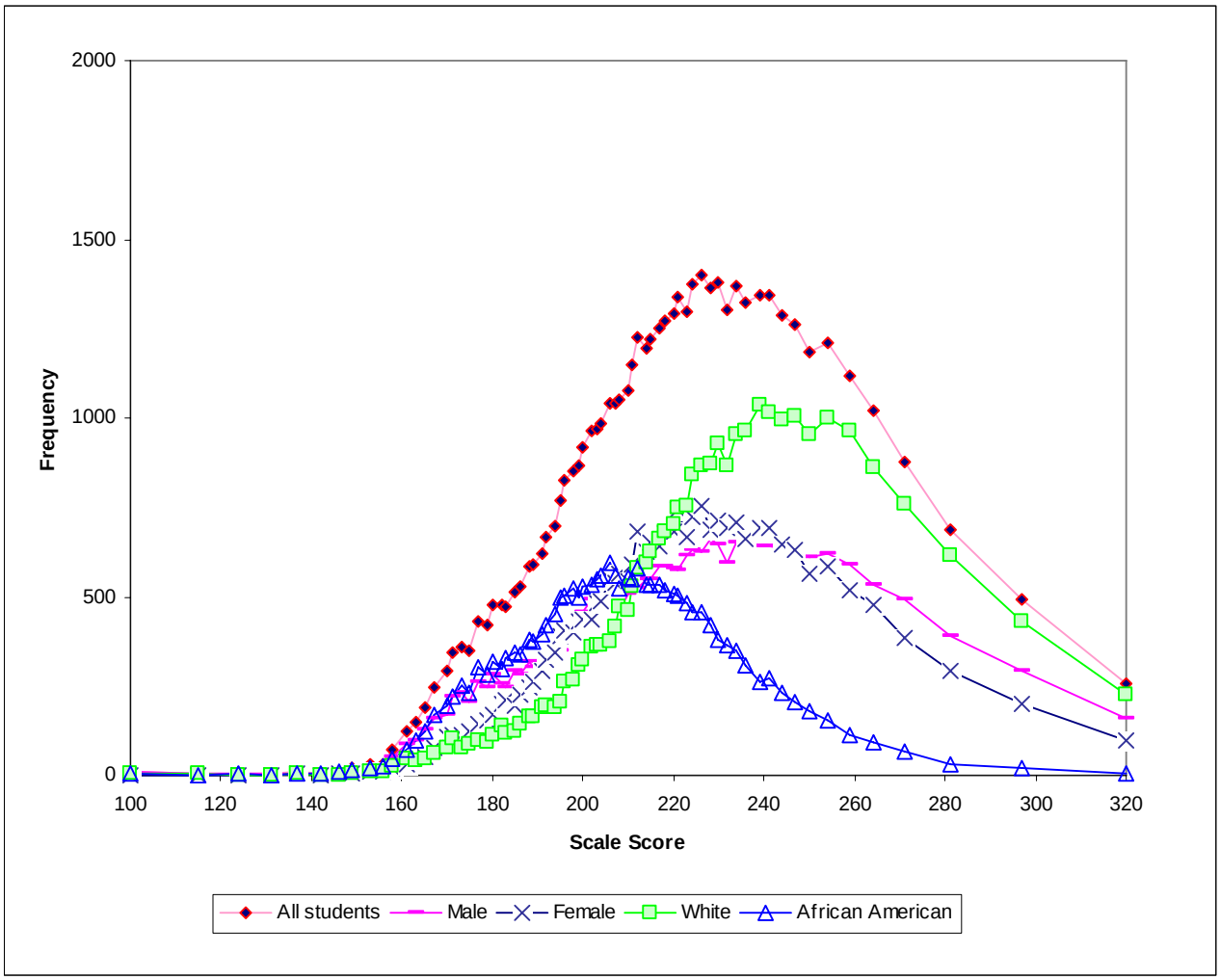
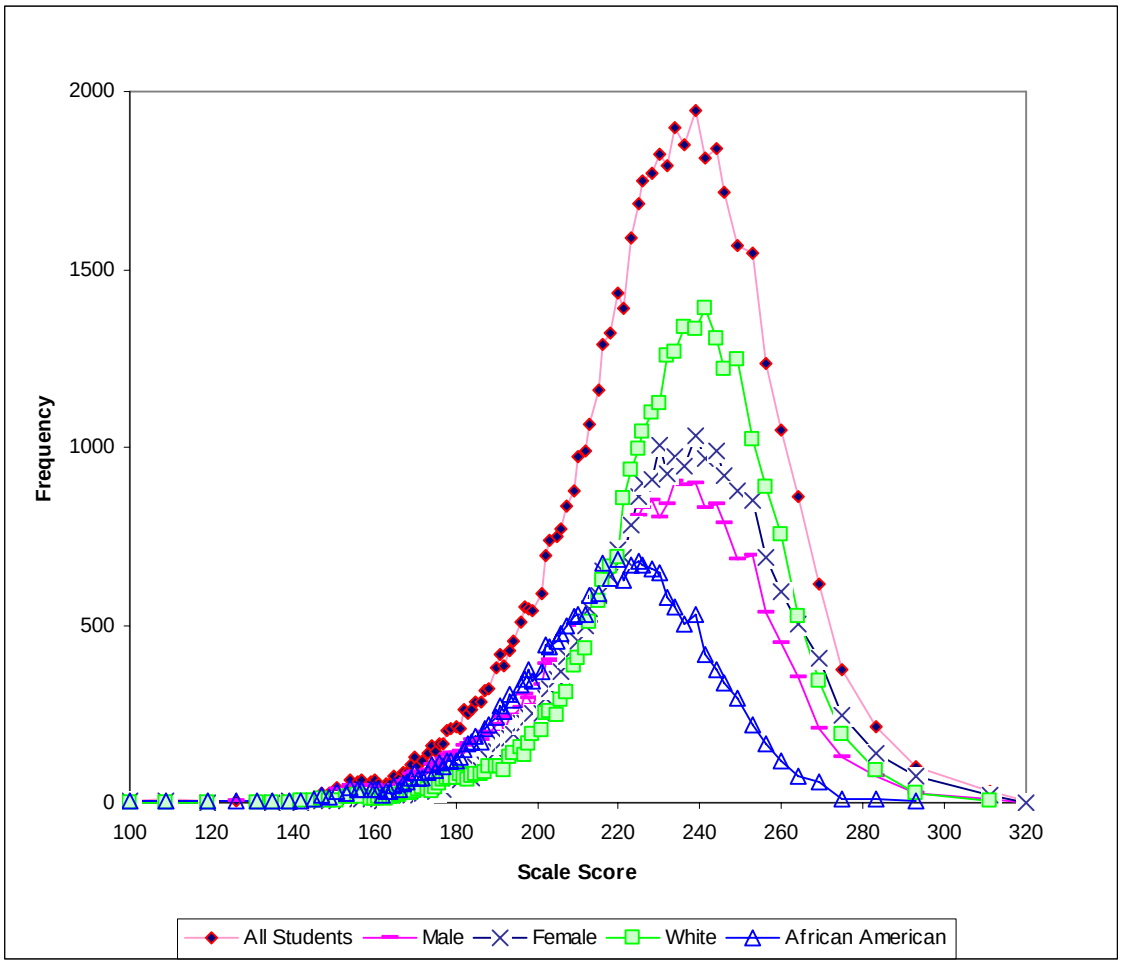


FIGURE 2

Scale Score Distribution for English Language Arts



Chapter 8

CONFIRMATION OF ACHIEVEMENT LEVELS

8.1 OVERVIEW

AIR conducted the achievement level standard-setting sessions July 21–25, 2003, in Columbia, South Carolina, for the HSAP mathematics and English language arts examinations. Although the No Child Left Behind Act regulations required reporting group scores in 2003 to determine AYP (adequate yearly progress), the recommended cut scores at the workshops were considered preliminary, and SDE intended to reexamine the achievement levels set using the 2003 field-test data by adding the data from the spring 2004 operational HSAP administration.

A set of confirmatory analyses, comparing student performance in 2003 and 2004, was requested primarily because of suspected differences between field-test conditions and operational conditions. Student motivation and awareness were likely to be greater for the 2004 operational administration than for the 2003 field-test administration. Similarly, heightened emphasis on the HSAP by schools and districts likely had a positive effect on student scores in 2004.

During its meetings in August and December 2003, the Technical Advisory Committee (TAC) examined the standard-setting procedures and results and discussed possible analyses for the confirmatory process. The TAC suggested approaching the process with two precepts in mind: performance standards should be adjusted if there is evidence of a need for change, and the process used for possible adjustments should be contextualized within a broader confirmatory process.

Further, the TAC noted that the ELA extended-response (ER) items were not included in the ordered item booklet used by the standard-setting panel. Although the TAC acknowledged that a direct comparison of setting the standard with and without the ER items was not possible, the TAC did conclude that some investigation of the impact of the ER items on the total scores was worthwhile. The TAC suggested that AIR conduct analyses evaluating the impact of including and excluding the ER items on student scale scores and, by extension, on student passing status.

At the TAC meeting on June 30, 2004, the original cut scores were reviewed on the basis of the confirmatory analysis results provided by AIR.

8.2 CONFIRMATORY ANALYSES

The confirmatory procedures included the review of patterns of student scores to determine whether they reflected improvement from field test to operational administration. In addition, changes in student scores that might be due to alterations in the positions of constructed-response and extended-response items on the operational test forms were evaluated.

Comparisons of Percentages of Students in Achievement Levels

The percentages of students at or above each achievement level were compared between the spring 2003 census field-test data and the spring 2004 operational data. The anticipated improvements in student performance from the field test to the operational test were confirmed by the 2004 results, as shown in table 8.1.

TABLE 8.1

Percentages of Students in the Achievement Levels

Administration	N	Achievement Levels				Percentage at or Above	
		L1	L2	L3	L4	L2+	L3+
Mathematics							
Spring 2003 FT	45,471	44.2%	19.3%	21.8%	14.7%	55.8%	36.5%
Spring 2004 OT	52,913	22.8%	29.0%	28.0%	20.3%	77.2%	48.2%
English Language Arts							
Spring 2003 FT	46,541	36.2%	21.8%	21.9%	20.1%	63.8%	42.0%
Spring 2004 OT	53,222	17.4%	28.0%	30.3%	24.4%	82.6%	54.6%

For ELA, the percentage of students scoring Level 2 and above increased from 64 percent to 83 percent. Performance on mathematics also improved, with the percentage of students at Level 2 and above increasing from 56 percent to 77 percent. An examination of the *p*-values for the items revealed that the *p*-values for 60 of the 62 mathematics items increased, as did those for 48 of the 52 ELA multiple-choice items.

Score Distributions

The cumulative score distributions between the standard-setting impact data and the operational form were compared in order to examine the changes in student performance throughout the score scale. The score distribution for 2004 reflected higher performance in 2004 than in 2003.

Changes in the Test and Test Administration

Changes that were made in the test and test administration between the field test and operational test were also considered as possible causes of the score increases. Primary among these were test length (the field test was longer than the operational test), test order (the HSAP field test came after the Basic Skills Assessment Program exit examination), and the order of the subtests (the ELA extended-response item was moved from the end of the field test to the beginning of the operational forms). An examination of the item position data, however, did not establish a systematic relationship between item position and *p*-value.

Changes in Student Population

In addition to the changes in the test itself and the administration procedures noted above, there were differences between the 2003 and 2004 populations of examinees. With the implementation of HSAP, the definition of *tenth grader* was substantially changed from what it was under the

Basic Skills Assessment Program (BSAP). Under the BSAP exit examination guidelines, students were considered tenth graders only if they had earned at least one credit for graduation in both English and mathematics. However, federal guidelines, as applied to HSAP testing, require students to be tested in their second year of high school, whether or not they have earned any credits.

This requirement increased the number of students tested in mathematics from 45,471 in 2003 to 52,987 in 2004 and the number of those tested in ELA from 46,541 in 2003 to 53,304 in 2004. The increased number of examinees suggested that districts may not have tested all of the newly defined tenth graders in 2003. The relationship between changes in the numbers tested in the two years and school performance was examined and the results did not show a significant impact as a result of the increase in examinees.

Impact of Extended-Response Item

In an examination of the impact of omitting the extended-response items from ELA standard setting, test results with and without an extended-response item were compared for spring 2003 field-test impact data and spring 2004 operational data. The consistency of classification of students into levels indicates that the performance standards function very similarly with and without an ER item. The percentage of agreement for students classified as Level 2 or above with and without an extended-response item included was 95.2 percent in 2003 and 95.6 percent in 2004.

Collateral Data Review

In an examination of the reasonableness of the cut scores, standards in other assessments were compared with the HSAP standards. The South Carolina tenth graders' performance on the BSAP, eighth graders' performance on the PACT and on the National Assessment of Educational Progress (NAEP), and tenth graders' performance on the HSAP were compared across years. In addition, the relationships between student performance on the PACT and the HSAP and between the BSAP and the HSAP were examined. The students who took the PACT (eighth graders) and the HSAP (tenth graders) were matched and used to construct an expectancy table between HSAP and PACT achievement-level classifications. The students who took both the PACT and the HSAP were matched: the 2001 PACT with the 2003 HSAP field test and the 2002 PACT with the 2004 HSAP. The students who took both the BSAP and the HSAP in 2003 were also matched.

Equating Model for the HSAP

The stability of item difficulty estimates between each item's 2004 operational parameter estimate and its bank value obtained from the 2003 census field test was examined. The 2004 operational item parameter estimates were transformed to the 2003 difficulty scale by applying linking constants. The linking constants were computed after excluding the item parameters that showed a significant difference between administrations. For each subject area, two linking constants were computed; one based on MC items only and one with MC and CR items. The ER item was not included in computing a linking constant and was treated as a new item because its item position had changed from the end to the beginning of the test. Tables 8.2 and 8.3 present the percentages at or above each level, using the bank (pre-equating) and the adjusted (post-equating) values.

TABLE 8.2

Pass Rates for the HSAP Mathematics Test

Linking Constant Based on MC Only		
Level	**Bank**	**Adjusted**
2	77.2%	77.2%
3	57.5%	59.8%
4	25.3%	27.9%
Linking Constant Based on MC+CR		
Level	**Bank**	**Adjusted**
2	77.2%	75.4%
3	57.5%	57.5%
4	25.3%	27.9%

TABLE 8.3

Pass Rates for the HSAP English Language Arts Test

Linking Constant Based on MC Only		
Level	**Bank**	**Adjusted**
2	82.5%	81.4%
3	64.8%	62.3%
4	38.4%	38.4%
Linking Constant Based on MC+CR		
Level	**Bank**	**Adjusted**
2	82.5%	80.1%
3	64.8%	62.3%
4	38.4%	38.4%

On the basis of the observed drift for some of the item parameters between the 2003 and 2004 test administrations as well as the assumption that item parameters obtained from an operational administration are better estimates than those obtained from a field test, AIR recommended that the SDE consider adopting an equating model that paralleled the post-equating model used for the PACT program. To the degree that item parameter drift occurs, the equating statistics may need to be updated.

AIR suggested that after each HSAP administration, a set of operational items be identified to serve as an appropriate link back to the item bank. Any item with significantly different performance results between the field-test administration and operational administration would be treated as a new item and would therefore not be included as part of the equating analysis.

The TAC, however, identified numerous practical drawbacks to using post-equating for the HSAP. Graduation Express (the scoring system used for graduating seniors) requires an extremely tight turnaround time; therefore, any delay in scoring to conduct the required posttesting analyses could seriously hamper the timely notification of students regarding their graduation status. Furthermore, fall and summer administrations include a higher proportion of low-scoring examinees (i.e., students who failed the HSAP exam at least once) than do the spring administrations. The TAC was concerned that the item parameter estimates based on a low-scoring sample (fall and summer) might not be of sufficient quality.

The TAC advised that equating was unlikely to change the test results very much. The TAC acknowledged that AIR's recommendation for post-equating would be desirable if there were more time, but given the time constraints and the different examinees in fall and summer administrations, the TAC's recommendation was against post-equating. However, the TAC did recommend that the item parameter drift be monitored, that items with statistics that change significantly be routinely tracked, and that, if necessary, the bank be recalibrated to itself. AIR agreed that if it detected meaningful drift it would bring that fact to SDE's attention.

Recommended Final Cut Scores

After reviewing the results of confirmatory analyses and after consultation with the TAC, the SDE recommended the following Level 3 and Level 4 cut scores. Table 8.4 provides final cut scores, as presented in the Rasch ability scale.

TABLE 8.4

Final Cut Scores in Rasch Ability Scale

Subject	Level 2	Level 3	Level 4
Mathematics	-0.224	0.658	1.584
English language arts	0.015	0.978	1.731

Chapter 9

RELIABILITY

In this chapter, three types of reliability indexes are presented: reliability of raw scores, overall standard error of measurement, conditional standard error of measurement, and decision consistency at each achievement level.

9.1 RELIABILITY OF RAW SCORES

For the HSAP assessments, the reliability coefficients were computed using stratified Cronbach *alpha*. As mentioned, the HSAP assessments included mixed item types: multiple choice, constructed response, and extended response. Although there are various techniques for estimating the reliability of test scores with multiple item types or parts (Feldt and Brennan 1989; Lee and Frisbie 1999; Qualls 1995), studies indicated (Qualls 1995; Yoon and Young 2000) that the use of Cronbach *alpha* underestimates the reliability of test scores for a test with mixed item types. The stratified coefficient *alpha* (Qualls 1995) is defined as

$$_{strat\ \alpha}\rho_{XX'} = 1 - \frac{\sum \sigma^2_{Y_j}(1 - {}_{\alpha}\rho_{Y_jY_j'})}{\sigma^2_X}$$

where, σ^2_X = the total score variance; $\sigma^2_{Y_j}$ = the score variance for a part-test j; ${}_{\alpha}\rho_{Y_jY_j'}$ = reliability of the part-test j.

Table 9.1 presents the reliability coefficients and standard errors of measurement for mathematics and English language arts for all students and subgroups. The maximum possible score is 71 in mathematics and 93 in ELA.

TABLE 9.1

Reliability Coefficients and Standard Errors of Measurement for Raw Scores

	All Students	Males	Females	Whites	African Americans
Mathematics					
Reliability	0.94	0.94	0.93	0.93	0.92
Standard error of measurement	3.42	3.40	3.41	3.20	3.67
English Language Arts					
Reliability	0.95	0.95	0.93	0.93	0.94
Standard error of measurement	3.27	3.33	3.20	2.98	3.56

9.2 OVERALL AND CONDITIONAL STANDARD ERRORS OF MEASUREMENT

Table 9.2 presents the classical test-theory standard error of measurement (SEM) and the IRT-based conditional SEM at the scale score cutoff points. The classical SEM is defined as $s_x\sqrt{1-r_{xx}}$, where s_x is the standard deviation of the scale score and r_{xx} is the reliability coefficient. IRT-based conditional SEM at the scale score cutoff points are defined as the reciprocal of the square root of the test information function at the point on the ability continuum that corresponds to the final scale score cutoff points (Hambleton, Swaminathan, and Rogers 1991). Although classical SEM and IRT conditional SEM both serve the same role, the value of IRT-based conditional SEM varies with ability levels, whereas the classical SEM does not.

TABLE 9.2

Classical and Conditional Standard Errors of Measurement

Subject	Classical SEM	IRT-Based Conditional SEM		
		L2	L3	L4
Mathematics	6.5	5.5	5.9	7.8
English language arts	5.5	5.6	6.4	7.7

Note: The SEM metric is in scale score points.

9.3 CONSISTENCY OF ACHIEVEMENT LEVELS

When student performance is reported in terms of achievement categories, a reliability index is computed in terms of the probabilities of consistent classification of students as specified in the standard 2.15 in the Standards for Educational Psychological Testing (AERA 1999). This index considers the consistency of classifications for the percentage of examinees that would, hypothetically, be classified in the same category on a second HSAP administration using either the same form or an alternate, equivalent form.

Although a number of procedures are available for estimating misclassification errors (Livingston and Lewis 1995; Hanson and Brennan 1990; Huynh 1976; Subkoviak 1976), AIR used the *beta* binomial distribution method (Huynh 1979; Huynh, Meyer, and Barton 2000).

Table 9.3 presents a summary of agreements between the operational test classifications—that is, the percentages of students who are consistently classified in the same achievement levels on two equivalent administrations of the test.

TABLE 9.3

Consistency Indexes for Achievement Levels for the Spring 2004 HSAP Operational Test

Subject	Level 2	Level 3
Mathematics	94.3%	92.2%
English language arts	94.5%	91.7%

Chapter 10

VALIDITY

Three types of validity evidence are reported in this section: test content, item fairness, and internal structure. Evidence on content validity is presented using the distribution of item content across content areas and the alignment of the spring 2004 HSAP operational test items with reference to the state curriculum standards. Evidence on item fairness is examined with the information on differential item functioning (DIF). Evidence on internal structure is provided in correlations among content areas.

10.1 ITEM DISTRIBUTION ACROSS STRANDS

The HSAP operational test forms were constructed from the precalibrated item pools that were created based on the 2003 census field-test administration. These items measured the specific assessment standards that have been approved by the SDE. All items in the operational forms were reviewed by the CRCs and the SRC and were approved by the SDE. The spring 2004 HSAP test specifications are presented in section 4.2, above, in terms of distribution of score point values by content area.

10.2 ITEM DEVELOPMENT

All HSAP items were developed with reference to the South Carolina curriculum standards and measurement guidelines. Various committees reviewed all items; only items reviewed by these committees and approved by the SDE were included in the operational forms. The embedded field-test items in ELA were also thoroughly reviewed before being included in the operational forms. AIR reviewed the field-test items internally before sending them to the SDE for review. After the SDE's review, the items were reviewed by the CRCs and the SRC at their December 2003 meetings.

10.3 DIFFERENTIAL ITEM FUNCTIONING

A critical issue in statewide high-stakes testing is whether the test is "fair" to all test takers; therefore, an important goal of item and test development is a pool of items that are fair to all students.

All HSAP items were reviewed for bias and differential item functioning. The SRC reviewed the HSAP items for potential bias, including language that might disadvantage a group, be offensive to members of a particular group, or present obstacles to a group due to factors unrelated to content and processes specified in the standards.

After data were collected, the differential item functioning (DIF) statistics were produced for the statistical review. A psychometric definition of the term "test fairness" is the degree to which an item performs differently for one group of examinees than it performs for another group of equally able examinees. DIF refers to statistical properties of an item in two equally able groups and is subject to later interpretation and judgment. Once an item is flagged for a significant DIF, judgment should be used to decide whether the difference in difficulty shown by the DIF index is unfairly related to group membership. The DIF statistics should be seen not as indicators of bias

or unfairness but as indicators of relative strengths and weaknesses of the two groups being compared when the overall ability that the test is intended to measure has been controlled.

As with other statistical methodologies, there are numerous widely accepted approaches to detecting potential unfairness in test items. Many of these methods fall under the general category of DIF analyses. Regardless of the statistical method adopted for identifying DIF, all DIF procedures have the same goal: to detect and examine items that might favor one group or present a disadvantage to another group.

Procedure

The procedures that AIR selected for detecting DIF were the Mantel-Haenszel (MH) chi-square for dichotomous items (MC items) and Mantel's chi-square for polytomous items (CR and ER items). AIR calculated the Mantel-Haenszel statistic (MH D-DIF) for MC items (Holland and Thayer 1988) and standardized mean difference (SMD) for CR items (Zwick, Donoghue, and Grima 1993) to measure the degree and magnitude of DIF.

The examinee group of interest is the *focal* group; the group to which performance on the item is being compared is the *reference* group. In this report, the focal groups for DIF were female and African American.

Based on the DIF statistics, items were separated into one of three categories (Holland and Thayer 1988; Dorans and Holland 1993): negligible DIF (A), intermediate DIF (B), and large DIF (C). The items in category C, which exhibit significant DIF, are of primary concern.

For MC items, positive values of *delta* indicate that a given item is easier for the focal group, suggesting that the item favors the focal group. A negative value of *delta* indicates that a given item is more difficult for the focal group. Similarly, for CR items, a positive SMD value implies that, conditional on the matching variable (i.e., a total score), the focal group has a higher mean item score than the reference group, thereby favoring the focal group.

For MC items, the item classifications are based on the Mantel-Haenszel chi-square and the MH delta (Δ) value as follows:

- The item is classified as C category if the absolute value of the MH delta value (i.e., $|\Delta|$) is significantly greater than 1 and also greater than or equal to 1.5.
- The item is classified as B category if the MH delta value (Δ) is significantly different from 0 and either the absolute value of the MH delta ($|\Delta|$) is less than 1.5 or the absolute value of the MH delta ($|\Delta)|$ is not significantly different from 1.
- The item is classified as A category if the delta value (Δ) is not significantly different from 0 or the absolute value of delta ($|\Delta|$) is less than or equal to 1.

For constructed-response items, the item classifications are based on the Mantel chi-square and the SMD index as follows:

- The item is classified as C category if the Mantel chi-square *p* value is less than .05 and the absolute value of SMD divided by standard deviation of the item score (i.e., |SMD/SD|) is larger than .25.
- The item is classified as B category if the Mantel chi-square *p* value is less than .05 and the absolute value of SMD divided by standard deviation of the item score (i.e., |SMD/SD|) is larger than .17.
- All other items will be classified as A category.

The number of items in DIF categories for the spring 2004 mathematics and ELA operational items and ELA field-test items is summarized in tables 10.1 and 10.2, respectively.

When items for the operational forms were selected, each item's statistics from the initial field test were reviewed and approved by the SDE. The inclusion of any "flagged" items on an operational form (i.e., items classified as C category) was possible only when the SDE approved the inclusion of such items. For the spring 2004 operational forms, one multiple-choice item with C+ (gender) in mathematics and no items with a C category in English language arts were included.

When the operational test data were analyzed, the item with the C category (gender) in mathematics disappeared, but a different item exhibited C- in gender. In English language arts, the operational test exhibited four items with the C category.

TABLE 10.1

Summary of Differential Item Functioning for Mathematics and English Language Arts Operational Items

Subject	Item Type	Reference Group	Focal Group	Total N of Items	DIF Classification		
					A	B	C
Mathematics	Multiple choice	Male	Female	62	60	1	1
	Multiple choice	White	Black	62	62	0	0
	Constructed response	Male	Female	3	3	0	0
	Constructed response	White	Black	3	2	1	0
English language arts	Multiple choice	Male	Female	57	52	3	2
	Multiple choice	White	Black	57	49	5	3
	Constructed response	Male	Female	2	0	0	2
	Constructed response	White	Black	2	1	1	0
	Extended response	Male	Female	8	8	0	0
	Extended response	White	Black	8	6	2	0

TABLE 10.2

Summary of Differential Item Functioning for English Language Arts Field-Test Items

Subject	Item Type	Reference Group	Focal Group	Total N of Items	DIF Classification		
					A	B	C
English language arts	Multiple choice	Male	Female	80	72	6	2
	Multiple choice	White	Black	80	72	6	2

10.4 CORRELATIONS AMONG REPORTING CATEGORIES

Reporting categories for mathematics include the following five areas: Algebra (AL), Number and Operations (NO), Measurement and Geometry (MG), Data Analysis and Probability (DP), and Integrated Responses (IR). English language arts also includes five reporting categories: Reading Process and Comprehension (RC), Analysis of Texts (AT), Word Study and Analysis (WS), Research (RS), and Writing (WR). Table 10.3 reports the correlation matrices among the areas.

TABLE 10.3

Correlations among Reporting Categories

Mathematics (N = 52,913)					
Reporting Category	**NO**	**AL**	**MG**	**DP**	**IR**
NO	1.00	0.78	0.74	0.68	0.76
AL		1.00	0.74	0.67	0.74
MG			1.00	0.67	0.75
DP				1.00	0.67
IR					1.00
English Language Arts (N = 53,222)					
Reporting Category	**RC**	**AT**	**WS**	**RS**	**WR**
RC	1.00	0.75	0.70	0.64	0.66
AT		1.00	0.70	0.63	0.64
WS			1.00	0.58	0.57
RS				1.00	0.53
WR					1.00

References

AERA. 1999. *Standards for Educational and Psychological Testing: American Educational Research Association, American Psychological Association, National Council on Measurement in Education.* Washington, DC: American Educational Research Association.

Dorans, Neil J., and Paul W. Holland. 1993. "DIF Detection and Description: Mantel-Haenszel and Standardization." In *Differential Item Functioning*, edited by Paul W. Holland and Howard Wainer. Hillsdale, NJ: Lawrence Erlbaum Associates.

Feldt, Leonard S., and Robert L. Brennan. 1989. "Reliability." In *Educational Measurement*, edited by Robert L. Linn. 3rd ed. Washington, DC: American Council on Education.

Hambleton, Ronald K., Hariharan Swaminathan, and H. Jane Rogers. 1991. *Fundamentals of Item Response Theory*. Newbury Park, CA: Sage Publications.

Hanson, Bradley A., and Robert L. Brennan. 1990. "An Investigation of Classification Consistency Indexes Estimated under Alternative Strong True Score Models." *Journal of Educational Measurement* 27:345–59.

Holland, P. W., and Dorothy T. Thayer. 1988. "Differential Item Performance and the Mantel-Haenszel Procedure." In *Test Validity*, edited by Howard Wainer and Henry I. Braun. Hillsdale, NJ: Lawrence Erlbaum Associates.

Huynh, Huynh. 1976. "On the Reliability of Decisions in Domain-Referenced Testing." *Journal of Educational Measurement* 13:253–64.

———. 1979. "Computational and Statistical Inference for Two Reliability Indices Based on the Beta-Binomial Model." *Journal of Educational Statistics* 4:231–246.

Huynh, H., J. Patrick Meyer III, and Karen Barton. 2000. *Technical Documentation for the 1999 Palmetto Achievement Challenge Tests of English Language Arts and Mathematics, Grades Three through Eight.* Columbia: South Carolina Department of Education.

Lee, Guemin, and David A. Frisbie. 1999. "Estimating Reliability under a Generalizability Theory Model for Test Scores Composed of Testlets." *Applied Measurement in Education* 12, no. 3:237–55.

Linacre, John M., and Benjamin D. Wright. 2003. *WINSTEPS Rasch-Model Computer Program.* Chicago: MESA Press.

Livingston, Samuel A., and Charles Lewis. 1995. "Estimating the Consistency and Accuracy of Classifications Based on Test Scores." *Journal of Educational Measurement* 32:179–97.

Masters, Geofferey N. 1982. "A Rasch Model for Partial Credit Scoring." *Psychometrika* 47:149–74.

Millman, Jason, and Jennifer Greene. 1989. “The Specification and Development of Tests of Achievement and Ability.” In *Educational Measurement,* edited by Robert L. Linn. 3rd ed. Washington, DC: American Council on Education.

Qualls, Audrey L. 1995. “Estimating the Reliability of a Test Containing Multiple Item Formats.” *Applied Measurement in Education* 8, no. 2:111–20.

Rasch, Georg. 1980. *Probabilistic Models for Some Intelligence and Attainment Tests*. Rev. ed. Chicago: University of Chicago Press.

SDE. 2004a. *HSAP Test Administration Manual.* Columbia: South Carolina Department of Education.

———. 2004b. *HSAP District Test Coordinator’s Supplement.* Columbia: South Carolina Department of Education.

Subkoviak, Michael J. 1976. “Estimating Reliability from a Single Administration of a Criterion-Referenced Test.” *Journal of Educational Measurement* 13:265–76.

Wright, Benjamin D., and Geoff Masters. 1982. *Rating Scale Analysis*. Chicago: MESA Press.

Wright, Benjamin D., and Mark H. Stone. 1979. *Best Test Design.* Chicago: MESA Press.

Yoon, Bokhee, and Michael J. Young. 2000. “Estimating the Reliability for Test Scores with Mixed Item Formats: Internal Consistency and Generalizability.” Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Zwick, Rebecca, John R. Donoghue, and Angela Grima. 1993. “Assessment of Differential Item Functioning for Performance Tasks.” *Journal of Educational Measurement* 30, no. 3:223–51.

The South Carolina Department of Education does not discriminate on the basis of race, color, national origin, sex, or disability in admission to, treatment in, or employment in its programs and activities. Inquiries regarding the nondiscrimination policies should be made to the director of the Office of Human Resources, 1429 Senate Street, Columbia, South Carolina 29201, 803-734-8505.