

MICHIGAN MERIT EXAMINATION
Technical Manual

2009 Testing Cycle

January 27, 2010

Preface

This purpose of this manual is to document the technical characteristics of the 2009 Michigan Merit Examination (MME) based on the results of the 2009 operational administration. Analytic results are provided by Michigan's Office of Educational Assessment and Accountability (OEAA), ACT, Inc., Measurement, Inc., and Pearson Educational Measurement (PEM). This manual includes information regarding: (1) changes implemented in the 2009 MME administration, (2) background to the test, (3) test development analyses, (4) erasure analyses, (5) ACT writing scoring analyses, (6) model fit analyses, (7) scaling and equating information related to linking across MME forms, (8) reliability and validity information, (9) item analysis information, (10) standard setting information, and (11) information related to Adequate Yearly Progress and Education YES.

The Michigan Merit Examination (MME) is used to assess Grade 11 and eligible Grade 12 students on Michigan's English language arts (ELA), mathematics, science, and social studies high school content standards and expectations. It is designed differently than other statewide assessments in that the MME has three distinct components: (1) the ACT Plus Writing college entrance examination, (2) WorkKeys job skill assessments in *Reading for Information*, *Applied Mathematics*, and *Locating Information*; and (3) Michigan-specific assessments in mathematics, science, and social studies. Each component is administered on a different day. The ACT Plus Writing component is administered on Day 1, the WorkKeys component is administered on Day 2, and the Michigan component is administered on Day 3.

We encourage individuals who are interested in receiving more detailed information on topics discussed in this manual, or on related topics, to contact the Office of Educational Assessment and Accountability at the Michigan Department of Education.

Office of Educational Assessment & Accountability
Michigan Department of Education
608 W. Allegan Street
P.O. Box 30008
Lansing, MI 48909

Table of Contents

PREFACE	I
CHAPTER 1: CHANGES FOR THE MME 2009 ADMINISTRATION	1
CHAPTER 2: BACKGROUND TO THE MICHIGAN MERIT EXAMINATION	2
School Structure and MME Administration	2
Students to Be Tested	2
MME Assessment Components and Schedule.....	2
Appropriate Uses for Scores and Reports	4
School, District, Intermediate School District, and State Reports.....	5
Organizations Involved in MME Testing	7
Michigan Department of Education (MDE) Office of Educational Assessment & Accountability (OEAA)	7
Center for Educational Performance and Information (CEPI)	7
Department of Information Technology (DIT)	8
Department of Educational Technology	8
Contractors and Subcontractors	9
Educators	9
Technical Advisory Committee	11
Michigan State Board of Education	11
CHAPTER 3: TEST DEVELOPMENT (ACT, WORKKEYS, AND MICHIGAN COMPONENTS)	12
Test Development for Day 1 and Day 2	12
The ACT Plus Writing Test	12
Philosophical Basis for the ACT	12
Description of the ACT Plus Writing	13
The English Test.....	14
The Mathematics Test	14
The Reading Test.....	15
The Science Test	15
The Writing Test	16
Test Development Procedures for the ACT Multiple-Choice Tests.....	16
Reviewing Test Specifications	16
Content specifications	16
Statistical specifications	17
Selection of Item Writers.....	22
Item Construction	22
Review of Items.....	22
Item Tryouts	23
Item Analysis of Tryout Units	23
Assembly of New Forms	23
Content and Fairness Review of Test Forms.....	24
Review Following Operational Administration.....	25
Test Development Procedures for the ACT Writing Test	26
Selection and Training of Prompt Writers.....	26
Prompt Construction.....	26
Content and Fairness Review of Prompts.....	26
Field Testing of Prompts	26
Review of Field Tests and Operational Administration	26
ACT Scoring Procedures.....	26
Technical Characteristics of the ACT Tests.....	28
The <i>WorkKeys</i> Assessments Components: Reading for Information, Applied Mathematics, and Locating Information.....	28
The WorkKeys Assessment Development Process	29
Skill Definition.....	29
Test Specifications	30
Prototyping	30
Pretesting	31

Operational Forms	31
WorkKeys Assessment Descriptions	32
<i>Applied Mathematics</i>	32
<i>Reading for Information</i>	34
<i>Locating Information</i>	37
Technical Characteristics of the WorkKeys Tests	38
Test Development for Day 3	39
Test Specifications	39
Step 1: Specification Development	39
Step 2: Item Writer Training	39
Step 3: Item Development	40
Step 4: Item Review	41
Step 5: Contractor Review	41
Step 6: Field Testing	41
Step 6: Field Test Item Review	42
Field Testing Procedures: Item Development, Review, Field Test Design, and Statistics	42
Field Testing Design	42
Field Test Sampling	42
Item Specifications	42
Item Statistics	43
For Multiple-Choice (MC) Items	43
Differential Item Functioning	43
Field Testing Embedding	43
Post-Field-Test Item Review	44
Field Test Item Statistics and Data	44
Statistics Prepared for Review Committees	45
Item Reviews	48
Bias/Sensitivity and Content Committee Review	48
Item Revision Procedures	48
Item Banking	49
Procedures	49
Existing Architecture	49
Acquisition of Legacy Files	49
Populating the Item Bank	49
Data Verification Process:	50
Quality Control Procedures for Item Bank	50
Quality Control of Items:	50
Quality Control of Metadata:	50
Quality Control of Item Statistics:	50
Quality Control of Functionality:	51
Data Included in Item Bank	51
Construction of Operational Test Forms	51
Assessment Blueprints	51
Mathematics	52
Science	52
Social Studies	52
Accommodated Formats	53
Item Selection	53
Select Assessment Items to Meet the Assessment Blueprints	53
Assess the Statistical Characteristics of the Selected Assessment Items	53
For Multiple-Choice (MC) Items	53
Review and Approve Test Forms	54
Accommodated Test Forms	55
Accommodated Format Production: Day 1 ACT Plus Writing	55
Large Print	55
Oral Presentation	56
Translated and Video Formats: State-Allowed Administrations	56
Accommodated Format Production: Day 2 WorkKeys	56
Translations	57

Braille.....	57
Other media.....	58
Accommodated Format Production: Day 3 Michigan Components	58
Braille.....	58
Large Print.....	58
Oral Administration	58
Bilingual Tests	59
CHAPTER 4: ADMINISTRATION	60
Preparation for Test Administration	61
MME Day 1 and Day 2: Materials Processing	64
Materials Orders—Day 1 and Day 2.....	64
Shipping—Day 1 and Day 2	64
Day 1 Processing.....	65
Day 2 Processing.....	65
Receipt and Processing—Day 1 and Day 2	65
Test Security—Days 1 and 2	66
Materials Return—Day 1 and Day 2.....	66
Materials Discrepancy Process—Day 1 and Day 2.....	67
Processing Assessment Materials Returned by Schools—Day 1 and Day 2	67
MME Day 3 Michigan Components: Materials Processing	68
Materials Orders—Day 3	68
Shipping—Day 3	69
Processing—Day 3.....	69
Materials Receipt and Processing—Day 3.....	69
Scorable Materials—Day 3.....	69
Non-Scorable Materials—Day 3.....	70
Test Security—Day 3.....	70
Scanning/Scoring—Day 3.....	71
Data Correction.....	71
Multiple-choice Scoring.....	72
Score Reporting	72
Description of Score Reports	73
Accommodations for Students with Disabilities (SWD) and English Language	76
Learners (ELL)	76
CHAPTER 5: TEST DEVELOPMENT ANALYSES.....	95
MME Components.....	95
Test Specifications and Alignment Between Contributing Components	95
Alignment of the 2009 MME with HSCes: Item Selection for Day 1 and Day 1 Scoring	96
Test Development for Michigan Components	98
Historical Alignment Analyses Prior to 2009 Administration	100
Post-Hoc Alignment Studies of the Pilot Michigan Merit Exam	100
CHAPTER 6: ERASURE ANALYSES.....	102
Description and Purpose	102
Data and Methods.....	102
Day 1 and 2 Analysis.....	103
Overview.....	103
Input.....	104
Output	104
Day 3 Erasure Analysis	106
CHAPTER 7: ACT WRITING TRAINING AND SCORING.....	111
Results of Constructed Response Scoring Procedures	111
Rangefinding.....	111
Rater Training.....	111
Scoring.....	112

Comment Codes	112
Rater Monitoring	112
Rater Validity Checks	112
Inter-Rater Reliability	113
CHAPTER 8: MODEL FIT	114
CHAPTER 9: SCALING AND EQUATING.....	128
Quality Control Protocols for MME Calibrations	128
Equating for ACT	129
Equating for <i>WorkKeys</i>	129
Equating for MME Social Studies.....	130
Equating for MME Writing, Reading, Mathematics, Science	131
Equating for MME ELA	135
Calibration Summary Reports	135
IRT Model Fit and Plots.....	135
Item Analysis.....	135
Theta Generation.....	136
IRT Models.....	136
Algorithms for the Scoring Programs.....	136
Results of Test Runs	137
Scoring Procedures for the Spring 2009 MME Administration	137
CHAPTER 10: SCORE PRECISION	139
Internal Consistency Reliability	139
Further Evidence of Reliability on ACT Writing	139
Empirical IRT Reliability	140
MME Scale Score Reliability	140
SEM/Information Curves with Cuts Scores (Imposed).....	141
Classification Consistency and Classification Accuracy	141
CHAPTER 11: VALIDITY	145
Construct Validity Evidence from Content and Curricular Validity	145
Relation to Statewide Content Standards	145
MME Alignment Studies	146
Educator Input.....	146
Construct-related Validity Evidence from Criterion Validity Analyses.....	147
Criterion-related Validity Evidence for MME Science.....	149
DIF Analyses of the Spring 2009 MME Data.....	150
Matching Criterion.....	151
Validity Evidence for the Day 1 Stand Alone Component: ACT Assessment.....	155
Measuring Educational Achievement	155
Making Admissions Decisions.....	155
Course Placement Decisions	155
Indicators of Educational Effectiveness	156
Evaluating Probable College Success	156
Validity Evidence for the Day 2 Stand Alone Component: the WorkKeys Assessments	156
Content-Related Evidence.....	156
Criterion-Related Evidence	156
Construct-Related Evidence.....	156
Gender and Race/Ethnicity Analyses	157
Fairness Review	157
Fairness Review for the ACT.....	157
Item Writing, Review, and Pretesting	157
Operational Forms Construction	158
Differential Item Functioning (DIF).....	158
Fairness review for WorkKeys.....	159
Conclusion	159

CHAPTER 12: ITEM ANALYSIS	160
Post-Field-Test Item Review	160
Data	160
Statistics and Graphs Prepared for Review Committees	161
 CHAPTER 13: STANDARD SETTING	 169
 CHAPTER 14: ADEQUATE YEARLY PROGRESS AND EDUCATION YES	 170
Legislative Grounding	170
Procedures for Using Assessment Data for Accountability.....	170
Achievement Status	172
Achievement Change	173
 CHAPTER 15: STATE SUMMARY DATA	 179
 CHAPTER 16: MME SCALE SCORE HISTORY	 206
 REFERENCES	 207
 APPENDICES	 209
Appendix A: Plots of PARSCALE Information Functions	210
Spring 2009 Writing Initial Form	210
Spring 2009 Reading Initial Form.....	211
Spring 2009 Mathematics Initial Form	212
Spring 2009 Science Initial Form	213
Appendix B: Data Created for Field-Test Items	214
Appendix C: Statistics and Terms Used on Item Labels for Item Review Committees	233
Appendix D: Guidelines for Bias Review of Field Test Item Data	235
Appendix E: Guidelines for Content Review of Field Test Item Data	243

Chapter 1: Changes for the MME 2009 Administration

Statewide assessments undergo periodic changes in their design, administration, and reporting. In the Spring 2009 MME test cycle, there were several such changes made to the MME. These changes are detailed below.

Changes to Test Design

To ensure MME scored items align more precisely with the Michigan High School Content Standards and that the MME test booklet form taken by each student would cover the standards identically regardless of the combination of Day 1, Day 2 and Day 3 test booklet forms the student took, a subset of ACT and *WorkKeys* test items were selected and were included in the MME score. The procedures for the item alignment and item selection are presented in Chapter 5, Test Development Analysis.

For the MME Day 2 administration, the *WorkKeys Locating Information* subtest was added to the *WorkKeys* test. A subset of *WorkKeys Locating Information* items was included in the mathematics score, and the former subset of ACT science items was removed from this score.

The Michigan-developed constructed response item was eliminated from the MME social studies test. Hence, the MME Writing and MME ELA scores included only one constructed response item, the ACT essay, as the Michigan-developed essay was dropped. Also, several selected *WorkKeys Locating Information* items were added to the MME social studies score, in order to better align the social studies test with the Michigan High School Content Standards.

A matrix sampling design was employed to assemble the Michigan-developed science forms.

Changes to Administration

The administration of the Michigan-developed mathematics items was moved from Day 2 to Day 3. Now, all three Michigan-specific assessments are administered on Day 3.

There are new contractors for MME. ACT, Inc. is the new contractor for the MME scoring. They are responsible for producing the final MME score from the three components (ACT Plus Writing, *WorkKeys*, and Michigan-specific assessments), including scoring, scaling and calibration, and item analysis. Measurement Inc. is the new contractor for MME Day 3 materials production and score reporting, including all materials production for MME Day 3, scoring MME Day 3 tests and reporting those scores to ACT, and generating score reports for parents and schools.

Chapter 2: Background to the Michigan Merit Examination

School Structure and MME Administration

Michigan's 57 intermediate school districts provide leadership, services, resources, and programs to Michigan districts and schools. The intermediate school districts include more than 550 public school districts, which consist of approximately 4,500 school buildings and approximately 125,000 students per grade. There were 1,149 high schools in operation during the Spring 2009 test cycle. Public school academies (charter schools) are required to administer the MME assessments. There are approximately 190 public school academies in the state. The MME assessments are administered to all Grade 11 and eligible Grade 12 students, including those with exceptional needs and English language learners.

There are approximately 2,200 home schooled students in the state of Michigan. These students are given the opportunity to be assessed at their local public school district. MME assessments are provided on an optional basis to nonpublic schools, which include approximately 1,100 buildings with approximately 16,000 students per grade for the lower grades and 10,000 students per grade for the upper grades. Michigan law requires the state to provide assessment opportunities to middle and high school students who attend nonpublic schools that do not administer the MME assessments. This is accomplished through administration of the assessment at auxiliary test centers throughout the state. Participation of nonpublic schools and students is voluntary.

Students to Be Tested

Schools must administer all three components of the MME to all students enrolled in Grade 11 during the Spring 2009 testing window. There are two exceptions:

1. A Grade 11 student is NOT to be tested on the MME if the student's IEP indicates that the student should take MI-Access, Michigan's alternate assessment. A student who takes MI-Access in Spring 2009 may not take any portion of the MME in Spring 2009.
2. A Grade 11 student (retained or reclassified as Grade 11) is NOT to be tested on the MME if the student has taken the complete MME in a previous year and has achieved a performance level of either 1,2,3, or 4 in **each** MME subject area, including reading, writing, mathematics, science, and social studies. A student who has a reported performance level of "N/A" or a blank performance level, in any MME subject area, is considered to have not yet taken the complete MME. These students **must** take the complete MME in Spring 2009.

Michigan law now requires that the complete MME be administered to a student once and only once. A Grade 12 student is only eligible to take the MME if either of the following is true:

- The student is a first-time tester who has not previously taken the MME.
- The student has taken the MME previously but received an invalid MME score (blank or "N/A" performance level) in any of the MME subjects tested, including reading, writing, mathematics, science, or social studies.

MME Assessment Components and Schedule

The MME is composed of three distinct components: (1) the ACT Plus Writing college entrance examination, (2) WorkKeys job skills assessments in *Reading for Information*, *Applied Mathematics*, and *Locating Information*, and (3) Michigan-specific assessments in mathematics, science, and social studies. Table 2.1 presents the MME assessment components for the 2008–2009 school year.

Table 2.1: MME Components and Sections

MME Day	MME Component	Sections	Reading	Writing	Mathematics	Science	Social Studies
Day 1	ACT Plus Writing	English		S			
		Mathematics			S		
		Reading	S				
		Science				S	
		Writing		A			
Day 2	WorkKeys	Reading for Information	S				
		Applied Mathematics			S		
		Locating Information			S		S
Day 3	Michigan Component	Mathematics			A		
		Science				A	
		Social Studies					A

Note: The shaded area shows the sections in each component that contribute to a student’s MME score in each subject area. An “A” means all operational items in that section contribute to the student’s MME score, and an “S” means select items in that section contribute to the MME score.

Appropriate Uses for Scores and Reports

Following administration of the MME assessment, reports are provided to help educators understand and use the MME assessment results. Under the No Child Left Behind act, all schools are required to demonstrate that their students are making progress and that student achievement is increasing. Information from the MME helps to make that determination. The reports provide educators, parents, and the public with an understanding of the educational progress of Michigan students.

Properly used, MME assessment results can:

- measure academic achievement as compared with content expectations, and the extent to which academic achievement is improving over time;
- evaluate programs and policies designed to improve academic achievement; and
- target academic help where it is needed.

Data Reporting Guidelines and Restrictions

In September 2009, the OEAA published an updated ethics document, the *Assessment Integrity Guide*, which replaced *Professional Assessment and Accountability Practices for Educators*, published in August 2005. Section 6 of this report provides the following specific instructions for appropriate and ethical data reporting that must be followed by school personnel when using data generated from the MME assessment:

“School personnel will:

1. Understand and comply with Michigan and United States laws that apply to the handling of family privacy and student data including but not limited to the Family Educational Rights and Privacy Act (1997) and the Michigan Freedom of Information Act (1996).
2. Focus on student achievement to improve individual and program performance.
3. Maintain student confidentiality at all times.
4. Ensure that the information is reported to parents and teachers as soon as possible to determine individual strengths and weaknesses.
5. Ensure that student information is accurate before placing it in the student’s permanent records.
6. Analyze student attainment and scores in conjunction with MDE Grade Level Content Expectations, or High School Level Content expectations, and Benchmarks.
7. Analyze results in the context of the school program as a whole, not in isolation.
8. Remind the community that various factors affect test performance and factors such as the following need be taken into consideration when analyzing test results: cultural backgrounds, health conditions, economic status, and former educational experiences.

School personnel will not:

1. Expose any personally identifiable information to anyone other than the student or parents/legal guardian or designated school personnel. (Public law requires the protection of student information).
2. Report on sub-groups of students that would lead to inadvertent identification of students. State results are reported for sub-group sizes of ten students per group or more. Smaller group sizes may inadvertently expose student identities.
3. Use names, student ID numbers, birthdates, gender, race or student ID numbers which may appear on reports on any public information. Names may be used on recognized achievement awards.
4. Falsify student records to alter the accuracy of reported results.
5. Misuse or misrepresent the meaning and interpretation of any student scores.”

Brief descriptions of MME score reports are provided below. More extensive descriptions with samples are included in the *Guide to Reports*, MME. The guides also outline information about the scale score, performance level, machine-scoring process, and hand-scoring process, and include notes for interpreting score report data. *Guide to Reports* are available at OEAA website: http://www.michigan.gov/documents/mde/MME_2009_Guide_to_Reports_283213_7.pdf.

Individual Student Reports

The Individual Student Report (ISR) provides a detailed description of each student's performance in the content areas assessed on the MME. The ISR is designed to help educators identify their students' academic strengths and areas that may need improvement. Schools may include these reports in student record files.

At the top of the ISR, a student performance summary appears to the right of the student's name and demographic information. The student performance summary contains the student's scale score, and the performance level attained for each subject.

The main section of the ISR presents detailed information on the individual student's performance for each high school content standards. Selected ACT items, selected *WorkKeys* items and all Michigan operational items are included. The number of points earned out of the total number of possible points is reported for each content standard assessed. ACT and *WorkKeys* scores are also included on the ISR.

The Parent Report presents a summary description of the student's performance by high school content standard for each subject area assessed on the MME, as well as scale scores and performance level information. ACT and *WorkKeys* scores are also included on the Parent Report. One copy of the Parent Report is produced for schools to distribute to the parent/guardian.

Student Record Labels present summaries of individual scale scores and performance levels in all content areas in label format. The labels are distributed to the schools for placement in the student record files (CA-60). The Parent Report and Student Record Label are printed for each student who is administered the MME.

School, District, Intermediate School District, and State Reports

There are four different types of reports generated for schools, districts, ISDs and the state: (1) summary reports at the school, district, and state level; (2) comprehensive reports at the school and district level, provided to districts and ISDs; (3) Demographic Reports at the school, district, and state level, and (4) Student Rosters, provided to schools and summarizing information at the high school content standard level.

Summary Reports are produced at the school, district, ISD, and state level, and provide a comparative set of mean scale score information summarized by school, district, and state. The Summary Reports are generated for three student populations:

- all students;
- students with disabilities (SWD);
- all except students with disabilities (AESWD).

The top section of each Summary Report identifies the title of the report, the level of aggregation (school, district, state), the student population included in the report, the grade level and the assessment cycle, School and district names and codes are included as applicable.

The first page of the Summary Report shows summary data for each content area. The three recent years' summary data are reported, including the number of students who received valid scores, the mean scale score for each school and each previous year, the scale score margin of error, the percentage of students attaining each performance level in each year, and the percentage of students who are advanced or proficient within each subject area in each year.

The second page of the MME Summary Report includes subscore information for each student population in each subject area for the current test cycle. The subscore data include the number of students assessed, the mean points earned, the total number of possible points, and the percentage of students earning each raw score range for each content area assessed. This summary data includes aggregate and mean data for all students using the assessment form assigned to the school.

Comprehensive Reports are produced for the district and ISD and provide the mean scale score information and percentage of students attaining each performance level by subject for each school within the district and for the district as a whole. The ISD report provides aggregated data for the ISD, for each public school district, and each PSA in the ISD.

Demographic Reports provide a summary of scores by demographic subgroup for each subject area assessed. Summary data reported includes the number of students assessed in each subgroup, the mean scale score for that subgroup, the percentage of students attaining each performance level, and the percentage of students who are advanced or proficient in each subject area. The Demographic Report is generated for three student populations:

- all students;
- students with disabilities (SWD);
- all except students with disabilities (AESWD).

The top section of each Demographic Report identifies the title of the report, the level of aggregation (school, district, state), the student population included in the report, the grade level, and the assessment cycle. School and district names and codes are included as applicable.

The main section of each Demographic Report lists the demographic subgroups and the total student population reported. Ethnicity subgroups are defined by federal requirements. The remaining categories are reported by a *yes* or *no* response. No summary scores are provided for subgroups with fewer than ten students. The demographic subgroups reported are:

- gender;
- ethnicity;
- economically disadvantaged (ED);
- English language learners (ELL);
- formerly limited English proficient (FLEP);
- migrant;
- homeless
- students with accommodations. MME reports accommodation conditions for the following:
 - standard (all);
 - nonstandard (all);
 - standard (ELL only);
 - nonstandard (ELL only).

Student Rosters are distributed to schools for each subject area. The Student Rosters present summary score information by grade for each high school content standard assessed within each subject area. Each student's name, identifying information, subject area scale score, performance level, and detail information are listed.

The top section of the School Roster identifies the grade level reported, the assessment cycle, and the subject area. The school name and code, and the district name and code are also provided.

Schools are not required to submit teacher name and/or class/group code, but if a school chooses to do so, that information would be included on the Student Roster and School Roster as well.

The main section of the Student Roster lists each student's name and identifying information, each student's scale score and performance level for the subject area, as well as for each high school content standard assessed.

Organizations Involved in MME Testing

Michigan Department of Education (MDE)

Office of Educational Assessment & Accountability (OEAA)

A primary function of the Office of Educational Assessment and Accountability (OEAA), located within the Michigan Department of Education (MDE), is to establish, develop, and conduct a state assessment system that fairly and accurately reflects the state's content standards. These assessments include Michigan Educational Assessment Program (MEAP), MI-Access, MEAP-Access, English Language Proficiency Assessment, and the Michigan Merit Examination (MME).

The OEAA staff directs and manages the implementation of the statewide assessment programs. In addition to planning, scheduling, and directing all assessment activities, the staff is extensively involved in item reviews, security, and quality control procedures.

OEAA is also responsible for assessment and accountability reporting, including

- the State of Michigan's *Education Yes!*;
- the federal No Child Left Behind Act (NCLB);
- the National Assessment of Educational Progress (NAEP);
- special reports for legislators, educators, and other stakeholders;
- data for MDE programs and other state agencies;
- external research requests;
- the federal Individuals with Disabilities Education Act (IDEA).

Center for Educational Performance and Information (CEPI)

The Center for Educational Performance and Information (CEPI) collects and reports data about Michigan K–12 public schools. CEPI initiatives in data collection and reporting facilitate school districts' compliance with NCLB and the MDE's accreditation plan, *Education Yes!* CEPI is located in the Department of Management and Budget.

State and federal laws require Michigan's K–12 public schools to collect and submit data about students, educational personnel, and individual schools. Districts report their data to CEPI via the Michigan Student

Data System (MSDS)—formerly known as Single Record Student Database (SRSD), the School Infrastructure Database (SID) and the Registry of Educational Personnel (REP). These data are used for several purposes, including:

- determining state aid payments;
- calculating adequate yearly progress;
- determining school accreditation
- generating graduation/dropout rates;
- documenting teacher qualifications;
- measuring what constitutes a “safe” school.

CEPI maintains the State of Michigan’s database of school directory information, the Educational Entity Master (EEM). The Report Card data that comes from the EEM includes:

- names of superintendents and principals;
- school and district addresses and telephone numbers;
- e-mail and Web site addresses.

Department of Information Technology (DIT)

Formed in October 2001 by executive order, the Department of Information Technology (DIT) was created to centralize and consolidate the technology resources in Michigan government. DIT’s strategic plan outlines five major goals:

- expand Michigan’s services to reach anyone at any time from anywhere;
- transform Michigan services through sharing and collaboration;
- manage technology to provide better service and faster delivery;
- make Michigan a “Great Workplace” and the employer of choice for technology professionals;
- create a statewide community of partnerships.

Staff members from the DIT assist with preparing the School Report Card, including providing a process for reviewing and resolving appeals from elementary, middle, and high schools.

Department of Educational Technology

In March 2006, the MDE published the report *Leading Educational Transformation for Today’s Global Society*, which outlines Michigan’s educational technology plan. The report calls for leadership at all levels to meet a single goal: preparing Michigan students to become productive citizens in a global society. The report specifies eight objectives with strategies, performance indicators, and action steps to focus current efforts and to utilize available state- level resources.

Objective 4 states: “Every Michigan educator will use data effectively for classroom decision making and school improvement planning through an integrated local and statewide decision support system.”

The Center for Educational Performance and Information will leads a collaboration of State of Michigan agencies to plan and implement comprehensive educational data management to meet federal and state reporting requirements and time lines.

Contractors and Subcontractors

OEAA has several contractors for the MME test program. The contractors for the MME Spring 2009 test cycle include ACT Inc., Measurement Inc., Pearson Educational Measurement (PEM) and Cheeney Media Concepts Corporation (Cheeney Media).

ACT is responsible for Day 1 (ACT Plus Writing—college readiness) and Day 2 (*WorkKeys*—work skills assessment) materials, administration, and calibration, scaling, and equating, as well as the derivation of MME subject scores using scores from all three assessment components. Measurement Inc. is in charge of MME Day 3 (Michigan components), including administration, Day 3 scoring, and MME Day 3 reporting. PEM, under the direction of **OEAA**, is responsible for test development for Day 3 Michigan components.

Cheeney Media produces the accommodated formats for Day 3 Michigan components, including large print, Braille, reader scripts, audio accommodations (audio DVDs, video DVDS, audio cassettes) and translated formats in Spanish and Arabic. Subcontracted under Cheeney Media, American Printing House for the Blind, Inc. (APH) creates the Braille and enlarged print versions for the MME assessments. APH assists test developers, including state and federal departments of education, with best practices and appropriate accommodations for assessing blind and visually impaired students.

Educators

The purpose of the Michigan Merit Examination is to accurately measure and report student achievement as measured by knowledge of the Michigan high school content standards. Educators who assist in developing and administering the assessments play crucial roles in helping to achieve fair and accurate student results.

The development of the Michigan Merit Examination is a meticulous process involving thousands of Michigan administrators, teachers, and curriculum experts. The Michigan Revised School Code and the State School Aid Act require the establishment of educational standards and the assessment of students' academic achievement. Accordingly, the State Board of Education, with the input of educators throughout Michigan, approved a system of academic standards and a framework within which local school districts could develop and implement curricula.

The MME assessment is based on the state High School Content Expectations with the exception of social studies which is based on the Michigan Curriculum Framework until Spring 2011. In 2004, the State Board of Education and the Michigan Department of Education embraced the challenge to initiate a “high school redesign” project. Since then, the national call to create more rigorous learning for high school students has become a major priority for state leaders across the country. The Cherry Commission Report (2005) highlighted several goals for Michigan including the development of high school content expectations (HSCEs) that reflect both rigorous and a relevant curricular focus. Dovetailing with this call to “curricular action” is Michigan’s legislative change in high school assessment. The Michigan Merit Exam, based on rigorous high school learning standards, was implemented in 2007 and will be fully aligned with these standards by 2011.

The Michigan Department of Education’s Office of Educational Improvement and Innovation (OEII, formerly Office of School Improvement) led the development of grade level content expectations (for grades K-8) and high school content expectations (for grades 9-12). Content area work groups of academicians chaired by nationally known scholars in the respective field, were commissioned to conduct a scholarly review and identify content standards and expectations. These content standards and expectations

went through an extensive field and national review and reflect best practices and current research in the teaching and learning of respective field. They not only build from the *Michigan Curriculum Framework Standards and Benchmarks* (1996), the *Career and Employability Skills Content Standards and Benchmarks* (2001), and include the *Michigan State Board of Education's Policy on Learning Expectations for Michigan Students* (2002), but are also closely aligned with national standards and frameworks in the respective subjects. The Michigan State Board of Education approved the English Language Arts and Mathematics High School Content Expectations in April, 2006, the Science High School Content Expectations in December, 2006, and the Social Studies High School Content Expectations in October, 2007. More related information can be found at www.michigan.gov/osi.

Advisory and Review Committees

The OEAA actively seeks the input and feedback via advisory and review committees in the development and implementation of assessment and accountability systems to further the educational goal of improving what students know and can do in relation to the state content standards. Programs that utilize these committees include:

- the MME and MEAP, the assessments for most students in K–12 education programs throughout the State of Michigan;
- the alternate assessments for students with disabilities (MI-Access and MEAP-Access);
- the English Language Proficiency Assessment (ELPA), a screening assessment for students new to this country who have limited English proficiency.

This ensures that Michigan assessments are high quality and gives Michigan educators a valuable professional development opportunity that increases their familiarity with the high school content standards and thereby enhances their teaching experience.

All committees use structured, nationally recognized processes for their work to ensure that the advice provided is clear, well documented, and efficiently obtained. The time and expertise of the committee members are valued. Below are brief descriptions of some of the key committees, their purposes, and characteristics of members.

- **Content Advisory Committees (CACs)** review assessment items and key content decisions to ensure that the content of items and tests measure important elements of the state content standards in each subject and that each item is clearly worded for the students, has one clear “best” answer, and is factually consistent with the most current knowledge in the field. Based on the advice of these committees and other key information, the OEAA will accept test items for use, drop items from further consideration, or edit items. Separate Content Advisory Committees are required for each subject tested. Committee members are very familiar with the subject, the related state content standards, benchmarks, and expectations, and hold detailed knowledge of the students being assessed. Some committee members have in-depth content expertise such as mathematicians who serve on the mathematics committee. Child development experts serve on several committees. The majority of the committee members must be current teacher experts at the high school level and in the subject tested.
- **Bias and Sensitivity Review Committees (BSCs)** review each text selection, item, and writing prompt for fairness, to assure that no group is unfairly advantaged or disadvantaged compared with any other group by any MME content. The committee rejects items it considers inappropriate, suggests revisions to some, and passes on the majority of the items to the next review committee.
- **Standard Setting Committees** are charged with establishing the performance levels used to report student results on the state assessments. For the MME, these committees develop the performance level descriptors (PLDs) and recommend cut scores for the four levels. The standard setting committee is an advisory committee that looks at actual student assessments, item by item. The

committee decides the performance level these assessment results represent, using clearly defined performance level descriptions, and discusses the rationale for the decision with the other committee members. The majority of standard setting committee members must be current teacher experts at the grade and subject tested. Other committee members include administrators, curriculum specialists, counselors, parents, and business leaders. Committees represent the geographic and ethnic diversity of the state.

- **Professional Practices Committee** assisted with developing the document, *Professional Assessment and Accountability Practices for Educators*. This document, published in August 2005, presents the expected ethical conduct of educators who administer the assessments. For assessments to yield fair and accurate results, they must be given under the same standardized conditions to all students. *Professional Assessment and Accountability Practices for Educators* is intended to be used by districts and schools in the fair, appropriate, and ethical administration of the assessments.

Technical Advisory Committee

The Technical Advisory Committee (TAC) independently monitors all assessment development and implementation processes, including information gathered in field tests and review of item development. The TAC may make recommendations for revisions in design, administration, scoring, processing, or use in the examination.

The TAC was first established in 1993 to assist the MDE in developing a high school proficiency assessment as a requirement for high school graduation, required by PA 118 of 1991. At that time, the purpose of the TAC was to assist the MDE in implementing provisions of the law. The TAC continues to advise and assist the OEAA to ensure the MME assessments are developed in keeping with technical guidelines that meet national standards. The TAC is composed of individuals from Michigan and across the nation who are recognized experts in developing or reviewing high stakes assessment programs.

Michigan State Board of Education

The State Board of Education provides leadership and general supervision over all public education, including adult education and instructional programs in state institutions, with the exception of higher education institutions granting baccalaureate degrees. The State Board of Education serves as the general planning and coordinating body for all public education, including higher education, and advises the legislature concerning the financial requirements of public education.

The State Board of Education established the standards at key checkpoints during Michigan students' academic careers. With the input of educators throughout Michigan, the State Board of Education approved a system of academic standards and a framework within which local school districts could develop and implement curricula. MME assessment results show how Michigan students and schools perform compared to standards established by the State Board of Education. MME assessments are criterion-referenced assessments, meaning that student performance is measured against a set standard—in this case, the High School Content Expectations—and results are reported relative to that standard. The standards are developed by Michigan educators and approved by the State Board of Education.

Chapter 3: Test Development (ACT, *WorkKeys*, and Michigan Components)

The MME tests consist of three components (i.e., Day 1 ACT Plus Writing, Day 2 *WorkKeys* and Day 3 Michigan Component). The following section deals with the test development process for Day 1 and Day 2 affiliated with ACT. The second half of the chapter documents the test development process for the Day 3 Michigan component, which includes a wide range of related topics (e.g., item development, item review, field testing, post-test field test item review, operational test construction, item bank, and test form construction). Test alignment information is presented in Chapter 5 (“Test Development Analyses”).

Test Development for Day 1 and Day 2

The ACT Plus Writing Test

The ACT Test Program is a comprehensive system of data collection, processing, and reporting designed to help high school students develop postsecondary educational plans and to help postsecondary educational institutions meet the needs of their students. One component of the ACT Test Program is the ACT Plus Writing Test, a battery of four multiple-choice tests: English, Mathematics, Reading, and Science, and a Writing Test. The ACT Test Program also includes an interest inventory, and it collects information about students’ high school courses and grades, educational and career aspirations, extracurricular activities, and special educational needs. The ACT Plus Writing is taken under standardized conditions.

ACT Test data are used for many purposes. High schools use ACT data in academic advising and counseling, evaluation studies, accreditation documentation, and public relations. Colleges use ACT results for admissions and course placement. States use the ACT Test as part of their statewide assessment systems. Many of the agencies that provide scholarships, loans, and other types of financial assistance to students tie such assistance to students’ academic qualifications. Many state and national agencies also use ACT data to identify talented students and award scholarships.

Philosophical Basis for the ACT

Underlying the ACT tests of educational achievement is the belief that students’ preparation for college is best assessed by measuring, as directly as possible, the academic skills that they will need to perform college-level academic work. The required academic skills can be assessed most directly by reproducing as faithfully as possible the complexity of college-level work. Therefore, the tests of educational achievement are designed to determine how skillfully students solve problems, grasp implied meanings, draw inferences, evaluate ideas, and make judgments in content areas important to success in college.

Accordingly, the ACT tests of educational achievement are oriented toward the general content areas of college and high school instructional programs. The test questions require students to integrate the knowledge and skills they possess in major curriculum areas with the information provided by the test. Thus, scores on the tests have a direct and obvious relationship to the students’ educational achievement in curriculum-related areas and possess a meaning that is readily grasped by students, parents, and educators. Tests of general educational achievement are used in the ACT because, in contrast to other types of tests, they best satisfy the diverse requirements of tests used to facilitate the transition from secondary to postsecondary education. By comparison, measures of examinee knowledge of specific course content (as opposed to curriculum areas) do not readily provide a common baseline for comparing students for the purposes of admission, placement, or awarding scholarships because high school courses vary extensively.

In addition, such tests might not measure students' skills in problem solving and in the integration of knowledge from a variety of courses.

Tests of educational achievement can also be contrasted with tests of academic aptitude. The stimuli and test questions for aptitude tests are often chosen precisely for their dissimilarity to instructional materials, and each test within a battery of aptitude tests is designed to be homogeneous in psychological structure. With such an approach, these tests may not reflect the complexity of college-level work or the interactions among the skills measured. Moreover, because aptitude tests are not directly related to instruction, they may not be as useful as tests of educational achievement for making placement decisions in college.

The advantage of tests of educational achievement over other types of tests for use in the transition from high school to college becomes evident when their use is considered in the context of the educational system. Because tests of education achievement measure many of the same skills that are taught in high school, the best preparation for tests of educational achievement is high school course work. Long-term learning in school, rather than short-term cramming and coaching, becomes the best form of test preparation. Thus, tests of educational achievement tend to serve as motivators by sending students a clear message that high test scores are not simply a matter of innate ability but reflect a level of achievement that has been earned as a result of hard work.

Because the ACT stresses such general concerns as the complexity of college-level work and the integration of knowledge from a variety of sources, students may be influenced to acquire skills necessary to handle these concerns. In this way, the ACT may serve to aid high schools in developing in their students the higher-order thinking skills that are important for success in college and later life.

The tests of the ACT therefore are designed not only to accurately reflect educational goals that are widely accepted and judged by educators to be important, but also to give educational considerations, rather than statistical and empirical techniques, paramount importance.

Description of the ACT Plus Writing

The ACT Plus Writing contains four multiple-choice tests—English, Mathematics, Reading, and Science—and a Writing Test. These tests are designed to measure skills that are most important for success in postsecondary education and that are acquired in secondary education.

The content specifications describing the knowledge and skills to be measured by the ACT were determined through a detailed analysis of relevant information: First, the curriculum frameworks for grades seven through twelve were obtained for all states in the United States that had published such frameworks. Second, textbooks on state-approved lists for courses in grades seven through twelve were reviewed. Third, educators at the secondary and postsecondary levels were consulted on the importance of the knowledge and skills included in the reviewed frameworks and textbooks.

Because one of the primary purposes of the ACT is to assist in college admission decisions, in addition to taking the steps described above, ACT conducted a detailed survey to ensure the appropriateness of the content of the ACT tests for this particular use. College faculty members across the nation who were familiar with the academic skills required for successful college performance in language arts, mathematics, and science were surveyed. They were asked to rate numerous knowledge and skill areas on the basis of their importance to success in entry-level college courses and to indicate which of these areas students should be expected to master before entering the most common entry-level courses. They were also asked to identify the knowledge and skills whose mastery would qualify a student for advanced placement. A series

of consultant panels were convened, at which the experts reached consensus regarding the important knowledge and skills in English and reading, mathematics, and science, given current and expected curricular trends.

Curriculum study is ongoing at ACT. Curricula in each content area (English, reading, mathematics, science, and writing) in the ACT tests are reviewed on a periodic basis. ACT's analyses include reviews of tests, curriculum guides, and national standards; surveys of current instructional practice; and meetings with content experts (see ACT, *ACT National Curriculum Survey*[®] 2005–2006, 2007a).

The tests in the ACT are designed to be developmentally and conceptually linked to those of EXPLORE (Grades 8 and 9) and PLAN (Grade 10). To reflect that continuity, the names of the content area tests are the same across the three programs. Moreover, the programs are similar in their focus on thinking skills and in their common curriculum base. The test specifications for the ACT are consistent with, and should be seen as a logical extension of, the content and skills measured in EXPLORE and PLAN.

The English Test

The ACT English Test is a 75-item, 45-minute test that measures understanding of the conventions of standard written English (punctuation, grammar and usage, and sentence structure) and of rhetorical skills (strategy, organization, and style). Spelling, vocabulary, and rote recall of rules of grammar are not tested. The test consists of five prose passages, each accompanied by a sequence of multiple-choice test items. Different passage types are employed to provide a variety of rhetorical situations. Passages are chosen not only for their appropriateness in assessing writing skills, but also to reflect students' interests and experiences. Most items refer to underlined portions of the passage and offer several alternatives to the portion underlined. These items include "NO CHANGE" to the underlined portion in the passage as one of the possible responses. Some items are identified by a number or numbers in a box. These items ask about a section of the passage, or about the passage as a whole. The student must decide which choice is most appropriate in the context of the passage, or which choice best answers the question posed.

Three scores are reported for the English Test: a total test score based on all 75 items, a subscore in Usage/Mechanics based on 40 items, and a subscore in Rhetorical Skills based on 35 items.

The Mathematics Test

The ACT Mathematics Test is a 60-item, 60-minute test that is designed to assess the mathematical reasoning skills that students across the United States have typically acquired in courses taken up to the beginning of Grade 12. The test presents multiple-choice items that require students to use their mathematical reasoning skills to solve practical problems in mathematics. Knowledge of basic formulas and computational skills are assumed as background for the problems, but memorization of complex formulas and extensive computation are not required. The material covered on the test emphasizes the major content areas that are prerequisite to successful performance in entry-level courses in college mathematics. Six content areas are included: pre-algebra, elementary algebra, intermediate algebra, coordinate geometry, plane geometry, and trigonometry.

The items included in the Mathematics Test cover four cognitive levels: knowledge and skills, direct application, understanding concepts, and integrating conceptual understanding. "Knowledge and skills" items require the student to use one or more facts, definitions, formulas, or procedures to solve problems that are presented in purely mathematical terms. "Direct application" items require the student to use one or more facts, definitions, formulas, or procedures to solve straightforward problem sets in real-world

situations. “Understanding concepts” items test the student’s depth of understanding of major concepts by requiring reasoning from a concept to reach an inference or a conclusion. “Integrating conceptual understanding” items test the student’s ability to achieve an integrated understanding of two or more major concepts so as to solve nonroutine problems.

Calculators, although not required, are permitted for use on the Mathematics Test. Almost any four-function, scientific, or graphing calculator may be used on the Mathematics Test. A few restrictions do apply to the calculator used. These restrictions can be found in the current year’s *ACT User Handbook* or on ACT’s website at www.act.org.

Four scores are reported for the Mathematics Test: a total test score based on all 60 items, a subscore in Pre-Algebra/Elementary Algebra based on 24 items, a subscore in Intermediate Algebra/Coordinate Geometry based on 18 items, and a subscore in Plane Geometry/Trigonometry based on 18 items.

The Reading Test

The ACT Reading Test is a 40-item, 35-minute test that measures reading comprehension as a product of skill in referring and reasoning. That is, the test items require students to derive meaning from several texts by: (1) referring to what is explicitly stated and (2) reasoning to determine implicit meanings. Specifically, items ask students to use referring and reasoning skills to determine main ideas; locate and interpret significant details; understand sequences of events; make comparisons; comprehend cause-effect relationships; determine the meaning of context-dependent words, phrases, and statements; draw generalizations; and analyze the author’s or narrator’s voice or method. The test comprises four prose passages that are representative of the level and kinds of text commonly encountered in first-year college curricula; passages on topics in the social sciences, the natural sciences, prose fiction, and the humanities are included. Each passage is preceded by a heading that identifies what type of passage it is (e.g., “Prose Fiction”), names the author, and may include a brief note that helps in understanding the passage. Each passage is accompanied by a set of multiple-choice test items. These items focus on the complex of complementary and mutually supportive skills that readers must bring to bear in studying written materials across a range of subject areas. They do not test the rote recall of facts from outside the passage or rules of formal logic, nor do they contain isolated vocabulary questions.

Three scores are reported for the Reading Test: a total test score based on all 40 items, a subscore in Social Studies/Sciences reading skills (based on the 20 items in the social sciences and natural sciences sections of the test), and a subscore in Arts/Literature reading skills (based on the 20 items in the prose fiction and humanities sections of the test).

The Science Test

The ACT Science Test is a 40-item, 35-minute test that measures the interpretation, analysis, evaluation, reasoning, and problem-solving skills required in the natural sciences. The content of the Science Test is drawn from biology, chemistry, physics, and the Earth/space sciences, all of which are represented in the test. Students are assumed to have a minimum of two years of introductory science, which ACT’s National Curriculum Studies have identified as typically one year of biology and one year of physical science and/or Earth science. Thus, it is expected that students have acquired the introductory content of biology, physical science, and Earth science, are familiar with the nature of scientific inquiry, and have been exposed to laboratory investigation.

The test presents seven sets of scientific information, each followed by a number of multiple-choice test items. The scientific information is conveyed in one of three different formats: data representation (graphs, tables, and other schematic forms), research summaries (descriptions of several related experiments), or conflicting viewpoints (expressions of several related hypotheses or views that are inconsistent with one another).

The items included in the Science Test cover three cognitive levels: understanding, analysis, and generalization. “Understanding” items require students to recognize and understand the basic features of, and concepts related to, the provided information. “Analysis” items require students to examine critically the relationships between the information provided and the conclusions drawn or hypotheses developed. “Generalization” items require students to generalize from given information to gain new information, draw conclusions, or make predictions.

One score is reported for the Science Test: a total test score based on all 40 items.

The Writing Test

The ACT Writing Test is a 30-minute essay test that measures students’ writing skills—specifically those writing skills emphasized in high school English classes and in entry-level college composition courses. The test consists of one writing prompt that defines an issue and describes two points of view on that issue. The students are asked to respond to a question about their position on the issue described in the writing prompt. In doing so, they may adopt one or the other of the perspectives described in the prompt, or they may present a different point of view on the issue. The essay score is not affected by the point of view taken on the issue.

Taking the Writing Test does **not** affect a student’s score on the multiple-choice tests or the Composite score for those tests. Rather, two additional scores are provided: a Combined English/Writing score and a Writing subscore. Also provided are comments on the student’s essay.

Test Development Procedures for the ACT Multiple-Choice Tests

This section describes the procedures that are used in developing the four multiple-choice tests described above. The test development cycle required to produce each new form of the ACT tests takes as long as two and one-half years and involves several stages, beginning with a review of the test specifications.

Reviewing Test Specifications

Two types of test specifications are used in developing the ACT tests: content specifications and statistical specifications.

Content specifications

Content specifications for the ACT tests were developed through the curricular analysis discussed above. While care is taken to ensure that the basic structure of the ACT tests remains the same from year to year so that the scale scores are comparable, the specific characteristics of the test items used in each specification category are reviewed regularly. Consultant panels are convened to review both the tryout versions and the new forms of each test to verify their content accuracy and the match of the content of the tests to the content specifications. At these panels, the characteristics of the items that fulfill the content specifications are also reviewed. While the general content of the test remains constant, the particular kinds of items in a

specification category may change slightly. The basic structure of the content specifications for each of the ACT multiple-choice tests is provided in Tables 3.1 through 3.4.

Statistical specifications

Statistical specifications for the tests indicate the level of difficulty (proportion correct) and minimum acceptable level of discrimination (biserial correlation) of the test items to be used.

The tests are constructed with a target mean item difficulty of about 0.58 for the ACT population and a range of difficulties from about 0.20 to 0.89. The distribution of item difficulties was selected so that the tests will effectively differentiate among students who vary widely in their level of achievement.

With respect to discrimination indices, items should have a biserial correlation of 0.20 or higher with test scores measuring comparable content. Thus, for example, performance on mathematics items should correlate 0.20 or higher with performance on the relevant Mathematics Test subscore.

Table 3.1. Content Specifications for the ACT English Test

Six elements of effective writing are included in the English Test. These elements and the approximate proportion of the test devoted to each are given in the table.

Content/Skills	Proportion of test		Number of items
Usage/Mechanics	0.53		40
Punctuation ^a		0.13	10
Grammar and Usage ^b		0.16	12
Sentence Structure ^c		0.24	18
Rhetorical Skills	0.47		35
Strategy ^d		0.16	12
Organization ^e		0.15	11
Style ^f		0.16	12
Total	1.00		75

Scores reported: Usage/Mechanics

Rhetorical Skills

Total test score

^a*Punctuation.* The items in this category test the student's knowledge of the conventions of internal and end-of-sentence punctuation, with emphasis on the relationship of punctuation to meaning (for example, avoiding ambiguity, indicating appositives).

^b*Grammar and Usage.* The items in this category test the student's understanding of agreement between subject and verb, between pronoun and antecedent, and between modifiers and the words modified; verb formation; pronoun case; formation of comparative and superlative adjectives and adverbs; and idiomatic usage.

^c*Sentence Structure.* The items in this category test the student's understanding of relationships between and among clauses, placement of modifiers, and shifts in construction.

^d*Strategy.* The items in this category test the student's ability to develop a given topic by choosing expressions appropriate to an essay's audience and purpose; to judge the effect of adding, revising, or deleting supporting material; and to judge the relevancy of statements in context.

^e*Organization.* The items in this category test the student's ability to organize ideas and to choose effective opening, transitional, and closing sentences.

^f*Style.* The items in this category test the student's ability to select precise and appropriate words and images, to maintain the level of style and tone in an essay, to manage sentence elements for rhetorical effectiveness, and to avoid ambiguous pronoun references, wordiness, and redundancy.

Table 3.2. Content Specifications for the ACT Mathematics Test

The items in the Mathematics Test are classified with respect to six content areas. These areas and the approximate proportion of the test devoted to each are given in the table.

Content Area	Proportion of test	Number of items
Pre-Algebra ^a	0.23	14
Elementary Algebra ^b	0.17	10
Intermediate Algebra ^c	0.15	9
Coordinate Geometry ^d	0.15	9
Plane Geometry ^e	0.23	14
Trigonometry ^f	0.07	4
Total	1.00	60

Scores reported: Pre-Algebra/Elementary Algebra
 Intermediate Algebra/Coordinate Geometry
 Plane Geometry/Trigonometry
 Total test score

^a*Pre-Algebra.* Items in this content area are based on operations using whole numbers, decimals, fractions, and integers; place value; square roots and approximations; the concept of exponents; scientific notation; factors; ratio, proportion, and percent; linear equations in one variable; absolute value and ordering numbers by value; elementary counting techniques and simple probability; data collection, representation, and interpretation; and understanding simple descriptive statistics.

^b*Elementary Algebra.* Items in this content area are based on properties of exponents and square roots, evaluation of algebraic expressions through substitution, using variables to express functional relationships, understanding algebraic operations, and the solution of quadratic equations by factoring.

^c*Intermediate Algebra.* Items in this content area are based on an understanding of the quadratic formula, rational and radical expressions, absolute value equations and inequalities, sequences and patterns, systems of equations, quadratic inequalities, functions, modeling, matrices, roots of polynomials, and complex numbers.

^d*Coordinate Geometry.* Items in this content area are based on graphing and the relations between equations and graphs, including points, lines, polynomials, circles, and other curves; graphing inequalities; slope; parallel and perpendicular lines; distance; midpoints; and conics.

^e*Plane Geometry.* Items in this content area are based on the properties and relations of plane figures, including angles and relations among perpendicular and parallel lines; properties of circles, triangles, rectangles, parallelograms, and trapezoids; transformations; the concept of proof and proof techniques; volume; and applications of geometry to three dimensions.

^f*Trigonometry.* Items in this content area are based on understanding trigonometric relations in right triangles; values and properties of trigonometric functions; graphing trigonometric functions; modeling using trigonometric functions; use of trigonometric identities; and solving trigonometric equations.

Table 3.3. Content Specifications for the ACT Reading Test

The items in the Reading Test are based on the prose passages that are representative of the kinds of writing commonly encountered in college freshman curricula, including prose fiction, the social sciences, the humanities, and the natural sciences. The four content areas and the approximate proportion of the test devoted to each are given below.

Reading passage content	Proportion of test	Number of items
Prose Fiction ^a	0.25	10
Social Science ^b	0.25	10
Humanities ^c	0.25	10
Natural Science ^d	0.25	10
Total	1.00	40

Scores reported: Social Studies/Sciences (Social Science, Natural Science)
Arts/Literature (Prose Fiction, Humanities)
Total test score

^a*Prose Fiction*. The items in this category are based on short stories or excerpts from short stories or novels.

^b*Social Science*. The items in this category are based on passages in the content areas of anthropology, archaeology, biography, business, economics, education, geography, history, political science, psychology, and sociology.

^c*Humanities*. The items in this category are based on passages from memoirs and personal essays and in the content areas of architecture, art, dance, ethics, film, language, literary criticism, music, philosophy, radio, television, and theater.

^d*Natural Science*. The items in this category are based on passages in the content areas of anatomy, astronomy, biology, botany, chemistry, ecology, geology, medicine, meteorology, microbiology, natural history, physiology, physics, technology, and zoology.

Table 3.4. Content Specifications for the ACT Science Test

The Science Test is based on the type of content that is typically covered in high school science courses. Materials are drawn from the biological sciences, the Earth/space sciences, physics, and chemistry. The test emphasizes scientific reasoning skills rather than recall of specific scientific content, skill in mathematics, or skill in reading. Minimal arithmetic and algebraic computations may be required to answer some items. The three formats and the approximate proportion of the test devoted to each are given below.

Content area ^a	Format	Proportion of test	Number of items	
Biology		Data Representation ^b	0.38	15
Earth/Space Sciences		Research Summaries ^c	0.45	18
Physics		Conflicting Viewpoints ^d	0.17	7
Chemistry				
Total		1.00	40	

Score reported:

Total test score

^aAll four content areas are represented in the test. The content areas are distributed over the different formats in such a way that at least one passage, and no more than two passages, represents each content area.

^b*Data Representation.* This format presents students with graphic and tabular material similar to that found in science journals and texts. The items associated with this format measure skills such as graph reading, interpretation of scatter plots, and interpretation of information presented in tables, diagrams, and figures.

^c*Research Summaries.* This format provides students with descriptions of one or more related experiments. The items focus on the design of experiments and the interpretation of experimental results.

^d*Conflicting Viewpoints.* This format presents students with expressions of several hypotheses or views that, being based on differing premises or on incomplete data, are inconsistent with one another. The items focus on the understanding, analysis, and comparison of alternative viewpoints or hypotheses.

Selection of Item Writers

Each year, ACT contracts with item writers to construct items for the ACT. The item writers are content specialists in the disciplines measured by the ACT tests. Most are actively engaged in teaching at various levels, from high school to university, and at a variety of institutions, from small private schools to large public institutions. ACT makes every attempt to include item writers who represent the diversity of the population of the United States with respect to ethnic background, gender, and geographic location.

Before being asked to write items for the ACT tests, potential item writers are required to submit a sample set of materials for review. Each item writer receives an item writer's guide that is specific to the content area. The guides include examples of items and provide item writers with the test specifications and ACT's requirements for content and style. Included are specifications for fair portrayal of all groups of individuals, avoidance of subject matter that may be unfamiliar to members of certain groups within society, and nonsexist use of language.

Each sample set submitted by a potential item writer is evaluated by ACT Test Development staff. A decision concerning whether to contract with the item writer is made on the basis of that evaluation.

Every item writer under contract is given an assignment to produce a small number of multiple-choice items. The small size of the assignment ensures production of a diversity of material and maintenance of the security of the testing program, since any item writer will know only a small proportion of the items produced. Item writers work closely with ACT test specialists, who assist them in producing items of high quality that meet the test specifications.

Item Construction

The item writers must create items that are educationally important and psychometrically sound. A large number of items must be constructed because, even with good writers, many items fail to meet ACT's standards.

Each item writer submits a set of items, called a *unit*, in a given content area. Most Mathematics Test items are discrete (not passage-based), but occasionally some may belong to sets composed of several items based on the same paragraph or chart. All items on the English and Reading Tests are related to prose passages. All items on the Science Test are related to passages and/or other stimulus material (such as graphs and tables).

Review of Items

After a unit is accepted, it is edited to meet ACT's specifications for content accuracy, word count, item classification, item format, and language. During the editing process, all test materials are reviewed for fair portrayal and balanced representation of groups within society and for nonsexist use of language. The unit is reviewed several times by ACT staff to ensure that it meets all of ACT's standards.

Copies of each unit are then submitted to content and fairness experts for external reviews prior to the pretest administration of these units. The content review panel consists of high school teachers, curriculum specialists, and college and university faculty members. The content panel reviews the unit for content accuracy, educational importance, and grade-level appropriateness. The fairness review panel consists of experts in diverse educational areas who represent both genders and a variety of racial and ethnic

backgrounds. The fairness panel reviews the unit to help ensure fairness to all examinees. Any comments on the units by the content consultants are discussed in a panel meeting with all the content consultants and ACT staff, and appropriate changes are made to the unit(s). All fairness consultants' comments are reviewed and discussed, and appropriate changes are made to the unit(s).

Item Tryouts

The items that are judged to be acceptable in the review process are assembled into tryout units for pretesting on samples from the national examinee population. These samples are carefully selected to be representative of the total examinee population. Each sample is administered a tryout unit from one of the four academic areas covered by the ACT tests. The time limits for the tryout units permit the majority of students to respond to all items.

Item Analysis of Tryout Units

Item analyses are performed on the tryout units. For a given unit the sample is divided into low-, medium-, and high-performing groups by the individuals' scores on the ACT test in the same content area (taken at the same time as the tryout unit). The cutoff scores for the three groups are the 27th and the 73rd percentile points in the distribution of those scores. These percentile points maximize the critical ratio of the difference between the mean scores of the upper and lower groups, assuming that the standard error of measurement in each group is the same and that the scores for the entire examinee population are normally distributed (Millman & Greene, 1989).

Proportions of students in each of the groups correctly answering each tryout item are tabulated, as well as the proportion in each group selecting each of the incorrect options. Biserial and point-biserial correlation coefficients between each item score (correct/incorrect) and the total score on the corresponding test of the regular (national) test form are also computed.

Item analyses serve to identify statistically effective test items. Items that are either too difficult or too easy, and items that fail to discriminate between students of high and low educational achievement as measured by their corresponding ACT test scores, are eliminated or revised for future item tryouts. The biserial and point-biserial correlation coefficients, as well as the differences between proportions of students answering the item correctly in each of the three groups, are used as indices of the discriminating power of the tryout items.

Each item is reviewed following the item analysis. ACT staff members scrutinize items flagged for statistical reasons to identify possible problems. Some items are revised and placed in new tryout units following further review. The review process also provides feedback that helps decrease the incidence of poor quality items in the future.

Assembly of New Forms

Items that are judged acceptable in the review process are placed in an item pool. Preliminary forms of the ACT tests are constructed by selecting from this pool items that match the content and statistical specifications for the tests.

For each test in the battery, items for the new forms are selected to match the content distribution for the tests shown in Tables 3.1 through 3.4. Items are also selected to comply with the statistical specifications described in a previous section of this chapter. The distributions of item difficulty levels obtained on recent

forms of the four tests are displayed in Table 3.5. The data in Table 3.5 are taken from random samples of approximately 2,000 students from each of the six national test dates during the 2006–2007 academic year. In addition to the item difficulty distributions, item discrimination indices in the form of observed mean biserial correlations and completion rates are reported.

Table 3.5. Difficulty^a Distributions and Mean Discrimination^b Indices for ACT Test Items, 2006–2007

	Observed difficulty distributions (frequencies)			
	English	Mathematics	Reading	Science
Difficulty range				
0.00–0.09	0	0	0	0
0.10–0.19	2	20	0	3
0.20–0.29	11	37	6	11
0.30–0.39	23	46	14	34
0.40–0.49	44	51	30	45
0.50–0.59	77	65	68	49
0.60–0.69	121	58	55	35
0.70–0.79	108	57	42	37
0.80–0.89	62	24	25	24
0.90–1.00	2	2	0	2
Number of items ^c	450	360	240	240
Mean difficulty	0.63	0.52	0.60	0.56
Mean discrimination	0.54	0.60	0.56	0.49
Avg. completion rate ^d	90	90	94	95

^aDifficulty is the proportion of examinees correctly answering the item.

^bDiscrimination is the item-total score biserial correlation coefficient.

^cSix forms consisting of the following number of items per test: English 75, Mathematics 60, Reading 40, Science 40.

^dMean proportion of examinees who answered each of the last five items.

The completion rate is an indication of how “speeded” a test is for a group of students. A test is considered to be speeded if most students do not have sufficient time to answer the items in the time allotted. The completion rate reported in Table 3.5 for each test is the average completion rate for the six national test dates during the 2006–2007 academic year. The completion rate for each test is computed as the average proportion of examinees who answered each of the last five items.

Content and Fairness Review of Test Forms

The preliminary versions of the test forms are subjected to several reviews to ensure that the items are accurate and that the overall test forms are fair and conform to good test construction practice. The first review is performed by ACT staff. Items are checked for content accuracy and conformity to ACT style. The items are also reviewed to ensure that they are free of clues that could allow testwise students to answer the item correctly even though they lack knowledge in the subject areas or the required skills.

The preliminary versions of the test forms are then submitted to content and fairness experts for external review before the operational administration of the test forms. These experts are different individuals from those consulted for the content and fairness reviews of tryout units.

Two panels, a content review panel and a fairness review panel, are then convened to discuss with ACT staff the consultants' reviews of the forms. The content review panel consists of high school teachers, curriculum specialists, and college and university faculty members. The content panel reviews the forms for content accuracy, educational importance, and grade-level appropriateness. The fairness review panel consists of experts in diverse areas of education who represent both genders and a variety of racial and ethnic backgrounds. The fairness panel reviews the forms to help ensure fairness to all examinees.

After the panels complete their reviews, ACT summarizes the results. All comments from the consultants are reviewed by ACT staff members, and appropriate changes are made to the test forms. Whenever significant changes are made, the revised components are again reviewed by the appropriate consultants and by ACT staff. If no further corrections are needed, the test forms are prepared for printing.

In all, at least sixteen independent reviews are made of each test item before it appears on a national form of the ACT. The many reviews are performed to help ensure that each student's level of achievement is accurately and fairly evaluated.

Review Following Operational Administration

After each operational administration, item analysis results are reviewed for any anomalies such as substantial changes in item difficulty and discrimination indices between tryout and national administrations. Only after all anomalies have been thoroughly checked and the final scoring key approved are score reports produced. Examinees may challenge any items that they feel are questionable. Once a challenge to an item is raised and reported, the item is reviewed by content specialists in the content area assessed by the item. In the event that a problem is found with an item, actions are taken to eliminate or minimize the influence of the problem item as necessary. In all cases, the person who challenges an item is sent a letter indicating the results of the review.

Also, after each operational administration, DIF (differential item functioning) analysis procedures are conducted on the test data. DIF can be described as a statistical difference between the probability of the specific population group (the "focal" group) getting the item right and the comparison population group (the "base" group) getting the item right given that both groups have the same level of achievement with respect to the content being tested. The procedures currently used for the analysis include the standardized difference in proportion-correct (STD) procedure and the Mantel-Haenszel common odds-ratio (MH) procedure.

Both the STD and MH techniques are designed for use with multiple-choice items, and both require data from significant numbers of examinees to provide reliable results. For a description of these statistics and their performance overall in detecting DIF, see the ACT Research Report entitled *Performance of Three Conditional DIF Statistics in Detecting Differential Item Functioning on Simulated Tests* (Spray, 1989). In the analysis of items in an ACT form, large samples representing examinee groups of interest (e.g., males and females) are selected from the total number of examinees taking the test. The examinees' responses to each item on the test are analyzed using the STD and MH procedures. Compared with preestablished criteria, the items with STD or MH values exceeding the tolerance level are flagged. The flagged items are then further reviewed by the content specialists for possible explanations of the unusual STD or MH results. In the event that a problem is found with an item, actions will be taken as necessary to eliminate or minimize the influence of the problem item.

Test Development Procedures for the ACT Writing Test

This section describes the procedures that are used in developing essay prompts for the ACT Writing Test. These include many of the same stages as those used to develop the multiple-choice tests.

Selection and Training of Prompt Writers

ACT holds a prompt writing workshop each year in which new essay prompts are developed. The participants invited to take part in this prompt development process are both high school and post secondary teachers who are specialists in writing, and who represent the diversity of the U.S. population in ethnic background, gender, and geographic location.

Prompt Construction

Prompts developed for the Writing Test provide topics that not only offer adequate complexity and depth so that examinees can write a thoughtful and engaging essay, but also are within the common experiences of high school students. Topics are carefully chosen so that they are neither too vast nor simplistic, and so that they do not require specialized prior knowledge. The topics are designed so that a student should be able to respond to a topic within the 30-minute time constraint of the test.

Content and Fairness Review of Prompts

After Writing Test prompts are developed and then refined by ACT writing specialists, the prompts go through a rigorous review process by external experts. These fairness and bias experts carefully review each prompt to ensure that neither the language nor the content of a prompt will be offensive to a test taker, and that no prompt will disadvantage any student from any geographic, socioeconomic, or cultural background.

Field Testing of Prompts

New Writing Test prompts are field-tested throughout the United States every year. Students from rural and urban settings, small and large schools, and both public and private schools write responses to the new prompts, which are then read and scored by trained ACT readers.

Review of Field Tests and Operational Administration

Once scoring of the new Writing Test prompts has been completed, the prompts are analyzed for acceptability, validity, and accessibility. The new field-tested prompts are also reviewed to ensure that they are compatible with previous operational prompts, that they function in the same way as previous prompts, and that they adhere to ACT's rigorous standards.

ACT Scoring Procedures

For each of the four multiple-choice tests in the ACT (English, Mathematics, Reading, and Science), the raw scores (number of correct responses) are converted to scale scores ranging from 1 to 36.

The Composite score is the average of the four scale scores rounded to the nearest whole number (fractions of 0.5 or greater round up). The minimum Composite score is 1; the maximum is 36.

In addition to the four ACT test scores and Composite score, seven subscores are reported: two each for the English Test and the Reading Test and three for the Mathematics Test. As is done for each of the four tests, the raw scores for the subscore items are converted to scale scores. These subscores are reported on a score scale ranging from 1 to 18. The four test scores and seven subscores are derived independently of one another. The subscores in a content area do not necessarily add to the test score in that area.

In addition to the above scores, if the student took the Writing Test, the student's essay is read and scored independently by two trained readers using a six-point scoring rubric. Essays are evaluated on the evidence they demonstrate of student ability to make and articulate judgments; develop and sustain a position on an issue; organize and present ideas in a logical way; and communicate clearly and effectively using the conventions of standard written English. Essays are scored holistically—that is, on the basis of the overall impression created by all the elements of the writing. Each reader rates an essay on a scale ranging from 1 to 6. The sum of the readers' ratings is a student's Writing Test subscore on a scale ranging from 2 to 12. A student who takes the Writing Test also receives a Combined English/Writing score on a score scale ranging from 1 to 36. Writing Test results do not affect a student's Composite score.

Electronic scanning devices are used to score the four multiple-choice tests of the ACT, thus minimizing the potential for scoring errors. If a student believes that a scoring error has been made, ACT hand-scores the answer document (for a fee) upon receipt of a written request from the student. A student may arrange to be present for hand-scoring by contacting one of ACT's regional offices, but must pay whatever extra costs may be incurred in providing this special service. Strict confidentiality of each student's record is maintained.

If a student believes that a Writing Test essay has been incorrectly scored, that score may be appealed, and the essay will be reviewed and rescored (for a fee) by two new expert readers. The two new readers score the appealed essay without knowledge of the original score, and the new score is adjudicated by ACT staff writing specialists before being finalized.

For certain test dates (specified in the current year's booklet *Registering for the ACT*), examinees may obtain (upon payment of an additional fee) a copy of the test items used in determining their scores, the correct answers, a list of their answers, and a table to convert raw scores to the reported scale scores. For an additional fee, a student may also obtain a copy of his or her answer document. These materials are available only to students who test during regular administrations of the ACT on specified national test dates. If for any reason ACT must replace the test form scheduled for use at a test center, this offer is withdrawn and the student's fee for this optional service is refunded.

ACT reserves the right to cancel test scores when there is reason to believe the scores are invalid. Cases of irregularities in the test administration process—falsifying one's identity, impersonating another examinee (surrogate testing), unusual similarities in answers of examinees at the same test center, or other indicators that the test scores may not accurately reflect the examinee's level of educational achievement, including but not limited to examinee misconduct—may result in ACT's canceling the test scores. When ACT plans to cancel an examinee's test scores, it always notifies the examinee prior to taking this action. This notification includes information about the options available regarding the planned score cancellation, including procedures for appealing this decision. In all instances, the final and exclusive remedy available to examinees who want to appeal or otherwise challenge a decision by ACT to cancel their test scores is binding arbitration through written submissions to the American Arbitration Association. The issue for arbitration shall be whether ACT acted reasonably and in good faith in deciding to cancel the scores.

Technical Characteristics of the ACT Tests

ACT has conducted extensive analyses on the technical characteristics in the ACT – the score scale, norms, equating, and reliability of the tests. A carefully selected sample of examinees from one of the six national test dates each year is used as an equating sample. Scores on the alternate forms are equated to the score scale using equipercentile equating methodology. Summary statistics, based on the six national ACT administrations in 2005–2006, for scale score reliability coefficients and average standard errors of measurement for the ACT tests and subscores are given in Table 3.6. The technical characteristics of the ACT test are thoroughly documented in the ACT technical Manual (ACT, 2007b). The ACT Technical Manual can be acquired from ACT’s website at www.act.org.

Table 3.6. Scale Score Reliability and Average Standard Error of Measurement Summary Statistics for the Six National ACT Administrations in 2005–2006

Test/Subtest	Scale score reliability			Average SEM		
	Median	Minimum	Maximum	Median	Minimum	Maximum
English	.91	.89	.91	1.71	1.65	1.79
Usage/Mechanics	.86	.84	.88	1.36	1.25	1.39
Rhetorical Skills	.84	.81	.85	1.19	1.14	1.25
Mathematics	.91	.89	.92	1.47	1.43	1.56
Pre-Algebra/Elementary Algebra	.82	.81	.83	1.37	1.30	1.44
Intermediate Algebra/Coordinate Geometry	.72	.70	.75	1.47	1.38	1.54
Plane Geometry/Trigonometry	.74	.69	.78	1.52	1.34	1.66
Reading	.85	.85	.87	2.18	2.11	2.26
Social Studies/Sciences	.75	.73	.77	1.65	1.57	1.73
Arts/Literature	.77	.76	.78	1.75	1.67	1.89
Science	.80	.74	.83	2.00	1.90	2.12
Composite	.96	.95	.96	0.94	0.91	0.96

The WorkKeys Assessments Components: Reading for Information, Applied Mathematics, and Locating Information

In recent years, members of the business community as well as the general public have indicated concern that American workers, both current and future, lack the workplace skills needed to meet the challenges of rapidly evolving technical advances, organizational restructuring, and global economic competition. New jobs often require workers coming from high schools or postsecondary programs to have strong problem-solving and communication skills. Current trends in basic skill deficiencies indicate that American businesses will soon be spending more than \$25 billion a year on remedial training programs for new employees.

ACT designed *WorkKeys* to address this problem. The system serves businesses, workers, educators, and learners. As part of the development process, ACT listened to employers, educators, and experts in employment and training requirements to find out which employability skills are crucial in most jobs. Based on their insights, ACT developed the first nine *WorkKeys* skill areas: *Applied Technology*, *Applied*

Mathematics, Business Writing, Listening, Locating Information, Observation, Reading for Information, Teamwork, and Writing. Personal skills assessments have also recently been developed.

Each skill area has its own skill scale that measures both the skill requirements of specified jobs and the employability skills of individuals. Before *WorkKeys*, there were no scales that could measure both the skills a person has and the skills a job needs. Each *WorkKeys* skill scale describes a set of skill levels. This makes it possible to determine the proficiency levels students and workers already have and to design job-training programs that can help them meet the demands of the jobs they want. The *WorkKeys* system is based on the assumption that people who want to improve their skills can do so if they have enough time and appropriate instruction. Showing a direct connection between job requirements and education and training has a positive effect on learner persistence and achievement.

The WorkKeys Assessment Development Process

WorkKeys assessments are designed to cover a range of skills that is not too narrow and not too wide. If too narrow, a huge battery of tests would be needed to measure skills accurately; and if too wide, the number of items needed for validation would make the assessment too long and time-consuming. Thus, the *WorkKeys* assessments are designed to meet the following criteria:

- The way a skill is assessed is generally congruent with the way the skill is used in the workplace.
- The lowest level assessed is at approximately the lowest level for which an employer would be interested in setting a standard.
- The highest level assessed is at approximately the level beyond which specialized training would be required.
- The steps between the lowest and highest levels are large enough to be distinguished and small enough to have practical value in documenting workplace skills.
- The assessments are sufficiently reliable for high-stakes decision making.
- The assessments can be validated against empirical criteria.
- The assessments are feasible with respect to cost, administration time, and complexity.

The development process for a *WorkKeys* assessment consists of five phases: skill definition, test specifications development, prototyping, pretesting, and construction of operational forms. The process used to develop the *WorkKeys* multiple-choice test items is similar to that used for many standardized assessments including others developed by ACT (Anastasi, 1982; Crocker & Algina, 1986). Both stimuli and response alternatives meet basic requirements associated with high-quality skills.

Skill Definition

Before constructing the *WorkKeys* assessments, ACT defines the content domains and develops hierarchical *WorkKeys* skill descriptions. This process typically begins with a panel made up of employers, educators, and ACT staff. The panel first develops a broad definition of a skill area and identifies the lowest and highest level of the skill that is worthwhile to measure. The panel then identifies examples of tasks within this broadly defined skill domain and narrows that domain to those examples that are important for job performance across a wide range of jobs. Next, the tasks are organized into “strands,” which are aspects of the general skill domain, or skill area that pertain to a singular concept to be measured. The strands assessed in *Reading for Information*, for example, include “choosing main ideas or details,” “understanding word meanings,” “applying instructions,” and “applying information and reasoning.”

The strands are also divided into levels based on the variables believed to cause a task to be more or less difficult. In general, at the low end of a strand a few simple things must be attended to, whereas at the high end, many things must be attended to and a person must process information to apply it to more complex situations. In the “applying instructions” strand of *Reading for Information*, for example, employees need only apply instructions to clearly described situations at the lower levels. At the higher levels, however, employees must not only understand instructions in which the wording is more complex, meanings are more subtle, and multiple steps and conditionals are involved, but must also apply these instructions to new situations.

Test Specifications

Using the skill definitions described above, the ACT *WorkKeys* development team refines the specifications, outlining in more detail the skills the assessment will measure and how the items will become more complex as the skill levels increase. Each level is defined in terms of its characteristics, and exemplar test items are created to illustrate it. While it is sometimes appropriate to assign content to a unique level, in most cases the complexity of the stimulus and question determines the level to which a particular test item is assigned.

WorkKeys test specifications for the multiple-choice assessments are unlike the test blueprints used in education. They are not a list of the content topics or objectives to be covered and the number of test items to be assigned to each. Rather, they are more like scoring rubrics used for holistic scoring of constructed-response assessments (White, E. M., 1994). Similarly, the alternatives for a single multiple-choice question may include multiple content classifications, modeling a well-integrated curriculum, yet making the typical approach to test blueprints, which assume that each item measures only one objective, inappropriate.

Prototyping

After development of the general test specifications, ACT test development associates (TDAs) begin writing items for the prototype test. All the items must be written to meet the test specifications and must correspond to the respective skill levels of the test. A number of prototype test items sufficient to create long test form (75 items for RFI and AM, and 50 items for LI) for the skill area are produced.

Each prototype test form (one per skill area) is administered to at least two groups of high school students and two groups of employees. Typically, one group of students and one of employees will be from the same city. The second groups of students and employees will be found in another state with a different situation (for example, if the first groups are from a suburban setting, the second may be from an inner city). The number of examinees varies according to the test format, with more being used for multiple-choice tests than for constructed-response tests. Typically, at least 200 students and 60 employees are divided across the two administration sites for each multiple-choice prototype test form.

During the prototype process, TDAs interview the examinees to gather their reactions to the test instrument, which helps ACT evaluate the functioning of the test specifications. Questions such as whether the prototype items were too hard, too easy, or tested skills outside the realm of the specifications must be answered before development can move to the pretesting stage. The examinees are asked to provide comments and suggestions about the prototype test form, and educators and employers are also invited to review and comment on it. Based on all the information from prototype testing, the test specifications are adjusted if necessary, and additional prototype studies may be conducted. When the prototype process is completed satisfactorily, a written guide for item writers is prepared.

Pretesting

For the pretesting phase, ACT contracts with numerous freelance item writers who produce a large number of items, which ACT staff edit to meet the content, cognitive, and format standards. *WorkKeys* item writers must be familiar with various work situations and have insight into the use of a particular skill in different employment settings because both content and contextual accuracy are critically important for *WorkKeys*. A test question containing inaccurate content may be distracting even if the specific content does not affect the examinee's ability to respond correctly to the skills portion of the question. Inaccurate facts, improbable circumstances, or unlikely consequences of a series of procedures or actions are not acceptable. An examinee who knows about a particular workplace should not identify any of the assessment content, circumstances, procedures, or keyed responses as unlikely, inappropriate, or otherwise inaccurate.

Given the wide range of employability skills assessed, verifying content accuracy for *WorkKeys* is challenging. To help *WorkKeys* staff detect any possible problems, the item writers write a justification for the best response and for each distractor (incorrect response) for each test item. Both the items and the justifications are checked and, if necessary, the test items are modified.

After the test questions and stimuli have been created and edited, and before administration of the pretesting forms, all items are submitted to external consultants for content and fairness reviews. Qualified experts in the specific skill area being assessed, usually persons using the skills regularly on the job, check for content and contextual accuracy. Members of minority groups review the items to make sure they will not be biased against, or offensive to, racial, ethnic, and gender groups. ACT provides all the reviewers with written guidelines and receives written evaluations back from them.

To provide the data required for both classical and item response theory (IRT)-based statistics, each multiple-choice item is administered to a sample of about 2,000 examinees. For practical reasons, most of these examinees are students, although smaller samples of employees are also assessed for each pretest. Then ACT researchers evaluate the psychometric properties (such as reliability and scalability) of each item.

Additionally, statistical, differential item functioning (DIF) analyses of the items are carried out to determine whether items function differently for various groups of individuals (by seeing if responses to items can be correlated with the gender or ethnicity of the examinees). Items that show DIF are eliminated from the item pool. Based on the data collected during pretesting for each skill area, no items in the *WorkKeys* tests show DIF. Statistical studies can also locate problem items, which are identified during the analysis and are reevaluated by staff and, if necessary, outside experts.

Operational Forms

Pretest item analyses are considered carefully when constructing the forms for operational testing. Alternate and equivalent test forms for each assessment are developed from the pool of items that meet all the content, statistical, and fairness criteria. ACT staff construct at least two equivalent test forms for each assessment. In these forms, both the overall characteristics of the test and the within-level characteristics for content, complexity, and psychometric characteristics are made as similar as possible.

In addition to developing the job-profiling procedure to link the content of the *WorkKeys* assessments to a specific job, ACT achieves validity through creating well-designed tests. During the development of the assessments, ACT works to minimize the likelihood of adverse impact resulting from use of the *WorkKeys* tests. Specifically, the assessments are designed to be job-related and fair by ensuring that the items go through a series of screens before they are made available to employers:

- The assessments are criterion-referenced (they use job requirements as the scoring reference, rather than population norms);
- The test specifications are well-defined;
- Items are written by people with employment experience in the workplace and thus the items tap a domain of workplace skill;
- Items measure a particular workplace skill;
- Content and fairness experts review the items to determine possible differences in responses among racial groups and gender; and
- Statistical analyses (for example, differential item functioning) at the item and test level are conducted to monitor the performance of various subgroups.

WorkKeys Assessment Descriptions

Applied Mathematics

The *Applied Mathematics* skill involves the application of mathematical reasoning to work-related problems. The assessment requires the examinee to set up and solve the types of problems and do the types of calculations that actually occur in the workplace. This assessment is designed to be taken with a calculator. As on the job, the calculator serves as a tool for problem solving. A formula sheet that includes, but is not limited to, all formulas required for the assessment is provided. There are five skill levels, with Level 7 requiring the most complex and Level 3 requiring the least complex mathematical concepts and calculations. The details of different level descriptions can be found in the table below.

Table 3.7. Skill Definition for *Applied Mathematics*

Level	Characteristics of Items	Skills
3	<ul style="list-style-type: none"> • Translate easily from a word problem to a mathematics equation • All needed information is presented in logical order • No extra information 	<ul style="list-style-type: none"> • Solve problems that require a single type of mathematics operation (addition, subtraction, multiplication, and division) using whole numbers • Add or subtract negative numbers • Change numbers from one form to another using whole numbers, fractions, decimals, or percentages • Convert simple money and time units (e.g., hours to minutes)
4	<ul style="list-style-type: none"> • Information may be presented out of order • May include extra, unnecessary information • May include simple charts, diagrams, or graphs 	<ul style="list-style-type: none"> • Solve problems that require one or two operations • Multiply negative numbers • Calculate averages, simple ratios, simple proportions, or rates using whole numbers and decimals • Add commonly known fractions, decimals, or percentages (e.g., 1/2, .75, 25%) • Add three fractions that share a common denominator • Multiply a mixed number by a whole number or decimal • Put the information in the right order before performing calculations

<p>5</p>	<ul style="list-style-type: none"> • Problems require several steps of logic and calculation (e.g., problem may involve completing an order form by totaling the order and then computing tax) 	<ul style="list-style-type: none"> • Decide what information, calculations, or unit conversions to use to solve the problem • Look up a formula and perform single-step conversions within or between systems of measurement • Calculate using mixed units (e.g., 3.5 hours and 4 hours 30 minutes) • Divide negative numbers • Find the best deal using one- and two-step calculations and then comparing results • Calculate perimeters and areas of basic shapes (rectangles and circles) • Calculate percentage discounts or markups
<p>6</p>	<ul style="list-style-type: none"> • May require considerable translation from verbal form to mathematical expression • Generally require considerable setup and involve multiple-step calculations 	<ul style="list-style-type: none"> • Use fractions, negative numbers, ratios, percentages, or mixed numbers • Rearrange a formula before solving a problem • Use two formulas to change from one unit to another within the same system of measurement • Use two formulas to change from one unit in one system of measurement to a unit in another system of measurement • Find mistakes in items that belong at Levels 3, 4, and 5 • Find the best deal and use the result for another calculation • Find areas of basic shapes when it may be necessary to rearrange the formula, convert units of measurement in the calculations, or use the result in further calculations • Find the volume of rectangular solids • Calculate multiple rates
<p>7</p>	<ul style="list-style-type: none"> • Content or format may be unusual • Information may be incomplete or implicit • Problems often involve multiple steps of logic and calculation 	<ul style="list-style-type: none"> • Solve problems that include nonlinear functions and/or that involve more than one unknown • Find mistakes in Level 6 items • Convert between systems of measurement that involve fractions, mixed numbers, decimals, and/or percentages • Calculate multiple areas and volumes of spheres, cylinders, or cones • Set up and manipulate complex ratios or proportions • Find the best deal when there are several choices • Apply basic statistical concepts

Reading for Information

The *Reading for Information* skill involves reading and understanding work-related instructions and policies. The reading passages and questions in the assessment are based on the actual demands of the workplace. Passages take the form of memos, bulletins, notices, letters, policy manuals, and governmental regulations. Such materials differ from the expository and narrative texts used in most reading instruction, which are usually written to facilitate reading. Workplace communication is not necessarily well-written or targeted to the appropriate audience. Because the *Reading for Information* assessment uses workplace texts, the assessment is more reflective of actual workplace conditions. There are five skill levels, with Level 7 being the most complex and Level 3 the least complex. The details of different level descriptions can be found in the table below.

Table 3.8. Skill Definition for *Reading for Information*

Level	Characteristics of Stimuli and Items	Skills
3	<ul style="list-style-type: none"> • Reading materials include basic company policies, procedures, and announcements • Reading materials are short and simple, with no extra information • Reading materials tell readers what they should do • All needed information is stated clearly and directly • Items focus on the main points of the passages • Wording of the questions and answers is similar or identical to the wording used in the reading materials 	<ul style="list-style-type: none"> • Identify main ideas and clearly stated details • Choose the correct meaning of a word that is clearly defined in the reading • Choose the correct meaning of common, everyday and workplace words • Choose when to perform each step in a short series of steps • Apply instructions to a situation that is the same as the one in the reading materials
4	<ul style="list-style-type: none"> • Reading materials include company policies, procedures, and notices • Reading materials are straightforward, but have longer sentences and contain a number of details • Reading materials use common words, but do have some harder words, too • Reading materials describe procedures that include several steps • When following the procedures, individuals must think about changing conditions that affect what they should do • Questions and answers are often paraphrased from the passage 	<ul style="list-style-type: none"> • Identify important details that may not be clearly stated • Use the reading material to figure out the meaning of words that are not defined • Apply instructions with several steps to a situation that is the same as the situation in the reading materials • Choose what to do when changing conditions call for a different action (follow directions that include “if-then” statements)

<p>5</p>	<ul style="list-style-type: none"> • Policies, procedures, and announcements include all of the information needed to finish a task • Information is stated clearly and directly, but the materials have many details • Materials also include jargon, technical terms, acronyms, or words that have several meanings • Application of information given in the passage to a situation that is not specifically described in the passage • There are several considerations to be taken into account in order to choose the correct actions 	<ul style="list-style-type: none"> • Figure out the correct meaning of a word based on how the word is used • Identify the correct meaning of an acronym that is defined in the document • Identify the paraphrased definition of a technical term or jargon that is defined in the document • Apply technical terms and jargon and relate them to stated situations • Apply straightforward instructions to a new situation that is similar to the one described in the material • Apply complex instructions that include conditionals to situations described in the materials
<p>6</p>	<ul style="list-style-type: none"> • Reading materials include elaborate procedures, complicated information, and legal regulations found in all kinds of workplace documents • Complicated sentences with difficult words, jargon, and technical terms • Most of the information needed to answer the items is not clearly stated 	<ul style="list-style-type: none"> • Identify implied details • Use technical terms and jargon in new situations • Figure out the less common meaning of a word based on the context • Apply complicated instructions to new situations • Figure out the principles behind policies, rules, and procedures • Apply general principles from the materials to similar and new situations • Explain the rationale behind a procedure, policy, or communication
<p>7</p>	<ul style="list-style-type: none"> • Very complex reading materials • Information includes a lot of details • Complicated concepts • Difficult vocabulary • Unusual jargon and technical terms are used, but not defined • Writing often lacks clarity and direction • Readers must draw conclusions from some parts of the reading and apply them to other parts 	<ul style="list-style-type: none"> • Figure out the definitions of difficult, uncommon words based on how they are used • Figure out the meaning of jargon or technical terms based on how they are used • Figure out the general principles behind the policies and apply them to situations that are quite different from any described in the materials

Locating Information

The *Locating Information* skill involves the locating, comparative, summarization, and analytic skills people use when they work with work-related graphics. The types of graphics used as stimuli include tables, data graphs, forms, charts, flowcharts, diagrams, maps, floor plans, instrument gauges, and blueprints. These graphics are based on materials that reflect the range of locating information demands found in the workplace. Because the *Locating Information* assessment uses workplace graphics, the assessment is more reflective of actual workplace conditions. There are four skill levels, with Level 6 being the most complex and Level 3 the least complex. The details of different level descriptions can be found in the table below.

Table 3.9. Skill Definition for *Locating Information*

Level	Characteristics of Graphics	Skills
3	<ul style="list-style-type: none"> • Elementary graphics • Simple order forms, bar graphs, tables, flowcharts, maps, instrument gauges, and floor plans • One graphic used at a time 	<ul style="list-style-type: none"> • Find one or two pieces of information in a graphic • Fill in one or two pieces of information that are missing from a graphic
4	<ul style="list-style-type: none"> • Straightforward graphics • Basic order forms, diagrams, line graphs, tables, flowcharts, instrument gauges, and maps • One or more graphics are used at a time 	<ul style="list-style-type: none"> • Find several pieces of information in graphics • Notice how graphics are related to each other • Sum up information shown in straightforward graphics • Identify trends shown in straightforward graphics • Compare information and trends shown in straightforward graphics
5	<ul style="list-style-type: none"> • Complicated graphics with possibly unusual formats • Detailed forms, tables, graphs, diagrams, maps, and instrument gauges • One or more graphics are used at a time 	<ul style="list-style-type: none"> • Sort through distracting information • Sum up information shown in detailed graphics • Identify trends shown in detailed graphics • Compare information and trends shown in detailed graphics
6	<ul style="list-style-type: none"> • Complicated graphics containing large amounts of information; may also have challenging formats, technical terms, or symbols • Very detailed graphs, charts, tables, forms, maps, and diagrams • One or more graphics are used at a time 	<ul style="list-style-type: none"> • Analyze data in one complicated graphic or several related graphics • Apply the information to specific situations • Use the information to make decisions • Use the information to draw conclusions

Technical Characteristics of the WorkKeys Tests

ACT has conducted extensive psychometric analyses on the *WorkKeys* tests, including scaling and equating, reliability, and validity studies. Different equating methods are used in *WorkKeys*; the common-item nonequivalent groups equating method is used in MME-related work. As an important reliability index, internal consistency reliability was found to be high for the *Reading for Information* and *Applied Mathematics* tests or moderately high for the *Locating Information* test. ACT has used a multi-faceted approach to collect validity evidence of the *WorkKeys* tests. As part of criterion-related validity evidence, the studies showed positive correlations between the test scores of the three tests and job performance ratings ranging from 0.12 to 0.86, which compares favorably with the correlations found in the general research literature on criterion-related validity of employment tests. The technical characteristics—the score scale, equating, reliability, and validity—of the *WorkKeys* Tests is thoroughly documented in the *WorkKeys* Technical Manuals of respective tests (ACT, 2008a, 2008b, and 2008c). The *WorkKeys* Technical Manuals can be requested by calling 1-800/WORKKEY (967-5539) or from ACT’s website at **www.act.org**.

Test Development for Day 3

Test Specifications

As noted in the previous chapter, all MME Day 3 subject tests are based on the high school content standards. A general description of development activities for all MME Day 3 subject tests is provided below, followed by subject-specific descriptions.

MDE staff, contractors, and Michigan educators worked together to develop the tests. The test development cycle included the following steps:

- Specification Development
- Item Writer Training
- Item Development
- Item Review
- Field Testing
- Field Test Item Review
- Operational Test Construction

In addition to assessing student knowledge of subject-specific content, the MME tests also assess student thinking skills in each of the three components (i.e., Day 1 ACT Plus Writing, Day 2 *WorkKeys* and Day 3 Michigan Component). Critical thinking skills are a primary focus of each of the three components of the MME. These skills are assessed through both multiple choice (MC) and constructed-response (CR) items. CR items only appeared in ACT Writing section in the Spring 2009 test cycle. The blueprints included in the subject-matter sections of this document reflect the crossing of content with process.

Step 1: Specification Development

Following the yearly alignment process undertaken by MDE (see Chapter 5 for more information on the 2009 alignment process), MDE and its contractors develop Michigan Component test specifications. The test specifications identify the content and types of items to be included. These specifications include the High School Content Standards, general indicators of difficulty, and other psychometric characteristics as well as general physical indicators such as artwork parameters. Test item specifications are very detailed and identify content limits, item formats, and similar aspects of test items, typically including sample items of each format.

All MDE tests are designed to assess higher order thinking skills. Most items in all subject areas focus more on comprehension and application than on simple recall or recognition. Indeed, guidelines for item writing for each test clearly include admonitions for item writers to avoid simple recall of trivial or unrelated facts, and specific attention is given to ensure that tests include adequate higher-order thinking skills. The ways in which higher order thinking skills are included in each subject test is addressed by content area in the subsequent, content-specific sections.

Step 2: Item Writer Training

For Michigan-developed components, all item writers are Michigan educators who have curriculum and instruction expertise and who have been recommended by their administrators. All have relevant degrees and experience, and many have previous experience in MME-specific item writing.

Once test and item specifications are written, contractors and content consultants from OEAA use these materials to train item writers to produce items specifically for MDE. The item-writing process begins over the summer. Pearson holds one 3 day training in June for item writers, In some cases, veteran item writers are given their item writing assignments as much as a month before the actual face-to-face meeting. Teachers are trained by experienced Pearson Content Specialist team members. Teachers engage in peer review of the items and continue working on their items once they leave the training. Pearson Content Specialists provide extensive feedback to the writers and there is much back and forth in Pearson's web-based Item Authoring tool. Once the items are in-house, Pearson reviews them during various stages, to include up to five internal rounds (Content 1, Editorial 1, Content 2, Editorial 2, and Senior Review), schedule permitting.

Step 3: Item Development

The Michigan item writers draft test items as described above in accordance with specifications approved by OEAA. Once this is completed, Pearson prepares the items for the first OEAA review in September.

This internal review consists of items being evaluated using the following criteria:

Skill

- Item measures one skill level.
- Item measures skill in manner consistent with specification.
- Item uses appropriate (realistic) level of skill.
- Item makes clear the skill to be employed.

Content

- Item measures one benchmark.
- Item measures benchmark in manner consistent with specification.
- Item taps appropriate (important) aspect of content associated with benchmark.
- Item makes clear the benchmark or problem to be solved.

Relevance

- Item calls for a realistic application of process to content.
- Item is not contrived.
- Item is appropriate for the grade level to be tested.
- Item groups reflect instructional emphasis.

Accuracy

- Item is factually accurate.
- Item contains only one correct or best response.
- If item pertains to disputed content, context for correct answer is clearly defined (e.g., "According to... the correct solution is...").
- Item is unambiguously worded.

Format

- Item contains no extraneous material except as required by the benchmark.
- Vocabulary is grade-appropriate and clear.
- Item contains no errors of grammar, spelling, or mechanics.
- Item is clearly and conveniently placed on the page.
- Item contains adequate white space for calculations as needed.

- Physical arrangement of item is consistent with benchmark or common practice (e.g., horizontal vs. vertical addition and subtraction, slash vs. horizontal fraction bar, notation, symbols, etc.).
- Keys for sets of MC items are balanced (i.e., equal numbers of A's, B's, C's, and D's).

Bias

- Item is free of race and sex stereotypes.
- Item contains no material known or suspected to give advantage to any group.
- Item is free of insensitive language.
- Item sets that identify race or sex either directly or indirectly are balanced with reference to race and sex.
- Item content and format are accessible to students with disabilities.
- Item content and format are accessible to students with limited English proficiency.

Step 4: Item Review

After the internal review takes place, all items are reviewed by committees of Michigan educators and Michigan citizens. This consists of bias/sensitivity review meetings (involving 10-15 Michigan educators on-site, the Bias and Sensitivity Review Committee [BSC]) and content review meetings (involving roughly the same number of educators, the Content Advisory Committee [CAC]). This allows grade-level educators to spend more time focusing on the nuances of each item and adjusting the items when necessary. Contractor staff trains the CAC and BSC and monitors the reviews. All items are first reviewed by the BSC and then the CAC.

Any item rejected by the BSC does not get passed on the CAC for review. Each review is led by MDE and contractor staff, using prescribed guidelines and forms to indicate the final status of each item:

- **Accept:** Each of the following eight category conditions (importance, thematic, grammar, clarity, accuracy, validity, sound measurement, grade-appropriate) has been met or exceeded and the item appears suitable for field testing.
- **Modify:** One or more of the category conditions have not been met or the item needs minor changes to make it acceptable. Reviewers provide recommendations on changes to be made to the item that will make the item suitable for field testing.
- **Reject:** Several category conditions have not been met, or are suspect, or need radical changes to make the item acceptable. In such cases, the item may be vague or ambiguous, inappropriate, or not clearly related to the text or to the standard. Without severe modifications it is unlikely to be salvaged. Reviewers provide comments as to why the item should be rejected.

Step 5: Contractor Review

After this round of reviews, Pearson incorporates all changes into the items and the items are ready to field test. They are placed on forms, reviewed internally again, sent to the OEAA for review, returned to Pearson for editing and revision, returned to the OEAA for sign-off, and reviewed for overall quality control one final time before they are sent for printing.

Step 6: Field Testing

Items that have passed bias/sensitivity and content review are field tested. MME field testing is done in embedded operational test forms. All test forms consist of a certain number of operational items, along with

a number of field test items. Field test items are distributed amongst the test forms. This process is described in greater detail below.

Step 6: Field Test Item Review

After field testing, contractor staff analyzes item results and presents them to the same groups listed under Item Review above, which gives committee members the opportunity to review the items with field test statistics. Formerly, the Merit Award Board had final review authority over all test items. This function has recently shifted to the Department of Education. Once items and their field test results have been presented to the Department and the Department has accepted them, they go into banks of items from which future operational tests may be constructed. The processes for field test item reviews are presented in greater detail below, and results of field test item reviews that occurred during the 2008-2009 administration cycle are discussed later in this chapter.

Field Testing Procedures: Item Development, Review, Field Test Design, and Statistics

This section provides an overview of the field testing procedures, conducted by the development contractor. The specific item review process at various test development stages is described in other sections of Chapter 3.

Field Testing Design

OEA conducts field testing by embedding matrix-sampled field-test items across multiple forms of operational assessments such that in general each field-test item appears on only one operational form. The numbers of unique field test items embedded across 12 forms are given in Table 3.13.

Table 3.13. Number of Forms and Field Test Items by Subject

Subject	Number of Forms	Total Number of Field-test Items
Mathematics	12	154
Science	12	187
Social Studies	12	132

Field Test Sampling

It is critical that field test items be calibrated with operational items in such a way that the obtained item parameters represent those parameters that would result were the field test items administered to all students. For the MME, each form (1-10) is spiraled within each classroom in each school. Therefore, every school gets every form for some of its students, which helps to ensure that the field test item parameters are representative of those which would be obtained if the items were administered to all students.

Item Specifications

MDE employs *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999) as a primary source of guidance in the construction, field testing, and documentation of the tests. The introduction to the 1999 *Standards* best describes how those *Standards* are and were used in the

development and evaluation of tests: “Evaluating the acceptability of a test or test application does not rest on the literal satisfaction of every standard in this document, and acceptability cannot be determined by using a checklist. (*Standards*, p. 4).”

Thus, the terms ‘target’ and ‘goal’ are used when referring to various psychometric properties of the tests. For example, while it is a goal of test development for each high school test to have a reliability coefficient of .90 or greater, it is not our intention to eliminate a test with a reliability coefficient of .89. Instead, the test results would be published, along with the reliability coefficient and associated standard error of measurement.

Item Statistics

Because the MME tests are used in making individual decisions about students, they must be very reliable, particularly at cut points (the score points that separate adjacent achievement categories). Due to the fact that eligibility for Michigan Promise Scholarships¹ is involved at the high school level, the reliability at the scholarship cut score must be very high. Target reliability coefficients of .90 (or higher) are therefore set for each test. Other psychometric properties include item difficulty, item discrimination, and differential item functioning. General statistical targets are provided below.

For Multiple-Choice (MC) Items

- Percent correct: between 30 and 90 percent
- Point biserial correlation with total score: .25 or greater
- Mantel-Haenszel: Few Category C items

It should be pointed out that the point biserial correlations for MC items assume embedded field testing and employ the base test total score, which is independent of the field tested item.

Differential Item Functioning

Items that disadvantage any identifiable subgroup of students are considered biased and detract from the validity of the tests. While only human judges can determine whether or not an item is biased, item statistics can serve as a tool to help judges in their decisions. After field testing, the BSC reviews item statistics that detect differential item functioning (DIF). Specifically, Mantel-Haenszel statistics are used as measure of DIF. Mantel-Haenszel statistics are the industry standard methodology for DIF analyses, and correspond well with the categories used by ETS. These analyses are conducted after field testing. The Mantel-Haenszel statistics are generated for each item, which alert OEAA and the BSC committee to the possible presence of DIF. At this point, the BSC reviews the item further to substantiate the item statistic flag. If the item is found to indeed have DIF, it is not used in its current form in further assessments.

Field Testing Embedding

No released items are made for MME and pre-equating is not employed for the MME.

¹ Please note that, as of the spring of 2010, funding is not available for the Michigan Promise Scholarship. However, funding was available at the time of the 2009 test administration. Additionally, students are still encouraged to take the complete MME to establish their eligibility for the Promise Scholarship should funds become available.

Post-Field-Test Item Review

After field-test administration, an item review process is undertaken to evaluate the items for further use. This section describes the steps of that process, which include: (1) Preparing item statistics for internal use and for review committees, (2) internal and contractor review of statistics, (3) item review, including item statistics, by bias and content committees, and (4) potential item revisions.

Field Test Item Statistics and Data

All field-test items were embedded in the live test forms for each test. After the calibration of live test forms, field-test items were calibrated and put onto the same scale as the live operational items. The statistics for each field-test item can be summarized into nine categories.

1. General test information: test name, subject, grade, level;
2. Administration related information: year cycle, administration year, released position;
3. Specific item information: MME item ID, CID, item type, answer key, maximal score, maturity, item function, character code, number of forms the item appears on, form numbers, test position, n-count (total, male, female, white, and black students), percent for each comment code, percent for each condition code;
4. Content-related information: strand, benchmark, grade level expectation, depth of knowledge, domain, scenario;
5. Option analysis: percent for each option and each score point (total, male, female, white, and black students), p-value or item mean (total, male, female, white, and black students), adjusted p-value, difficulty flag, item standard deviation, item-total correlation, biserial/polyserial correlation, corrected point-serial correlation, item-total correlation flag, option point-biserial correlation, flag for potential miskeying;
6. DIF analysis: Mantel Chi-square, Mantel-Haenszel Delta and its standard error, signed and unsigned SMD, SMD signed effect size, DIF category, and favored group for male versus female comparison and white versus black comparison;
7. IRT parameters: b-parameter and its SE, step parameters and their respective SE, item information at cut points;
8. Fit statistics: mean-square infit, mean-square outfit, mean-square fit flag, misfit level;
9. Data for creating plots: conditional item mean for decile 1 to 10 for each student group (total, male, female, white, and black students) for creating conditional mean plots, 5th, 25th, 50th, 75th, 95th percentile for creating Box-and-Whisker plot for each student group (total, male, female, white, and black students) for each option and each score point.

The process of generating item statistics is as follows:

For Days 1 and 2, ACT and Pearson complete all scoring and produce raw scores, which they send to OEAA. For Day 3, Measurement Incorporated completes the scoring, and provides OEAA with raw scores and with any necessary erasure analyses. OEAA then creates a matched file, with data from Days 1, 2, and 3 and returns this to ACT. ACT calibrates the tests and calculates the scale scores, conducts IRT analyses and produces the statistics listed above, which they then provide to OEAA for further analyses and use by review committees and OEAA psychometricians. Finally, Measurement Incorporated produces the final score reports, using the scale scores and other information generated by ACT.

Statistics Prepared for Review Committees

From the analyses listed above, the following statistics were used to create item labels for the post-field-test reviews. Different sets of statistics were prepared for MC for review committee. Table 3.14 displays all the statistics prepared for MC items for review committee. These include six categories.

1. General administration information: test name, grade, subject, and administration time;
2. Item general information: CID, maturity, forms and positions;
3. Item specific information: item type, key, p-value, n-count, Rasch difficulty, difficulty flag, point-biserial correlation, point-biserial correlation flag, fit flag, option quality flag;
4. Breakout group descriptive statistics and optional analysis: percent of students selecting each option and omit, option point-biserial correlations, and n-count for all and subgroups: male, female, white, and black students;
5. Differential Item Functioning: flag, and favored group for male versus female and white versus black;
6. Review decision.

When the p-value for an MC item was out of the desired range, a difficulty flag was shown. When point-biserial correlation for an MC item was out of range, a point-biserial or item-total correlation flag was shown. If the DIF level for male versus female or white versus black comparison was higher than moderate, a DIF flag was turned on. When options did not function well or score point distribution was abnormal, a miskey flag was on. The criteria used for flagging an MC item are presented in Table 3.15.

Table 3.14. Item Label for a MC Item

MME Grade: 11 Subject: Math Admin: Spring 2009

ID: 3547537
Form: 0907
Position: 10

GLCE: L2.1.1

- | |
|---|
| <input type="checkbox"/> Accept as is |
| <input type="checkbox"/> Reject |
| <input type="checkbox"/> Accept with revision |

Scenario: NA

Table 1. Item Information

Type: MC	P-value: 0.11	Difficulty Flag: PL
Key: C	N-count: 10029	PB Correlation: 0.11
	Maturity: FT	PB Correlation Flag: CL
		Option Quality Flag: H P

Table 2. Breakout Group Descriptives and Option Analysis

		N-count	Percent of Students Selected Option				
			A	B	C *	D	Omit
Group	All	10029	10	71	11	8	0
	Male	4900	9	69	13	8	0
	Female	5129	11	73	8	8	0
	White	7770	9	74	10	7	0
	Black	1481	17	56	10	17	0
Option PB Correlations			-0.24	0.24	0.11	-0.25	-0.02

Table 3. Differential Item Functioning

Reference/ Focal Group	Male/ Female	White/ Black
Flag	B	
Favored Group	male	

**Note: IRT statistics were not provided during the most recent data review (July 2009).*

Table 3.15. Flagging Criteria

Statistic	Desired	Definition	Comments
P-value	>0.3 and < 0.9	The percentage of students who answered the item correctly.	Outside this range and the item may be too difficult or too easy. The desired overall mean P-Value on a test is around 0.6.
PB Correlation	> 0.25	The relationship between students' performance on an item and their performance overall on the test.	Any less than 0.25 and the item may be unreliable in discriminating well between the high and low achievers; if less, consider the response distribution and the item content.
DIF Flag	No Flag	DIF refers to the differences in performance on a studied item between the reference and the focal groups after the two groups have been matched by ability.	DIF only indicates that the examinees of equal proficiency from different subgroups have an unequal probability of responding correctly to an item. The items that exhibit DIF should be carefully examined for potential bias against particular groups.
Option Analysis	The option of key has the highest percentage	Option analysis (Score point distribution) shows the percentage of the total students and those in the gender and ethnicity subgroups who chose each option.	The keyed option should usually have the highest percentage. The keyed option point-biserial correlation should be larger than 0 while the non-key option should be smaller than 0.

Explanation of Flags:

- PL** ... p-value low
- PH** ... p-value high
- CL** ... correlation low between item and total
- B** ... moderate DIF
- C** ... substantial DIF
- H** ... highest percentage is not a keyed option
- L** ... low percentage of any option (*less than or equal to 2%*)
- P** ... positive pb-correlation for any non-keyed option
- N** ... negative pb-correlation for the keyed option
- O** ... omit has a positive pb-correlation greater than .03

Item Reviews

Bias/Sensitivity and Content Committee Review

Pearson planned and conducted Bias/Sensitivity committees (BSC) followed by Content Advisory committees (CAC) on field tested items that were flagged either because of Differential Item Functioning (DIF) or any other content related item property (see flagging criteria on Table 3.15). The goal of these committees was to identify items that are eligible to be used as scorable items in future operational assessments. In these meetings, items may be either: a) accepted “as is”, b) discarded, or occasionally c) sent back to the development contractor and OEAA for revision. OEAA identified the members of the BSCs and CACs using members previously involved in the development contractor committees.

Pearson prepared items following field testing for reviews by a Bias/Sensitivity Committee and a Content Advisory Committee. The administration contractor (Measurement, Inc.) scored all field test items and provided the raw data necessary for Pearson to generate the item statistics and data analyses necessary for BSC and CAC meetings. Pearson assembled all materials for the meetings including the items, data and analyses of the items, agenda, training materials, security agreements, sign in sheets, and the necessary records for committee sign off on each item. Each committee met for one or more days depending on the number of flagged items. The reviews were guided by checklists to ensure that the items meet the criteria for inclusion in the item bank and for potential use on future examinations. Pearson reviewed the flagging and review criteria with the OEAA to be sure that all nuances of acceptability are captured correctly. The review panel examined each item and determined if it is of high quality and matches the intended assessment objective. The items were reviewed to ensure they are appropriate for the grade level. The determination of accuracy of all material, checking that each question has only a single right answer, was part of the review process. The item statistics for each item were presented, along with a general orientation to interpretation and use of the data in item approval. The bias and sensitivity committee focused additionally on issues that ensure that the items have no stereotypical statements, present no unfair advantage or disadvantage to any group, and are free of bias for race, ethnicity, gender, age, disability status, and any other category of individuals for whom the item may be unfair. Following OEAA’s decision, these reviews occurred in face-to-face group meetings. Separate review sessions were created for content and bias/sensitivity review. Items were organized by content area and high school content expectation, and reviewers could comment individually on items or through a bulletin board discussion venue. OEAA staff had access to all items and comments during the review period. Reviewers might stop and start the review as their schedules permitted. Reports were generated that detail the reviewer's comments, and they have been evaluated by Pearson assessment specialists in conjunction with the OEAA. Pearson worked with OEAA staff to ensure that committee decisions were captured accurately.

Item Revision Procedures

It is Pearson’s policy to leave post-field test items as intact as possible since those items have data attached to them (i.e. item statistics, etc.). Making major changes can negate that data. However, there are circumstances where the item may be revised.

Generally, the field test data review committee examines items and either accepts them or rejects them. Occasionally, committee members suggest minor revisions that could improve the clarity or quality of the item. The OEAA must approve of any changes to the item, and if the committee or the OEAA believes significant changes are required to improve the item, it is rejected as ready for operational use.

The committee's recommendations are brought back to Pearson and entered into the system. At this time, the field tested items are available for use on operational forms. Items selected for operational use will be composed and reviewed by Pearson staff, then sent to the OEAA for review. Minor changes may be requested as the items are considered as a group. Once Pearson has made these revisions and the OEAA has approved the form, the Pearson quality assurance team reviews the form one final time before it is sent to the printer.

Item Banking

Procedures

After items have undergone the field testing procedures described above and are ready to become operational items, they are entered into an item bank. Pearson developed a secure item bank for grades K through 12 in the areas of Mathematics, English Language Arts, Science and Social Studies for OEAA. OEAA desires an electronic item bank software application that includes search, preview, and reporting functionality. This is a secure application, maintained by a password and other controls to be used for statewide high-risk assessment preparation.

Existing Architecture

File Maker Pro 7.0 software runs the Item Bank application. A Pearson developed Item Bank prototype has been modified to meet these requirements. The application runs as a secure web site. File Maker Pro 7.0 is a relational database with multiple tables. Each item and its assets are stored as an individual record.

Acquisition of Legacy Files

Measurement, Inc. delivered legacy items, passages, art, statistics, and metadata to Pearson. Items and previews were received as Microsoft Word documents separated by grade and status. Metadata was received as a Microsoft Excel document by grade and status. Statistics were received as an Excel document by year and season of administration. Art was received in one or more of the following formats grouped by grade level: .eps, .jpg, .tif, .ai, .cdr, .gif, .wmf, and .doc.

Populating the Item Bank

A unique, seven-digit Corporate Identification Code (CID) was assigned to each item. All passages were assigned a unique passage code. The source word document was a compilation of each item, and associated passage if applicable, at a designated grade level and status. To prepare for the population into the item bank, each item was saved as an individual word document.

The individual word documents were converted to three formats: .pdf, .gif, and .txt. The .pdf file is used to build reports. The .gif format displays the item as it appeared in the most recent test administration. The .txt document is stored, but visible to the user, in order to perform a search on any character within the item.

Data is imported in the Item Bank via Microsoft Excel spreadsheet. A unique relationship for all assets is defined by the Corporate Identification Code (CID). Image files are imported directly from the folders which they are grouped.

Data Verification Process:

The following process is followed by Pearson, Inc. to verify data entered into the item bank:

- Check for modifiable fields (exception is Comments field)
- Export data from FileMaker tables (into Excel or Access)
 - Look for blank fields
 - Look for duplicates
 - Look for anomalies in data
 - Assign a second reviewer to check data
- Compare counts to prepared list

Quality Control Procedures for Item Bank

Quality Control of Items:

- A hard copy of every item is available, typically from the test booklet or final PDF.
- A comparison of each item is completed by comparing the test booklet data and preview with the item text on the data screen and preview in the Item Bank for each item.
- The hard copy from the book only shows the reviewer the stem, art, and options.
- The item text (on the Item Data screen) does not contain any formatting. The actual content is verified as correct.
- Test Define is used to obtain the Metadata information for each item.
- A MAIN IDEA query is used to obtain the Metadata information for each item and to do a cross-check with the Metadata obtained from Test Define.
- The form number is stored in the statistics table and will be checked during the QC of the statistics.
- The Item Status and the "Item Not Appropriate" field information is verified against the "Released Position" field in the Test Define.

Quality Control of Metadata:

- To verify that data is linked together properly and that all data is accounted for, the contractor will export item metadata out of the item bank, import the data into an external database (Access), and perform queries (null, unmatched, duplicate, and cross tab).
- Missing data must be populated before a final statistical QC can take place. Once missing data has been accounted for and statistics have been populated, searches are performed based on administration (year and season), grade, and form number. Counts are verified against Test Defines for actual count of item type, item designation, and sequence numbers.
- Test booklets are compared to the set found in the item bank. Test defines may also be cross-referenced at this time. In comparing the representation of the item in the item bank to the test booklets, the item bank contractor should focus on:
 - Year, Season, Form, CID, Item Code, Grade, Item Type, Art, Passages, Item Designation (CR or FT), Item Status, Item Not Appropriate, Item Preview, and Correct Answer (Item vs. Item Preview screens, Item vs. Stats screen).

Quality Control of Item Statistics:

- Use the file from Psychometrics that contains all statistical information to QC the statistics information. This is the statistical data received from Pearson Educational Measurement.

- Export the records from the item bank (CID and all stats) into an external database (Access) and compare it with the file provided by Psychometrics. Any missing data, duplicate data, or anomaly in any of the fields is an indication that further review and verification is needed.

Quality Control of Functionality:

- Verify that the item bank functions as required:
 - Fields that should be searchable or modifiable should have that functionality.
 - Linking between tables of the item bank.
 - Built-in search functions, such as viewing items associated with a passage, etc.
 - Verify that selected sets can be saved and exported from the bank.
- A test-run of all reports should be executed to verify accurate performance.
- Verify that the saved forms, selected items, and found sets have been cleared before the final compile.

Data Included in Item Bank

The data included in the item bank is the same as the data prepared after field test administration. Please refer to the section on post field test item review to get detailed information.

An item bank performs three broad functions: (1) it acts as a repository for test items, passages, statistics, and associated attributes, such as metadata and art; (2) it permits the selection of items for test construction by multiple search criteria; and (3) it can generate a number of basic reports.

Construction of Operational Test Forms

The Michigan Office and Educational Assessment and Accountability (OEAA), Measurement Incorporated (M.I.; former contractor), and Pearson Educational Measurement (PEM) work collaboratively to develop and construct the operational test forms used to support the MME program.

Test form development entails the following steps:

- Review the assessment blueprints for the operational assessments
- Select assessment items to meet the content and process specifications of the assessment blueprints
- Assess the statistical characteristics of the selected assessment items
- Review and approve test forms

The following sections discuss essential aspects, include guidelines, and identify important references to follow through the four-step process.

Assessment Blueprints

As the name implies, the assessment blueprints identify the content and types of items to be included on the operational forms. These specifications include benchmark and content targets (limits), general indicators of difficulty and other psychometric characteristics, as well as general physical indicators such as passage length and artwork parameters.

All MME assessments are designed to assess higher order thinking skills. Most items in all subject areas focus more on comprehension and application than on simple recall or recognition. Indeed, specifications for each assessment clearly include admonitions to avoid simple recall of trivial or unrelated facts.

For 2009, the MME assessment used multiple-choice (MC) items only. Each item is aligned to a specific domain, standard, and objective. The alignment information is used during the forms construction process to help ensure the forms meet the blueprints.

This section provides an overview of the test blueprints for each subject, accommodated materials, and item specifications that guide the building of the operational test forms. The 2009 MME test contains three subject area tests: mathematics, science, and social studies. The test structures are summarized in this section.

Mathematics

The MME Mathematics Assessment is based on the Michigan High School Content Standards. For 2009, each mathematics form includes a common set of two MC items per Standard (maximum of 10 points from common items), plus a matrix of items (one item per standard), and Field Test items (as needed). Ten unique initial forms, one “makeup” form, and an accommodated form were constructed. In order to ensure comparability across all forms, each form is developed based on the carefully constructed test specifications and test development principles, outlined previously in this chapter. Equating methodologies are then used to ensure that the scales are on comparable levels (see Chapter 9 for more information on scaling and equating). These forms are spiraled within each classroom, so that all ten initial forms are distributed across schools and students. The test structure for MME mathematics assessment is summarized in Table 3.10.

Table 3.10. Test Structure for the Spring 2009 MME Mathematics Core Test

Subject	# Common Operational	#Matrix	#Field Test	Total Operational Items
Mathematics	10	10	14	20

Science

For the 2009 MME Science test, each form consists of a common set of HSCEs (one item per Standard, for a maximum of 16 points from common items), plus a matrix of items that cover the other HSCEs (one item per Standard), and Field Test items. Ten unique initial forms, one “makeup” form, and an Accommodated Form were constructed. As described in the mathematics section above, each form is comparable due to the test specifications and test development principles, and is then equated and scaled using the methodologies outlined in Chapter 9. The test structure for science tests is summarized in Table 3.11.

Table 3.11. Test Structure for the Spring 2009 MME Science Core Test

Subject	#Common Operational	#Matrix	#Field Test	Total Operational Items
Science	16	16	17	32

Social Studies

For the 2009 MME Social Studies tests, Ten unique forms, one “makeup” form, and an Accommodated Form were constructed. As described in the mathematics section above, each form is comparable due to the test specifications and test development principles, and is then equated and scaled using the methodologies outlined in Chapter 9. The test structure for social studies tests is summarized in Table 3.12.

Table 3.12. Test Structure for the Spring 2009 MME Social Studies Core Test

Subject	#MC Operational	#Matrix	#Field Test	Total Operational Items
Social Studies	28	N/A	14	28

Accommodated Formats

Each operational test is available to students who require accommodations according to their IEP, section 504 plan, or ELL instructional plan. Tests are available in Braille, large print, audio cassette, audio DVD, and video DVD. Form 12 is a unique form for accommodation for all the three components of the MME. Students testing with accommodations take the MME in sequence within a two-week accommodated testing window. For more detailed information regarding accommodated formats of the MME, see the *Spring 2010 MME Day 3 Administration Manual for Students Testing with Accommodations*, on the MME website at www.michigan.gov/mme.

Item Selection

In addition to the content coverage requirements, the forms must also meet certain statistical targets. These targets are outlined in the next three sections below.

Select Assessment Items to Meet the Assessment Blueprints

Following field testing, the items are submitted for review to both the Bias Review Committees (BRCs) and the Content Advisory Committees (CACs). These committees, composed of Michigan educators and Michigan citizens, sort the field tested items and identify which items are eligible for inclusion in the operational item pool. There is a separate pool for each subject assessed. It is from these pools that items are selected to meet the requirements outlined in the assessment blueprints.

Test forms are developed using the selected items. In addition to overarching content requirements for each test form developed, content experts and psychometricians consider requirements related to subdomains, graphics and other visual representations, passage and content dependent items, and clueing concerns.

Assess the Statistical Characteristics of the Selected Assessment Items

The statistical process begins with the work of the Content Advisory Committees and the Bias Review Committees following the field test. The committees evaluate the field test items using item statistics from classical measurement theory and item response theory models. From the work of these committees, a pool of items that are eligible to be used in constructing the operational forms is identified.

Because the MME assessments are used in making individual decisions about students, they must be very reliable, particularly at cut points (the score points that separate adjacent achievement categories). Particularly at the high school level, because Merit scholarships are involved, the reliability at the scholarship cut score must be very high. The targeted reliability coefficient is .90 (or higher) for each assessment. Other psychometric properties include item difficulty, item discrimination, and differential item functioning. General item and form level statistical targets are provided below:

For Multiple-Choice (MC) Items

- Percent correct: $.25 \leq p\text{-value} \leq .95$
- Point biserial: $\geq .25$
- Mantel-Haenszel: Few Category C items²

To help ensure adequate coverage of a full range of achievement on the operational assessments, the draft forms are evaluated to see whether the following targets are met (see Table 3.16). As necessary, items are replaced on the draft forms until this distribution is approached.

Table 3.16. Desired Range of Item Difficulty Distribution

Rasch Item Difficulty	% of items
-2.00 to -1.00	25
-0.99 to 0.00	25
0.01 to 1.00	25
1.01 to 2.00	25

Even with careful test form development, it is usually not possible to create alternate forms that are exactly equal with respect to difficulty. The MME assessments are being analyzed using Item Response Theory (IRT).

As the MME test forms are assembled, spreadsheets are used to track the statistics and other metadata (e.g., alignment) for the selected assessment items. Both classical and IRT statistics are included. The statistics listed on the spreadsheets include item p-values, correlations, and IRT item difficulties for multiple-choice items and item means, standard deviations, correlations, and IRT step difficulty estimates for constructed-response items.

The above two steps require an iterative process to create test forms that are a combination of the content and statistical information. Working together, PEM psychometricians and content experts replace items until both groups are satisfied with the forms. Through this iterative process of item selection, item content takes precedence over statistical characteristics.

Review and Approve Test Forms

Once PEM staff have reached consensus on a test form, the form and associated information is submitted to OEAA staff for review and approval. Included in the test matrices are open slots for embedded field test items. The OEAA reviews the test forms to determine whether both content and statistical requirements are met.

Guidelines for test forms review include:

- Confirm that all assessment items were accepted by the OEAA and the committees
- Confirm that all blueprint requirements are met
- Confirm that all content considerations including content/skill/topic balance, correct keys, no clueing, and correct graphics are met.
- Confirm that the item and mean difficulty levels are accurate and meet requirements
- Confirm that the assessments cover a full range of achievement levels

²For category C items, D's absolute value is significantly greater than or equal to 1.5.

As necessary, OEAA and PEM replace items that are identified by OEAA as problematic, either from a content or psychometric perspective. As items are replaced, the match of the newly revised test form to the specifications is updated and reviewed. This process continues until OEAA has approved each form.

Accommodated Test Forms

A testing accommodation is a change to the testing environment to assist a student with special needs so that assessment can mirror instruction as much as is possible without invalidating test results. District and campus testing coordinators are responsible for communicating information about testing accommodations to test administrators and other interested individuals. Information about testing accommodations is also included in the test administrator manuals.

The decision to use a particular accommodation with a student should be made on an individual basis and should take into consideration the needs of the student and whether the student routinely receives the accommodation in classroom instruction and testing. If a student receives special education services, all accommodations must be documented in the student's individualized education program (IEP), section 504 plan, or ELL instructional plan.

Typically, accommodations allow for a change in one or more of the following areas:

- Presentation format
- Test setting
- Scheduling or timing
- Response format

The following accommodated testing materials are provided for MME: Braille, Large Print, Oral Administration and Bilingual.

Accommodated Format Production: Day 1 ACT Plus Writing

For the MME Day 1 materials for the ACT Plus Writing, the Braille version is created from the unique accommodated form. This same form will be used for regular type and all alternate test formats. ACT test forms are designed from the outset according to principles of universal design, so that the tests are amenable to accommodations across the range of testing populations, conditions, and formats. ACT keeps tests as simple and straightforward as possible, consistent with curricular requirements—and this applies equally to vocabulary, graphics, typographic design, page layout, and the interrelationships among all these elements.

The accommodated form is provided to National Braille Press for production of the Braille version. NBP is responsible for Braille transcription and creation of the raised line drawings included in the booklet. ACT does an additional proof of the Raised Line Drawings, but otherwise, NBP is responsible for all quality control checks.

Large Print

The Large Print format is developed from the unique accommodated form. ACT generally maintains the item layout of the regular type test booklet where possible; sometimes the layout of certain enlarged graphics must be adjusted so that the graphics do not cross over the binding and become obscured, ACT standard is 18-point font for large type. ACT produces and proofs the copy in-house before delivering it to The Brandt Company for printing. Brandt performs quality control checks in addition to the ones at ACT.

Oral Presentation

Students approved for oral presentation have the tests read to them either “live” or from a recording, in the three formats outlined below.

Reader Scripts

Reader Scripts are used when a student will have the test read by a qualified member of the testing staff individually in a separate room. The Reader Script is created from the tapescript (see below) once the audiocassette masters have been approved. These scripts include detailed instructions to the reader on administration procedures, how items are to be read, and guidelines to ensure a standardized administration no matter who is reading the tests. Reader Scripts are created from the unique accommodated form by ACT Test Development and proofed extensively before being delivered to RR Donnelley for printing. Reader Scripts are currently scanned directly from final camera-ready copy.

Cassettes and Audio DVDs

The audio recordings for cassettes and Audio DVDs are created from the unique accommodated form, using a tapescript written by ACT Test Development. The narrator is chosen by ACT and the same recording is used for both cassette and audio DVD formats. The audio recordings include a recitation of each item, as well as a recitation of all directions (stop, turn the page, etc.). They also include instructions for students on how to recheck their work or refer to passages in the test booklet students follow along with as needed. The cassettes are created first, and once the masters are approved, a digital file is delivered to the audio DVD vendor to perform “tracking” that is unique to the audio DVDs. Tracking the discs enables students to efficiently refer back to items and recheck their work. Cassettes and Audio DVDs are only available in English for the ACT Plus Writing.

Translated and Video Formats: State-Allowed Administrations.

The ACT Plus Writing is also available as an English Video DVD. The video component consists of a test booklet on the screen, intertitles preceding the questions, and prominent arrows that follow the screen text in sync with the audio component, as a visual aid to students. Additionally, the ACT Mathematics and Science Tests, along with the directions for all tests, are translated into Spanish and Arabic, the top language groups represented in the state after English.

This translation is done from the unique accommodated form and is also presented as a video DVD using the same English video component as described above. ACT’s subcontractor, Metro Studios, contracts out the translations, and is also responsible for synching the English audio, translated audio, and English video components together. The translation team consists of a primary translator who also narrates the tests, a spotter who ensures the translated test is narrated exactly as shown in the Reader Script, and a proofer who compares the finished recording to the English version of the test and identifies any translation errors or questions. Metro Studios facilitates any necessary discussion between the original translator and the proofer, and revisions are made as needed. Metro Studios has primary responsibility for translation accuracy and performs quality control checks for all three video formats. Students testing with a translated format receive the accommodated test booklet, printed in English, with which to follow along. This booklet matches the one displayed on-screen in the video.

Accommodated Format Production: Day 2 WorkKeys

In a particular administration, initial testing, make-up testing, and accommodated testing typically have different sets of questions. The test forms, however, are built to identical specifications and are fully equated to the other test forms administered in that testing situation as well as to forms used by the general population. Regardless of the accommodation, the same test form is used for the translated forms, Braille forms, large print forms, reader scripts, and other accommodations.

Translations

The International Test Commission (www.intestcom.org/itc_projects.htm) has developed guidelines for test adaptation, especially across cultures. The guidelines reference *Adapting Educational and Psychological Tests for Cross-Cultural Assessment* by Ronald K. Hambleton regarding advantages and disadvantages of various translation methods. ACT's WorkKeys Development team has chosen to use the back translation method of quality control.

ACT works with a local company to produce translations of the WorkKeys accommodated forms. They use another outside company for the Spanish translations and a separate source for the Arabic translations. Back translations are done by different personnel than those doing the original translation, for quality control. The WorkKeys editor (MA, Foreign Language Education) then compares the back translation with the original reader script done by WorkKeys personnel.

The company doing the Spanish translations has three staff members whose native language is Spanish, two of those with college degrees in communications and languages. The Arabic translation is also done by a staff member whose native language is Arabic, is a certified Arabic teacher, and is chair of a department for English Language Learners.

Braille

ACT currently uses two vendors for the WorkKeys assessments. Both vendors follow the codes set forth by the Braille Authority of North America (BANA) and the guidelines for proofreading as used for the National Library Service. You can find the NLS specs and guidelines at: [http://www.loc.gov/nls/specs under Spec 800](http://www.loc.gov/nls/specs_under_Spec_800).

In short the procedures for proofreading Braille are to have the document read after translation by a team consisting of a sighted person and a blind person. The blind person reads aloud to the sighted person who follows the print. When that is completed, a correction sheet is returned to the translator. After those corrections are made, it goes to a different team to be re-read. All the text is re-read, keeping in mind the corrections from the first reading. After the test is corrected from the second reading and the corrections are approved, it goes to a QC person for review. The QC person compares the Braille to the hardcopy checking for any possible errors, which sometimes might be formatting Braille. When the document is approved, it is sent to have TA notes written, if applicable. The TA notes are then checked by the translator. When the test goes to the production floor, a percentage of the tests are checked by the same QC person before the final copies are produced.

It should also be noted that both vendors use transcribers and proofreaders that have been certified by the Library of Congress.

ACT also receives a proof copy of the Braille document. We review all tactile graphics and visually compare them to the original art to make sure everything is included.

Other media

Reader scripts are prepared by ACT for appropriate WorkKeys assessments, indicating how each item should be read aloud (e.g., pronunciations of names, format of numbers, mathematical statements). Audio recordings are created using the reader script with a “spotter” following the script during the recording, and a “proof” copy is created for further checking by a WorkKeys editor for exact match. For large print materials, text is typically enlarged 130%. As noted, all materials undergo several quality control checks.

Accommodated Format Production: Day 3 Michigan Components

Braille

For the MME Day 3 materials, the Braille test is created from the unique accommodated form. Items for the accommodated form are selected specifically because of their adaptability to Braille (in addition to meeting test specifications). Doing this ensures that items do not have to be dropped from the Braille form and replaced with other items, which helps ensure the comparability of the Braille form to the accommodated form, and thus to the other test forms.

Once the unique accommodated form is produced, it is provided to an independent subcontractor, Cheeney Media Corporation, who is responsible for the production of all accommodated formats. Cheeney in turn subcontracts the Braille and production of the Braille form to American Printing House. After American Printing House finishes translating the form into Braille, Cheeney Media conducts the appropriate quality control checks.

Large Print

Like the Braille format, the Large Print format is developed from the unique accommodated form. The items on this form are screened for adaptability to large print. Text is enlarged to one of four font sizes based on the degree of visual impairment. The font sizes offered reflect the sizes of print being used in current instructional situations. Mathematics diagrams requiring measurement are not enlarged.

Cheeney Media Corporation subcontracts the production of this format to American Printing House as well, and performs the appropriate quality control checks after American Printing House produces the Large Print forms.

Oral Administration

Students may have oral administrations by having a test administrator read the script aloud or by using a pre-recorded audio version of the scripted test.

Reader Scripts

Reader Scripts are created for each test component for each day, indicating exactly how each item should be read aloud without compromising the quality of the item. For example, if a problem requires students to indicate the largest number, the answers would not be read aloud. These scripts include detailed instructions to the reader regarding how to administer the assessment fairly. They also include phonetic spelling and other guidelines to ensure that each Reader reads the script in exactly the same way, as this is important for a standardized administration. Reader Scripts are produced from the unique accommodated format, and are carefully checked by Cheeney Media Corporation and by OEAA for accuracy (i.e. are all of the items the same on the accommodated form and in the Reader Script? Are there any errors in the spoken specification?)

Audio Recording

The audio recordings are created from the unique accommodated form, using the Reader Script as a script. The audio recordings include a recitation of each item, as well as a recitation of all direction (stop, turn the page, etc.). They also include instructions regarding how to review items if necessary. Audio recordings are available in both English, Spanish and Arabic.

Bilingual Tests

The MME is printed in English, and is translated into Spanish and Arabic, the top language groups represented in the state after English. This translation is done from the unique accommodated form. For MME Day 3, Cheeney Media Corporation uses an independent subcontractor to perform the initial translations. These translations are then re-translated by a separate independent subcontractor to ensure accuracy. If there are any discrepancies, Cheeney Media facilitates the discussion between translators and produces a maximally accurate translation.

For the test administration, students receive an accommodated form with the questions printed in English, but are then provided with a DVD with the translation in Spanish or Arabic. Students may have the test interpreted on the day of testing for languages where a recorded bilingual version is not available.

Chapter 4: Administration

A valid and reliable MME assessment requires that assessments are aligned with the Michigan High School Content Standards and then administered and scored according to sound measurement principles. The MME is composed of three primary elements:

- Day 1 – The ACT Plus Writing
- Day 2 – *WorkKeys (Reading for Information, Applied Mathematics, and Locating Information)*
- Day 3 – Michigan-specific assessments in mathematics, science, and social studies

Sound assessment practices require that schools administer all assessments in a consistent manner across the state so that all students have a fair and equitable opportunity for MME scores that accurately reflect their achievement in each of the MME content areas:

- Total English Language Arts
Reading
Writing
- Mathematics
- Science
- Social Studies

The schools play a key role in administering the MME assessment in a manner consistent with established procedures, monitoring the fair administration of the assessment, and working with OEAA to address deviations from established assessment administration procedures. School Test Supervisors, Backup Test Supervisors, Test Accommodations Coordinators, Room Supervisors, and Proctors play a key role in the fair and equitable administration of the MME Assessment.

Each public school and participating non-public school must designate the following testing staff: Test Supervisor, Backup Test Supervisor, and Test Accommodations Coordinator who meet the operational eligibility criteria to administer the MME. Each school building (or alternate testing facility) involved in administering the assessments must meet the established facility standards. The following manuals used during workshop training and provided to test day staff detail the procedures for administering the MME assessment for Spring 2009.

- *Spring 2009 – Supervisor’s Manual ACT Plus Writing – State Testing*
- *Spring 2009 – Supervisor’s Manual ACT Plus Writing – State Special Testing*
- *WorkKeys –Administration Manual for State Testing*
- *MME Administration Manual Spring 2009*
- *MME Administration Manual for Students Testing with Accommodations*

The MME Spring 2009 assessments were designed to be administered by eligible / trained school staff. School staff eligible to administer the MME Assessments must meet the following criteria:

- Test (and Backup) Supervisors – may NOT be related to any examinee taking the MME in 2008-2009 anywhere in Michigan.
- Room Supervisors and Proctors may NOT assist in a room where any relative is being tested.
- Test Accommodations Coordinators – may NOT be related to or guardian of any examinee participating in MME accommodations testing anywhere in Michigan during the testing year.

- Testing staff supporting accommodations – may not be involved in coaching high school athletics or college athletics (applicable only if student testing with accommodations participates in athletics).

Relatives include: children, stepchildren, grandchildren, nieces, nephews, siblings, in-laws, spouses and wards.

Depending on the number of students in each room, trained room supervisors and proctors were assigned to assist Test Supervisors or Test Accommodations Coordinators. The following staffing guidelines were required: A proctor may be used to assist a room supervisor or the Test Supervisor if fewer than 25 examinees are testing. A Proctor is required (in addition to the Room Supervisor) for every 25 examinees (or portion thereof) after the first 25 in a room. (For students testing with accommodations, the ratio is one Proctor to every 10 examinees, or portion thereof, after the first 10 in a room).

Roles and responsibilities of the Test Supervisor, Backup Test Supervisor, Test Accommodations Coordinator, Room Supervisor, and Proctor are specified in the *Spring 2009– Supervisor’s Manual ACT Plus Writing – State Testing*, *Spring 2009– Supervisor’s Manual ACT Plus Writing – State Special Testing*, *WorkKeys Administration Manual for State Testing*, and *MME Administration Manual Spring 2009 and the MME Administration Manual for Students Testing with Accommodations Spring 2009*.

Michigan has made the commitment that all public school students must be assessed as required by state policy and federal law and provided the opportunity for non-public schools and students to optionally participate. During the Spring 2009 administration all 11th graders were given the opportunity to take all of the MME assessment components, with the exception of students with IEPs that indicate that they should take MI-Access, Michigan’s alternate assessment.

For the spring of 2009 every 11th grade and eligible 12th grade student as defined by the local school district based on academic standing, was provided the opportunity to qualify for the Michigan Promise Scholarship.

Preparation for Test Administration

Due to the fact that the MME is a standardized assessment and that this assessment must be administered under identical conditions in all schools, there are a set of standardized procedure that must be followed for all components of the test.

School Establishment Process

Because the ACT Plus Writing college entrance examination is one component of the MME, and is administered on Day 1 as a national standardized assessment that can result in college-reportable scores, there are certain ACT rules that must be followed as part of administering the entire MME. Each Michigan high school, with its own Michigan School Code, MUST be established as an MME Test Center. Students enrolled at these schools cannot test at another school. In the fall of 2008, ACT sent out School Establishment Packets or Renewal Packets to Michigan high schools, to guide them through the approval process. As explained in those packets, in order for a school to be approved to administer the MME, they must (1) submit all required forms, and (2) complete required staff training.

Below is a description of the responsibilities of three key testing staff: the Test Supervisor, the Test Accommodations Coordinator, and the Back-up Test Supervisor. ACT requires that these individuals be identified and registered with ACT as part of the School Establishment process undertaken in the Fall of 2008 with all schools that will serve as Test Centers.

The Test Supervisor, Back-up Test Supervisor, and Test Accommodations Coordinator must assume important professional responsibilities to protect the integrity of all secure test materials and to ensure that all examinees at their school are tested under the same conditions as examinees at every other school administering the examination.

Qualifications and Requirements for Test Supervisors and Back-up Supervisors include:

1. **Not be related to or guardian of any examinee participating in State Testing with standard time anywhere in Michigan on either the initial or makeup test date this year. (Relatives or wards include children, stepchildren, grandchildren, nieces, nephews, siblings, in-laws, spouses, and persons under their guardianship.)**
2. Be proficient in English.
3. Be experienced in testing and measurement.
4. Be a staff member of the school.
5. Have control over locked, limited-access storage at the school to secure the test materials.
6. Ensure that the tests are administered in strict compliance with all policies and procedures as documented in each Supervisor's Manual (one for each day of testing).
7. Not be engaged in test preparation activities for the ACT at any time during the current testing year (September through August), except as specifically required by school contract. The normal duties of a counselor or teacher are **not** a conflict of interest, provided they are part of job responsibilities specifically defined by one's employer and the employer is not a commercial enterprise.

Primary Responsibilities

1. Newly appointed Test Supervisors and Back-up Test Supervisors must participate in a mandatory training session conducted by ACT and Michigan Department of Education staff.
2. Read and follow exactly all policies and procedures in each Supervisor's Manual (one for each day).
3. Arrange for all students to complete pre-test sections of their answer folders in a supervised session at school **before** test day. If applicable, affix barcode labels to examinee answer folders prior to test day.
4. Arrange for all students to test on the designated test dates with testing as the first activity of the morning. All room supervisors must begin reading the Verbal Instructions **no later than 9:00 a.m.** (for standard time testing).
5. Make arrangements for test rooms that meet standard testing requirements, including uncrowded seating facing the same direction, manageable security, good lighting and ventilation, adequate writing surfaces, and required space between examinees.
6. Ensure test rooms are free from distractions during the test session(s) (bells, public address system turned off, etc.) and separated from regular school activities.
7. Receive, check-in, and ensure security of test materials from receipt until return. Take steps to protect materials from damage, theft, or loss, and from conditions that could allow prior access to the tests.
8. Identify a sufficient number of qualified assistants to serve as room supervisors and proctors. One room supervisor is required per room, plus one proctor for every 25 examinees in the room after the first 25 (or one proctor for every 10 students testing with accommodations). All testing staff must be proficient in English, may **not** be involved in ACT test preparation outside of normal school duties, and may not be enrolled in high school. No room supervisor or proctor may assist in a room where a relative is testing.
9. Conduct training for all testing staff before the test dates, including a complete review of each Supervisor's Manual (one for each day).
10. Ensure all testing staff remain attentive to testing responsibilities throughout the entire administration, including accurate timing and monitoring for prohibited behavior.

11. Complete, verify, and return all required reports, seating diagrams, forms, answer folders, and test booklets immediately after testing.
12. Document all irregularities and consult directly with ACT, OEAA, and Measurement, Inc., as appropriate, regarding actions to be taken.
13. Cooperate fully with ACT, OEAA, and Measurement, Inc., if applicable, to investigate and resolve suspected or documented irregularities.

Qualifications and Responsibilities for Test Accommodations Coordinators:

1. **Not be related to or guardian of any examinee participating in State Testing with accommodations anywhere in Michigan this year during the two week testing window for accommodations. (Relatives or wards include children, stepchildren, grandchildren, nieces, nephews, siblings, in-laws, spouses, and persons under their guardianship.)**
2. Be proficient in English.
3. Be experienced in testing and measurement.
4. Be a staff member of the school.
5. Have control over locked, limited-access storage at the school to secure test materials.
6. Ensure that the tests are administered in strict compliance with all policies and procedures as documented in each Supervisor's Manual (one for each day of testing).

To avoid the appearance of a conflict of interest and to protect both the examinee and testing staff from allegations of impropriety, the Test Accommodations Coordinator must also:

1. Not be a private consultant or individual tutor whose fees are paid by a student (or the student's family) for whom accommodations are requested.
2. Not be engaged in test preparation activities for the ACT at any time during the current testing year (September through August), except as specifically required by school contract. The normal duties of a counselor or teacher are **not** a conflict of interest, provided they are part of job responsibilities specifically defined by one's employer and the employer is not a commercial enterprise.
3. Not be involved in coaching high school or college athletics (applicable only if any student requesting accommodations participates in athletics). This qualification is in place to protect testing staff who receive and handle secure test materials and who administer the test to students individually or in very small groups without other testing staff present.

Primary Responsibilities

1. Determine which students need to apply for accommodations on the ACT, complete a request form for each, gather required signatures, and compile documentation. Consult with appropriate school personnel to determine accommodations for Day 2 and Day 3 materials to be ordered on the OEAA Secure Site.
2. Ship completed accommodations request forms and the completed Test Accommodations Coordinator Header as a group to arrive at ACT no later than the required deadline provided to you on the Checklist of Dates.
3. Provide timely responses to requests from ACT for additional information about individual students.
4. Newly appointed Test Accommodations Coordinators must participate in a mandatory training session conducted by ACT and Michigan Department of Education staff.
5. Train staff assigned to assist with the administration of tests to students approved for accommodations.
6. Check-in all secure test materials shipped for students testing with accommodations and, in consultation with Test Supervisor, maintain security while materials are at the school.

7. Arrange for all students to complete pre-test sections of their answer folders in a supervised session at school **before** test day. If applicable, affix barcode labels to examinees answer folders prior to test day.
8. Arrange for all students to test within the designated accommodations testing window using only the authorized accommodations and materials assigned to each student.
9. Assign examinees to test rooms, separated by timing code with a room supervisor for each room. Separate students testing with different timing codes according to instruction provided in the *Supervisor's Manual for Students Testing with Accommodations*.
10. Complete, verify, and return all required reports, seating diagrams, forms, answer folders, and test booklets/alternate formats as directed immediately after the testing window.
11. Document all irregularities and consult directly with ACT, OEAA, and Measurement, Inc., as appropriate, regarding actions to be taken.
12. Cooperate fully with ACT, OEAA, and Measurement, Inc., if applicable, to investigate and resolve suspected or documented irregularities.

MME Day 1 and Day 2: Materials Processing

Materials Orders—Day 1 and Day 2

ACT utilized enrollment numbers provided by each school to develop preliminary standard materials quantities for Day 1 and Day 2 of the MME. Materials quantities for accommodated students were produced using numbers gathered through the accommodations application and request process for Day 1 and through the OEAA Secure Site for Day 2.

Shipping—Day 1 and Day 2

To provide the OEAA with secure and dependable services for the shipping of Michigan assessment materials, ACT's Distribution Center maintains the quality and security of material distribution and return by using such methods as sealed trailers and hiring reputable carriers with the ability to immediately trace shipments. ACT uses all available tracking capabilities to provide status information and early opportunities for corrective action.

Materials are packaged by school and delivered to the Test Supervisors. Each shipment to a school contains a shipping document listing each school's materials.

Final standard time Day 1 and Day 2 materials quantities are packaged using information provided by the Test Supervisors through OEAA's secure website. Michigan educators also provide ACT with the Pre-Identification information needed to print barcode labels which are affixed to each answer folder. Bar-coding of all secure materials during the pre-packaging effort allows the accurate tracking of these materials through the entire packing, delivery, and return process. It also permits us to inventory all materials throughout the packaging and delivery process along with the ability to provide the customer with status updates at any time.

For the Spring 2009 testing, secure and non secure materials shipped were reported in Tables 4.1. through 4.7.

Day 1 Processing

Table 4.1. Total Secure Documents Shipped in Spring 2009 Administration

Test Booklets	Quantities
ACT Multiple Choice-Initial	166,120
ACT Multiple Choice-Makeup	10,465
ACT Writing-Initial	171,150
ACT Writing-Makeup	14,680
Total Accommodations	10,491
Total	372,906

Table 4.2. Total Non Secure Documents Shipped in Spring 2009 Administration

Answer Folders	Quantities
ACT Plus Writing-Initial	170,400
ACT Plus Writing-Makeup	17,016
Total	187,416

Day 2 Processing

Table 4.3. Total Secure Documents Shipped in Spring 2009 Administration

Test Booklets	Quantities
<i>WorkKeys</i> Initial	162,534
<i>WorkKeys</i> Makeup	11,512
<i>WorkKeys</i> Accommodations	12,789
Total	186,835

Table 4.4. Total Non Secure Documents Shipped in Spring 2009 Administration

Answer Folders	Quantities
<i>WorkKeys</i> Initial	169,744
<i>WorkKeys</i> Makeup	11,455
Total	181,199

Receipt and Processing—Day 1 and Day 2

Each school's shipment included a copy of the packing list along with other shipping information to permit the accurate inventory of materials upon receipt by the Test Supervisor or the Test Accommodations Coordinator. Day 1 and Day 2 materials were shipped via a secure carrier, with traceable means, to pre-specified shipping addresses provided by the Test Supervisors and Test Accommodations Coordinators from each school. ACT standard shipping process does not allow for shipment to districts. These shipments were comprised of non-secure shipments followed by shipments of secure test materials. The nonsecure

shipments included administration manuals, pre-printed barcode labels, and the answer documents students needed prior to the Day 1 and Day 2 assessments to complete the noncognitive sections of the ACT and *WorkKeys* in a supervised in-school pre-test session.

ACT requested each school’s Test Supervisor or Test Accommodations Coordinator inventory the materials sent, verify the shipping contents and call ACT’s toll-free number to report any shipping problems or materials shortages. Instructions were provided for secure storage of materials until test day.

Table 4.5. Number of Students Tested in Spring 2009 (ACT / *WorkKeys* / MME)

	MME Testing – Spring 2009
Day 1 - ACT	126,394
Day 2 - <i>WorkKeys</i>	125,575
Day 3 - MME	127,127

Test Security—Days 1 and 2

Secure test materials include all ACT test booklets and used answer folders. The Test Supervisor is responsible for the security of all test materials from the time the carrier delivers them to the school to the time they are in the return carrier’s possession. The Test Supervisor must protect the materials from damage, theft, or loss, and from conditions that could allow prior access to the tests.

Test materials must be kept in a locked, secure area, such as a vault or non-portable cabinet in a locked, limited-access room. Only the Test Supervisor, Back-up Test Supervisor, Test Accommodations Coordinator, and possibly a few specifically authorized persons may have access to the area. If the security of test materials is compromised, ACT will not report scores.

ACT test booklets are copyrighted and cannot be photocopied or used for any purpose other than testing. Under no circumstances is a test booklet seal to be broken by anyone other than the examinee as instructed on test day. Testing staff and examinees are prohibited from disclosing test questions, essay topics, or response choices to anyone.

Directions in the manuals note that testing staff who observe a student engaging in one or more of the unethical practices should allow the suspected student(s) to finish the assessment and mark the student’s answer folder VOID and complete an Irregularity Report. The Assessment Administrator is instructed to immediately notify the Test Supervisor of the suspected prohibited practice. Adequate Yearly Progress (AYP) requires the use of a valid assessment score. A student without a valid assessment score is considered “not assessed” for AYP purposes.

Materials Return—Day 1 and Day 2

Schools were provided with “Return Kits” containing all of the necessary labels and documentation for returning their materials.

The tracking numbers of the FedEx return labels provided to each school were documented at the time of “Return Kit” production and those numbers were entered into our internal tracking system database.

Materials were prepared for return by the Test Supervisor. They packaged the materials and applied the self-adhesive return label that was supplied in the “Return Kit” from their original shipment. On the day after the initial test day, FedEx was dispatched to each school that had been sent Day 3 materials to retrieve test materials. This process was repeated for each school on the day after make-up testing.

For accommodated materials, all materials must be returned after the close of the accommodated testing window. Test Accommodations Coordinators are provided with specific return instructions similar to those provided to Test Supervisors for non-accommodated materials.

Test Supervisor manuals provide clear instructions on how to assemble, box, and return testing materials after test administration. Because of the criticality of used test materials and quantities often involved, safety is also a major concern, not only for the materials but for the people moving them. Only single column boxes are used to distribute and collect test materials, so the weight of each carton is kept to a reasonable and manageable limit.

Preaddressed, prepaid labels are provided. The labels facilitate accurate and efficient sorting of each carton and its contents upon receipt.

Day 1 and Day 2 materials were returned directly to ACT.

Materials Discrepancy Process—Day 1 and Day 2

ACT logged in the returned assessment materials from Day 1 and Day 2 of the MME during the check-in process. A check-in database was created for 2008-2009 to facilitate this process. The database tracked Day 1 and Day 2 standard time and accommodated formats. ACT followed up with schools to assure timely return of those testing materials, as well as tracked schools who did not return all testing materials.

Schools that have not returned any material

Detailed status reports are generated as test materials are received and checked in. These reports are monitored daily for missing or incomplete shipments and follow-up occurs with schools missing materials.

Schools that have returned incomplete shipments

Detailed status reports, listing the number of boxes received from each school, are reviewed daily. An ACT team member will follow up with a phone call on quantities appearing to be less than expected as compared with FedEx tracking information.

Schools with missing secure test materials

After secure materials are scanned, reports indicating missing materials are generated. These reports identify materials and serial numbers and are provided to the ACT team for follow-up with the affected schools.

Schools Returning Answer Documents After Established 4/3/09 Cutoff Date

Documents were processed in accordance with late receipts processing guidelines mutually agreed upon by ACT and OEAA.

Processing Assessment Materials Returned by Schools—Day 1 and Day 2

ACT logged in the returned assessment materials from Day 1 and Day 2 of the MME during the check-in process within 24 hours of receipt and the answer documents were prepared for scanning within 72 hours of receipt. ACT followed up with schools to assure timely return of all testing materials, as well as to track schools that had not returned all Day 1 and Day 2 testing materials. The status of each school was readily

discernable from the log files updated by check-in staff. ACT utilized standard processing procedures for the Day 1 and Day 2 assessments, in terms of transferring the documents from the check-in process to the scanning process.

MME Day 3 Michigan Components: Materials Processing

Materials Orders—Day 3

Schools ordered all Day 3 materials through the OEAA Secure Site. For the initial order of materials, schools identified the number of students taking the standard-time assessment and the number of students taking each accommodated format of the test.

For established schools which had not placed an initial order, OEAA used the greater of either the number of grade 11 students enrolled in the September MSDS (formerly SRSD) file or the pre-ID count to place an order. For grade 12, only the pre-ID student count and not the MSDS (formerly SRSD) file enrollment count was used.

Appropriate quantities of materials for each initial order were packed and shipped to each school in two shipments, one containing non-secure materials and one containing secure materials. The non-secure shipment included pre-ID student barcode labels, administration manuals (1 per 15 standard-time students), administration manuals for testing students with accommodations, school header sheets, class/group ID sheets, and materials necessary for packaging and return of test materials. The secure shipment included test books and accommodated formats.

Schools placed additional orders for specific quantities of certain items, rather than count of students testing. This was also true for orders for makeup materials.

For all orders, Measurement Inc. combined the pull of data and the processing of that data into one step. This provided immediate feedback to the OEAA Secure Site with respect to any orders that could not be filled immediately, and eliminated duplicates of process data showing up in the system.

Measurement Inc. warehouse staff utilized a packaging application to generate an on-demand pick list. The pick list documented the specific materials and quantities to be included in an order, but not the exact barcode ranges for secure materials. After warehouse staff picked the materials for the order, the materials and pick list were delivered to a packing station. The staff member at the packing station initiated the packing process by scanning the order number on the pick list. The packing station employee used a hand scanner to capture the barcode value on each secure material. As each material was scanned into the order, the packing application verified that it was the correct material and kept a running count of the quantity. After the packing station employee entered all order material information into the system, a validation check verified they had packed the proper quantity of the correct materials.

Any validation failures were displayed on the packing station screen. The packing station employee corrected any errors by scanning more materials into the order or removing materials by scanning the barcode of the material that needed to be removed. Once all validation failures had been corrected and the material types and quantities matched the order information, the packaging application printed a packing list and secure checklist to be included in the shipment.

Each packing station included a shipper tracking label printer to maintain order accuracy. By doing so, each order remained independent of other orders during packaging and sealing. The tracking label contained the

order number and address of the recipient for verification against the packing list. After verifying the shipper tracking label against the packing list, each box was sealed with heavy-duty plastic tape and the shipper tracking label applied. The FedEx label contained the order number, school number and the name of the school to which the materials were being delivered, and a Box n of X identifier to indicate the number of boxes shipped. At the time the FedEx number was created the application created a corresponding entry in the FedExTracking table in the MI database.

Shipping—Day 3

For MME Day 3, Measurement Inc. monitored the distribution of materials to schools by FedEx. Test Supervisors were instructed to inventory all test materials sent in order to assure that they received an adequate supply of assessment materials. Test Supervisors were instructed to use Measurement Inc.’s Call Center via a toll-free telephone number to report any problems. Additional orders for materials were placed via the OEAA Secure Site.

Processing—Day 3

Table 4.6. Total Secure Documents Shipped in Spring 2009 Administration

Test Booklets	Quantities
MME Initial	147,111
MME Makeup	9,710
MME Accommodations	13,256
Total	170,077

Table 4.7. Total Non Secure Documents Shipped in Spring 2009 Administration

Answer Folders	Quantities
MME Initial	128,991
MME Makeup	5,128
Total	134,119

Materials Receipt and Processing—Day 3

Upon arrival at Measurement Inc., all boxes were scanned into their tracking system database where they were logged in and checked against the tracking numbers that were pre-assigned to each school. This provided immediate information on the number of boxes received and their points of origin. Boxes marked with a “scorable” label were separated from boxes marked with a “non-scorable” label. Boxes without either label were processed as “scorable”.

Scorable Materials—Day 3

Boxes of materials marked as “scorable” were opened sequentially to remove used answer documents. Answer documents, along with school header and any class/group ID sheets, were then placed into bar-coded tote boxes for IT Operations. (If there was no school header, a cover sheet for the correct school was generated.)

As materials were transferred to IT Operations for scanning, the tracking number from the shipping box was scanned along with the barcode for each tote box into which materials from that shipping box are placed. This procedure provided a permanent link between the shipping box in which materials were received and the tote box containing the answer documents from that box. A scan batch ID sheet was placed on top of the tote box when full, and was linked to the scan batch ID label on the tote box.

Non-Scorable Materials—Day 3

Boxes containing non-scorable materials were examined to remove any scorable materials that may have been returned there in error. A separate or “redundancy” check was performed on each box by a second individual at this time to assure that all scorable materials were located. Any scorable materials located during these searches were placed immediately into the appropriate tote boxes according to the procedure outlined for other scorable materials. The tote boxes of used answer documents were then forwarded to our IT department for scanning and processing.

The security check-in process for the secure materials from boxes marked “non scorable” captured the security barcode number for each booklet or accommodated format material returned. These items were unpacked and then scanned at a workstation equipped with a barcode reader and a PC. The barcode of the box into which the booklets was to be stored was linked to each set of scanned secure items.

A report was produced listing any security barcodes in the master database, but not found during check-in.

Overall, 99.82% of secure materials sent were returned and checked in. Of 188, 940 secure materials sent, 188, 605 were checked in, leaving 335 items “missing” for further investigation at the school level by OEAA and Measurement Inc.

Note: if any boxes, scorable or non-scorable, are found to contain MME Day 1 (ACT) or Day 2 (*WorkKeys*) materials, the worker will first ensure that any Day 3 materials are removed from the box. The box will then be sent to ACT, following procedures in the “Process for Handling Misdirected MME Materials” document.

Test Security—Day 3

Procedures related to test security are the same on Day 3 as on Days 1 and 2. Test materials must be kept in a locked, secure area, such as a vault or non-portable cabinet in a locked, limited-access room. Only the Test Supervisor, Back-up Test Supervisor, Test Accommodations Coordinator, and possibly a few specifically authorized persons may have access to the area. Test booklets cannot be photocopied or used for any purpose other than testing. Under no circumstances is a test booklet seal to be broken by anyone other than the examinee as instructed on test day. Testing staff and examinees are prohibited from disclosing test questions, essay topics, or response choices to anyone.

One difference between Days 1 and 2, and Day 3 is the procedure should a Room Supervisor observe a student engaging in unethical practices. In Day 1 testing, prohibited behavior results in a voided Answer Document. For prohibited behavior on Day 3, Room Supervisors mark the “Prohibited Behavior” circle on the Answer Document and return the Answer Document. Measurement Incorporated with scan, but not score, the Answer Documents.

Scanning/Scoring—Day 3

Once they were logged into the Operations department, scan bins were forwarded to the cutting area, where one scan bin at a time was removed from the cart for cutting. The cutting operation converted the multi-page answer document into a stack of single sheets ready for scanning. When the answer documents were printed, each sheet was imprinted with a lithocode value unique to that document. Both a scannable and human-readable version of the lithocode were printed on every sheet of each answer document. In the unlikely event that a scan bin was dropped at the cutting or pre-scanning stage, the unique lithocode allowed the answer documents to be reassembled, and answer document integrity to be verified at the scanner and project database once the data was transferred. Software validations at the scanner ensured that all pages of each student’s answer document were accounted for; thus, any pages that are out of order could be easily corrected prior to any further processing.

MI image scanned all pages of a student’s answer document at the same time using BancTec IntelliScan XDS color image scanners. The BancTec IntelliScan XDS is rated to scan 190 sheets per minute at an optical resolution of 240 dots per inch (DPI) and creates both JPEG and TIFF images for every page. These scanners utilize precision camera assemblies pressurized to minimize dust. This, plus low maintenance LED camera illumination, reduces the need for rescans. The scanner features a completely open paper path to dramatically improve document throughput. This paper path reduces the time to recover from paper jams and other complications that are common for scanners with more restrictive paper paths. Both sonic and vacuum double-sheet detection technology ensure that every sheet is scanned. In addition, BancTec has designed custom document integrity software for Measurement Incorporated. This application detects out-of-sequence pages. The scanner will stop to allow operator correction before imaging, thus eliminating post scanning corrective action.

To ensure that all sheets in the scan bin are scanned, the last sheet in every bin was an “End of Batch” sheet. If the End of Batch record does not appear in the data file an error alert was generated, and a technician made a visual check of the scan bin to verify all answer documents had been scanned. If necessary, the data file was opened again, and any missing sheet(s) appended to the file creating a complete data file.

Data Correction

Once all of the scanned data was combined to create the student records, data validation routines were executed. These routines analyzed the data and created error tables for answer documents containing questionable data. Common error detection routines included checks for the following situations:

- Inconsistencies in school, grade, or form
- Inconsistencies in headers and answer documents
- Duplicate student barcodes within the same bin or another bin of answer documents
- Missing student barcodes
- Missing or incomplete demographics (such as a blank name)
- Double marks in the demographic and/or multiple-choice grids

Measurement Incorporated utilized a double data correction process. Data correction operators used our data correction application that retrieves flagged data records and highlights the problem field on a computer screen so it can be resolved. The operator compared the highlighted data to the scanned image of the answer document, and made any necessary correction. Once an operator corrected a flagged record, the same flagged record was routed to a second data correction operator who repeated the data correction process.

After a flagged record was edited by two operators, the data correction application checked that both operators have made identical corrections. In the event that the two corrections differed, the record was routed to a supervisory staff member for a third and final resolution. This process continued until all flagged records are examined.

To ensure accuracy, once a correction had been written to the database, the document was validated again to ensure the corrected edit had not created another error. All edits were recorded and tracked in Measurement Inc. databases, along with the user ID of the staff member making the edits.

Multiple-choice Scoring

After all flagged data is reviewed and corrected, student selected responses are scored against the item answer keys. The Test Maps table called TestMaps pulled from the OEAA database in Michigan. That data was converted into a set of 36 records, each with its own set of correct answers (or answer key strings). Then those answer keys were applied to student responses to produce a string of ones and zeroes, indicating right and wrong answers. A validation process (key check) was used to detect any potential answer key problems. The students' selected responses and correct answer indicators were transmitted to MDE in a data file.

Score Reporting

The master student roster that identified all students for whom reports should be produced was a student data "Match File" transmitted to Measurement Incorporated from OEAA. This matched file included information from Days 1, 2 and 3 of testing, and had been analyzed by ACT to produce scale scores, item statistics, and other psychometric analyses. It was then returned to OEAA and to Measurement Incorporated for the production of score reports.

The reports included: Individual Student Reports, Parent Reports, Student Roster, Student Record Labels, ISD Comprehensive Report and District Comprehensive Report, State Demographic Report, ISD Demographic Report, District Demographic Report, School Demographic Report, State Summary Report, District Summary Report, and School Summary Report.

Measurement Inc. provided each of the reports as a static or dynamic Adobe Acrobat PDF on the Electronic Report Hosting website. These PDF files were electronically transmitted to the Michigan MME. The PDF files were split into batches based on the report type, and the PDF files of each batch were placed in their own sub-directories. HOV produced PDFs separated by school. In addition to the electronic distribution of report, the PDF of reports was printed and distributed to the schools.

For schools in districts that selected the green option for reporting, the Individual Student Reports, Parent Reports, and Student Record Labels were printed. All other reports were available only online as PDFs.

The PDFs were extensively reviewed before the preliminary reports were printed and mailed. Labels and district reports were printed inline and sorted with the related school reports. Labels were printed on inventory label material. The reports were segmented by color card stock.

Reports packages were shrink-wrapped and packaged for traceable ground delivery throughout the state. Depending on the size of the report, the reports were packaged in appropriate shipping boxes or envelopes. Packages were matched against a distribution list for accuracy and completeness of the cycle run.

The MME Guide to Reports provided samples of the various reports, along with descriptions of how users could better understand and use those reports.

HOV printed the Guide-to-Reports Handbooks for Michigan MME. These handbooks were printed separately and included with each report package. In addition to the printed handbook, the information was also posted on the MME Web page, found at http://www.michigan.gov/documents/mde/9_MME_2009_Guide_to_Reports_283213_7_283862_7.pdf.

Description of Score Reports

Parent Report

The Parent Report presented individual test results for all students in grades 11 and 12 who tested in a subject.

The Parent Report contained the following information:

- Scale score for each subject area
- Performance level for each subject area
- Subscore values for each subscore strand for each subject area, including the number of points the student earned, the number of points possible, and the percent correct
- Text, including a letter from the superintendent, performance level definitions, subject descriptions, assessment descriptions, and ACT and *WorkKeys* descriptions
- Scale score graphs
- ACT test scores
- Work Skills level scores

The Parent Report provided information for the following subjects: MME Mathematics, MME Science, MME Reading, MME Writing, MME Total ELA, and MME Social Studies.

Individual Student Report

The Individual Student Report (ISR) provided a detailed description of each student's performance in the subject areas assessed on the MME. This report was designed to help educators identify the academic strengths of their students and the areas that may need improvement. Schools may include these reports in student record files.

The Individual Student Report was a report on a student's achievement in four subject areas: English Language Arts, Mathematics, Science, and Social Studies. English Language Arts is divided into Reading and Writing Sections.

Student Roster

The Student Roster presented individual test results for all students in grades 11 and 12 who tested in a subject. It listed those students by class/group who took the test in the subject - regardless of what form they took.

The Student Roster contained the following information for each subject area:

- Scale Score
- Performance Level
- Subscore values for each subscore strand, including the number of points the student earned and the number of points possible

The last line of each subject of the report showed the number of students assessed, which was the number of students reported on the roster for that group.

The Student Roster provided information for the following subjects: MME Reading, MME Writing, MME Total ELA, MME Mathematics, MME Science, and MME Social Studies.

Student Record Labels

The Student Label provided summary description of each student's performance in the subject areas assessed on the MME.

The Student Labels consisted of the following information for each student:

- Demographic information
- Scale score and performance level for subjects tested

Student Labels provided student information in different subjects in the following order:

- ELA Total
- Reading
- Writing
- Mathematics
- Science
- Social Studies

State Demographic Report, ISD Demographic Report, District Demographic Report, School Demographic Report

The Demographic Report was a statistical summary of twenty student demographic areas for all the subjects in a grade, aggregated in a student group. There were eighteen types of student groups, arrived at by combining the following modes, populations, and grades:

Modes:

- State
- District
- School

Student populations:

- All Students
- Students with Disabilities
- All Except Students with Disabilities

Grades:

- 11
- 12

The Demographic Report provided data for the following subjects: MME Reading, MME Writing, MME Total ELA, MME Mathematics, MME Science, and MME Social Studies.

In calculating the percent of students with scale scores at a certain performance level, both the numerator and denominator were expressed as a float, and the result of that calculation was rounded in the manner of the SQL function ROUND (numeric_expression, length).

State Summary Report, District Summary Report, School Summary Report

The Summary Report consisted of two pages:

- 1 - A summary of performance levels achieved compared to previous years
- 2 - A distribution of scores by subject and strand in a grade

Both pages of reports were produced for all eighteen types of student groups reported, which were arrived at by combining the following modes, populations, and grades:

Modes:

- State
- District
- School

Student populations:

- All Students
- Students with Disabilities
- All Except Students with Disabilities

Grades:

- 11
- 12

The Summary Reports provided data for the following subjects: MME Reading, MME Writing, MME Total ELA, MME Mathematics, MME Science, and MME Social Studies.

Any mean score in this report was the average score calculated by summing the applicable scores and dividing that sum by the total number of those scores. Percentages were calculated by dividing the number in a category by the total number of students assessed. In any division calculation, both the numerator and denominator were expressed as a float, and the result of that calculation was rounded in the manner of the SQL function ROUND (numeric_expression, length).

ISD Comprehensive Report and District Comprehensive Report

The Comprehensive Report provided summary score data by subject and grade for public schools, aggregated in a student group. The District Comprehensive Report listed data for the district, followed by data for each school within the district. The ISD Comprehensive Report listed data for the ISD, followed by data for each district.

There were twelve types of student groups, arrived at by combining the following modes, populations, and grades:

Modes:

- ISD
- District

Student populations:

- All Students
- Students with Disabilities
- All Except Students with Disabilities

Grades:

- 11
- 12

The Comprehensive Report provided data for the following subjects: MME Reading, MME Writing, MME Total ELA, MME Mathematics, MME Science, and MME Social Studies.

In calculating the mean scale score or the percent of students with scale scores at a certain performance level, both the numerator and denominator were expressed as a float, and the result of that calculation was rounded in the manner of the SQL function ROUND (numeric_expression, length).

Accommodations for Students with Disabilities (SWD) and English Language Learners (ELL)

All students are to participate in the assessment programs approved by the State Board of Education. For some students, accommodations that are customarily used during routine classroom activities may be considered for use during the administration of the MME assessments. The State Board of Education has approved standard and nonstandard assessment accommodations for the Michigan Educational Assessment System including MME, MI-Access, and ELPA.

The MME Accommodation Summary Table (Table 4.11 beginning next page) identifies standard and nonstandard accommodations for students with disabilities, Section 504 students, and/or students with limited English proficiency (also referred to as English language learners, or ELL). Standard accommodations do *not* change the construct that the assessment is measuring and *do* provide a valid score. Nonstandard accommodations change the construct that the assessment is measuring, rendering scores that are not valid. Accommodations not listed in the table are considered nonstandard.

The Michigan Merit Examination (MME) consists of three major components administered over three days: the ACT Plus Writing, three *WorkKeys* tests (*Reading for Information*, *Applied Mathematics*, and *Locating Information*), and Michigan developed items for mathematics, science and social studies. Table 4.8 outlines the Spring 2009 test organization.

Table 4.9 outlines which components contribute to each MME score. The MME scores will play a role in qualifying for the Michigan Promise scholarship and will be the foundation for the No Child Left Behind (NCLB) calculation of Adequate Yearly Progress (AYP) and EdYES! accountability reports for high schools.

Table 4.8. Spring 2009 Test Organization

Spring 2009 Test Organization						
Day*	Assessment	Subject Session	Number of Parts	Total Items	Testing Time (minutes)	Estimated Time Required for Administration
Day 1 March 10 (Makeup March 24)	ACT Plus Writing	English	5	75 MC items	45	Total test time including check in, instructions breaks, and collection of materials - 5 hours
		Mathematics		60 MC items	60	
		Reading		40 MC items	35	
		Science		40 MC items	35	
		Writing		1 Prompt	30	
Day 1 Standard Testing Time 205 minutes (3 hrs / 25 minutes)						
Day 2 March 11 (Makeup March 25)	WorkKeys	<i>Reading for Information</i>	3	33 MC Items	45	Total test time including check in, instructions breaks, and collection of materials – 3.5 hours
		<i>Applied Mathematics</i>		33 MC Items	45	
		<i>Locating Information</i>		38 MC items	45	
Day 2 Standard Testing Time 135 minutes (2 hour / 15 minutes)						
Day 3 March 12 (Makeup March 26)	Michigan	Mathematics	3	34 MC items	40	Total test time including check in, instruction, breaks and collection of materials - 3 hours
		Science		49 MC items	40	
		Social Studies		42 MC items	40	
Day 3 Standard Testing Time 120 minutes (2hours)						
<i>*More detailed information about this schedule and the MME program is available on the MME Web page at www.mi.gov/mme.</i>				TOTAL Minutes	460	
				TOTAL hours	7.67	

Table 4.9. Components Contributing to MME Score

				Components Contributing to MME Scores					
Day	Test	Subject Session	Parts	ELA	Reading	Writing	Mathematics	Science	Social Studies
Day 1	ACT Plus Writing	English	1	Selected items		Selected items			
		Mathematics	1				Selected items		
		Reading	1	Selected items	Selected items				
		Science	1					Selected items	
		Writing	1	X		X			
Day 2	WorkKeys	<i>Reading for Information</i>	1	Selected items	Selected items				
		<i>Applied Mathematics</i>	1				Selected items		
		<i>Locating Information</i>	1				Selected items		Selected items
Day 3	Michigan	Mathematics	1				X		
		Science	1					X	
		Social Studies	1						X

MME Test Accommodations Window

All accommodated testing must be administered within the two-week window that begins on the initial test date for that component of the MME and ends on the makeup date for that component. Testing may be scheduled on any days during the window, but each student must take the tests in prescribed order – Day 1 (the ACT Plus Writing), followed by the Day 2 tests, followed by the Day 3 tests. All testing staff must meet ACT’s requirements. If testing occurs outside the authorized window, or with procedures that conflict with ACT directions, or under supervision of testing staff who do not meet ACT’s requirements, the answer documents will not be scored or scores cancelled.

ACT-Approved versus State-Allowed Accommodations on the ACT (Day 1 of the MME)

ACT is committed to ensuring that official ACT scores reported to colleges and other entities from MME testing are comparable to scores earned through other forms of ACT testing involving the application of ACT’s test accommodations policies. Therefore, ACT supports the following two forms of accommodations on the ACT when it is administered as Day 1 of the MME:

- 1) **ACT-approved accommodations** result in ACT scores that are fully reportable to colleges, scholarships, and other entities *in addition to* being used for MME scores. Only students with professionally diagnosed and documented disabilities who receive accommodations in school should apply for ACT-approved accommodations.
- 2) **“State-allowed” accommodations** result in ACT scores that are not college reportable; they are used only for MME scores. English language learners who do not have a disability but receive accommodations in school should request State-Allowed accommodations.

Requesting Accommodations on the ACT (Day 1 of the MME)

In general, all accommodations on the ACT must be requested and reviewed by ACT. However, there are limited exceptions. For example, because testing will normally occur at the local school rather than a separate test center, some arrangements do **not** require review or prior approval from ACT (e.g., placement at the front of the room). Such arrangements are noted on the attached accommodations summary table as “local decision” meaning they do **not** require ACT review or approval.

All schools must appoint a Test Accommodations Coordinator (TAC) who will submit requests for accommodations to ACT. The TAC has access to two different forms specifically designed for the MME administration of the ACT:

- 1) **ACT-Approved Accommodations** – This form is used to request ACT approval of accommodations on the MME for students who meet ACT eligibility requirements. (See information about ACT’s review of these requests in the next section below.)
- 2) **State-Allowed Accommodations** – This form is used to order test materials for students who will test with “State-Allowed” accommodations. These students are those who do not meet ACT’s eligibility requirements (e.g., English language learners with no disabilities) or whose requests for ACT approval have been denied. ACT will ship the materials ordered for each student; no review or approval process will be conducted.

ACT Review of Requests for ACT-Approved Accommodations on the ACT (Day 1 of the MME)

ACT will review requests for ACT-Approved Accommodations by applying the Americans with Disabilities Act (ADA) standards that are used for all such requests. Not every request for an accommodation listed on the attached accommodations summary table as available will be approved. Approval is dependent on submission of all required documentation by the stipulated deadline and review by ACT. It is possible for ACT to approve an accommodation for one student, while the same accommodation may be denied for a different student. ACT’s decision whether to approve the requested accommodations under the ADA will

determine whether resulting ACT scores can be reported to colleges *in addition to* being used for MME scores.

Ordering State-Allowed Accommodations Materials for the ACT (Day 1 of the MME)

Students who do not meet ACT eligibility requirements (e.g., English language learners with no disabilities) or whose requested accommodations are denied by ACT have two options: 1) Test under standard conditions, or 2) Submit an order for “State-Allowed” accommodations materials. **IMPORTANT NOTE:** Students must submit and order for “State-Allowed” accommodations so that ACT can ship the correct ACT test materials – which are *different* from those used by examinees testing with ACT-Approved accommodations.

ACT scores resulting from testing with “State-Allowed” accommodations are **not** be college reportable, but will be used for MME scores. Thus, some students will achieve ACT scores that are college reportable because their accommodations have been approved by ACT, while others using the same accommodations will achieve ACT scores that are *not* college reportable because their use of those accommodations was not approved by ACT.

Local Decision for Accommodations on *WorkKeys* and Michigan Components (Day 2 and Day 3 of the MME)

There is **no** separate request form for accommodations on *WorkKeys* or the Michigan components of the MME. ACT’s approval of accommodations applies **only** to the administration of the ACT Plus Writing (Day 1). Schools are advised to use ACT’s approval as a guideline for ordering alternate formats (e.g., audio versions, large print) of the *WorkKeys* tests and Michigan components of the MME. Because there is no issue of reporting scores to colleges, schools may provide accommodations on the *WorkKeys* and Michigan components of the MME consistent with the accommodations listed in the “MME Day 2 and Day 3” columns of the attached accommodations summary table, even if the student tests without those accommodations on the ACT.

MME Accommodations Summary Table

The explanation of the Spring 2009 Michigan Merit Examination (MME) Accommodations Summary Table is arranged in columns in Table 4.10. The summary is presented in Table 4.11.

Table 4.10. Spring 2009 MME Accommodations Summary Table Explanation

Column	Explanation
Accommodation	Each accommodation that appears on the Assessment Accommodations Summary Table approved by the Michigan State Board of Education is listed.
MME Day 1 (The ACT Plus Writing)	
May Request	<p>ACT has indicated whether each accommodation may be requested for the ACT Plus Writing (Day 1 of the MME), or whether State-Allowed accommodated formats may be ordered.</p> <ul style="list-style-type: none"> • Accommodations for which local decisions may be made without a request to ACT are specifically noted (4, 5, 6, 7, 9, 10, 11, 12, 14, 15, 16, 33, 34, 35, 47, 48, 53). • Some formats or accommodations are noted as State-Allowed only (18, 19, 20, 26, 27, 28, 29, 30, 31, 44, 71). • A few accommodations are not allowed for the ACT (23, 37, 54). • Some accommodations do not apply to ACT (21, 24, 25, 36, 38, 49, 50, 51, 52, 62, 65). • Additional details about some accommodations are needed before a decision can be made for an individual request (13, 17, 56, 69, 76, 77, 78).
ACT Comments	These comments clarify ACT’s understanding of each accommodation and any associated restrictions.
College Reportable ACT Scores	<p>ACT has noted whether each accommodation that requires approval will result in ACT scores that are fully reportable to colleges and other entities <i>when approved by ACT for an individual student with disabilities</i>. If specific restrictions must be met or documentation from the test administration provided, these are also noted. The use of accommodations that require approval and which have not been approved by ACT for an individual student are eligible for State-Allowed accommodations testing. Taking the ACT Plus Writing with State-Allowed accommodations will result in ACT scores that are reportable <i>only</i> for MME scores (i.e., “State-Allowed” accommodations). If a student uses a combination of accommodations and <i>any</i> of those accommodations are State-Allowed (not ACT-Approved), the resulting scores will <i>not</i> be college reportable but can be used for MME Scores.</p> <p>NOTE 1: State-Allowed accommodations must be requested (ordered) from ACT so that the student receives accommodated test forms.</p> <p>NOTE 2: The use of accommodations considered Standard (S) for MME Day 2 and Day 3(see next column) will result in valid MME scores that may be used for the Michigan Promise scholarship and accountability. This is true for both ACT-Approved and State-Allowed accommodations.</p>

Table 4.11. Spring 2009 MME Accommodations Summary Table

Spring 2009 Michigan Merit Examination (MME) Accommodations Summary Table						
Accommodation	MME Day 1 (The ACT Plus Writing)			MME Day 2 and Day 3		
	May Request	ACT Comments	College Reportable ACT Scores ³	IEP	504	ELL
A. Timing/Scheduling						
1. Extended assessment time within reason (approximately 1½ times the estimated assessment time) NOTE: All MME tests are timed. Timing codes are assigned by ACT for Day 1. For Day 2 and Day 3, schools may allow time-and-a-half, double time, or a maximum of 3 hours for each test.	Yes Yes	Time-and-a-half in single self-paced session using regular or large-print. For certain formats and disabilities, ACT will assign a timing code for the ACT based on the test format and disability, up to triple time (and testing over multiple days, one test per day). Oral presentation (e.g., audiocassette, audio DVD, or reader), and Braille normally <i>require</i> triple time.	Yes Yes – only if ACT timing guidelines are followed	S	S	S
2. Frequent or appropriate supervised breaks	Yes	Interpreted as “stop-the-clock” breaks; normally available only with standard time. If requested with extended time, must provide documentation to support need for “stop-the-clock” breaks <i>in addition to</i> extended time.	Yes	S	S	S
3. Administration of the assessment at a time most beneficial to the student, with appropriate supervision	Yes	Must be within the designated two-week window that begins on initial state test day for that component and ends on the makeup day for that component. Testing may be scheduled for any days during the window, but each student must take the components of the MME in prescribed order, with all of Day 1 tests (ACT) completed before proceeding to Day 2 tests (<i>WorkKeys</i>) and all of Day 2 tests completed prior to beginning Michigan components (Day 3).	Yes	S	S	S

**Spring 2009 Michigan Merit Examination (MME)
Accommodations Summary Table**

Accommodation	MME Day 1 (The ACT Plus Writing)			MME Day 2 and Day 3		
	May Request	ACT Comments	College Reportable ACT Scores ³	IEP	504	ELL
4. Clock or method of informing students of remaining time	Local decision-but must adhere to all ACT directions	Five minutes remaining announcement routinely part of verbal instructions for <u>all</u> students on ACT. Students approved for time extensions on the ACT are given hourly announcements of time. No other assistance in monitoring time is allowed.	Yes	S	S	S

**Spring 2009 Michigan Merit Examination (MME)
Accommodations Summary Table**

Accommodation	MME Day 1 (The ACT Plus Writing)			MME Day 2 and Day 3		
	May Request	ACT Comments	College Reportable ACT Scores ³	IEP	504	ELL
B. Setting						
5. Placement of student where he/she is most comfortable (e.g., front of the room, back of the room)	Local decision unless requesting off-site	Because testing will be at own school rather than national test center, arrangement does not require ACT approval if no other accommodations requested. If setting is off-site, appropriate off-site application must be approved by ACT.	Yes	S	S	S
6. Administration of the assessment in a Bilingual/English as a Second Language (ESL) setting	Local decision unless requesting off-site	If setting is off-site, appropriate off-site application must be approved by ACT.	Yes	S	S	S
7. Administration of the assessment in a special education setting	Local decision unless requesting off-site	If setting is off-site, appropriate off-site application must be approved by ACT.	Yes	S	S	NA
8. Provision for assessment administration at home when student is homebound or in a care facility when medically necessary, with appropriate supervision by a school district professional.	Yes	Appropriate off-site application must be approved by ACT.	Yes	S	S	NA
9. Administration of assessment in a distraction free space or alternate location (e.g., separate room, or location within the room) with appropriate supervision	Local decision unless requesting off-site	Because testing will be at own school rather than national test center, arrangement does not require ACT approval if no other accommodations requested. If setting is off-site, appropriate off-site application must be approved by ACT.	Yes	S	S	NA
10. Provision for assessment administration to student in an interim alternative education setting with appropriate supervision of a school district professional.	Local decision unless requesting off-site	If setting is off-site, appropriate off-site application must be approved by ACT.	Yes	S	S	NA
11. Administration of the assessment in a small group	Local decision unless requesting off-site	Because testing will be at own school rather than national test center, arrangement does not require ACT approval if no other accommodations requested. If setting is off-site, appropriate off-site application must be approved by ACT.	Yes	S	S	S

**Spring 2009 Michigan Merit Examination (MME)
Accommodations Summary Table**

Accommodation	MME Day 1 (The ACT Plus Writing)			MME Day 2 and Day 3		
	May Request	ACT Comments	College Reportable ACT Scores ³	IEP	504	ELL
12. Administration of the assessment individually	Local decision unless required by approved accommodation	Because testing will be at own school rather than national test center, arrangement does not require ACT approval if no other accommodations requested. Note that individual testing is required for selected accommodations (e.g., if approved accommodations could disturb others or if approved for a reader).	Yes	S	S	NA
13. Tools to assist with concentration	Submit details with request	Requests considered individually based on documentation submitted. Approval and reportable status depend on detailed information about the tools proposed for use.	Depends on details	S	S	NA
14. Qualified person familiar to the student administers the assessment	Local decision-staff must meet all ACT requirements	Only if not a relative or athletic coach (if student is an athlete). See also #34 and #35.	Yes	S	S	S
15. Appropriate seating, special lighting, or furniture	Local decision	Provided by the school.	Yes	S	S	NA
16. Able to move, stand or pace during assessment in a manner where others' work cannot be seen and is not distracting to others	Local decision	Because testing will be at own school rather than national test center, arrangement does not require ACT approval if no other accommodations requested.	Yes	S	S	S
17. Background music or noise buffers	Submit details with request	Requests considered individually based on documentation submitted. Music and earplugs not normally approved. Approval and reportable status depend on detailed information about the buffers proposed.	Depends on details	S	S	NA
C. Presentation						
18. Use of bilingual word-for-word non-electronic translation glossary for English language learners	Yes (State-Allowed only)	Provided by school or student.	No	S	S	S
19. Use of bilingual dictionaries that define or explain words or terms	Yes (State-Allowed only)	Provided by school or student.	No	NS	NS	NS

**Spring 2009 Michigan Merit Examination (MME)
Accommodations Summary Table**

Accommodation	MME Day 1 (The ACT Plus Writing)			MME Day 2 and Day 3		
	May Request	ACT Comments	College Reportable ACT Scores ³	IEP	504	ELL
20. Use of dictionary, thesaurus, spelling book, or grammar book for mathematics, science, social studies, and English language arts	Yes (State-Allowed only)	Provided by school or student.	No	NS	NS	NS
21. Use of screen reader for English language arts reading assessment	NA		NA	NS	NS	NS
22. Use of an abacus	Yes	Provided by school or student; student must test individually.	Yes	S	S	NA
23. Use of arithmetic tables	No	Arithmetic tables not allowed for the ACT or <i>WorkKeys</i> .	NA	NS	NS	NS
24. Use of actual coins and bills	NA	Items do not involve this kind of manipulation.	NA	S	S	NA
25. Use of manipulatives for mathematics assessments, such as base 10 blocks	NA	Items do not involve this kind of manipulation.	NA	NA	NA	NA
26. Use of state-produced video or audio version of assessment, for English language learners, <u>read in English</u> for a student who is dominant in a native language other than English or determined to be at the basic or lower intermediate English language proficiency levels in the content areas of mathematics, science, and social studies. Also the writing section of the MEAP ELA or MI-Access ELA Expressing Ideas assessment.	Yes (State-Allowed only)	If student's reason for accommodations is English language proficiency, student must order "State-Allowed" accommodations materials.	No	S	S	S

**Spring 2009 Michigan Merit Examination (MME)
Accommodations Summary Table**

Accommodation	MME Day 1 (The ACT Plus Writing)			MME Day 2 and Day 3		
	May Request	ACT Comments	College Reportable ACT Scores ³	IEP	504	ELL
27. Use of state-produced video or audio version of the assessment, for English language learners, <u>read in English</u> for a student who is dominant in a native language other than English or determined to be at the basic or lower intermediate English language proficiency levels in the reading components of the English language arts assessment.	Yes (State-Allowed only)	If student’s reason for accommodations is English language proficiency, student must order “State-Allowed” accommodations materials.	No	S	S	S
28. Use of state-produced video or audio version, for English language learners, of the mathematics, science, or social studies assessments <u>read in Arabic or Spanish</u> for a student whose dominant language is Arabic or Spanish or who is determined to be at the basic or lower intermediate English language proficiency levels, and provided that the student is receiving bilingual instruction (e.g., transitional, two-way, or dual language) using the student’s native languages in the school setting.	Yes (State-Allowed only)	If student’s reason for accommodations is English language proficiency, student must order “State-Allowed” accommodations materials.	No	S	S	S

**Spring 2009 Michigan Merit Examination (MME)
Accommodations Summary Table**

Accommodation	MME Day 1 (The ACT Plus Writing)			MME Day 2 and Day 3		
	May Request	ACT Comments	College Reportable ACT Scores ³	IEP	504	ELL
29. Reading all directions to the student in the <u>student's native language</u> , provided that the student is dominant in a native language other than English or has been determined to be at the basic or lower intermediate English language proficiency levels and provided that the student is receiving bilingual instruction (e.g., transitional, two-way or dual language) using the student's native language in the school setting.	Yes (State-Allowed only)	If student's reason for accommodations is English language proficiency, student must order "State-Allowed" accommodations materials.	No	S	S	S
30. Provision for student restatement of directions in the student's own words	Yes (State-Allowed only)	Only if tested individually.	No	S	S	S
31. Students asking for clarification of directions	Yes (State-Allowed only)	Only if tested individually.	No	S	S	S
32. Directions provided using sign language	Yes	Applies only to <u>spoken</u> instructions exactly as provided in the administration manual.	Yes	S	S	NA
33. Administration of assessment by Bilingual/ESL staff, or similarly qualified person	Local decision-staff must meet all ACT requirements	Only if all directions for test administration are read verbatim in English with no clarifications in another language.	Yes	S	S	S
34. Administration of the assessment by person familiar to the student	Local decision-staff must meet all ACT requirements	Only if not a relative or athletic coach (if student is an athlete). See also #14 and #35.	Yes	S	S	S

**Spring 2009 Michigan Merit Examination (MME)
Accommodations Summary Table**

Accommodation	MME Day 1 (The ACT Plus Writing)			MME Day 2 and Day 3		
	May Request	ACT Comments	College Reportable ACT Scores ³	IEP	504	ELL
35. Any assessment administration not directly supervised by a school district professional	Local decision-staff must meet all ACT requirements	For state testing, ACT administration manual states that testing staff may be “current or retired faculty members, school administrative or clerical employees, substitute teachers, student teachers, and teachers’ aides.” Staff may not be “volunteers.” In addition: “High school students and lower-division undergraduates may not work as testing staff. Anyone who intends to take the ACT within the next 12 months must not administer the test in any capacity.” Additional restrictions regarding relatives and athletic coaches also apply.	Yes	NS	NS	NS
36. Reading the MEAP English Language Arts Listening assessment to the student in his/her native language	NA		NA	NS	NS	NS
37. Administer assessment sections in any order for English language arts, science, and social studies	No	ACT tests must always be administered in prescribed sequence.	NA	S	S	S
38. Administer assessment sections in any order for Mathematics	NA	ACT Mathematics test is not in sections.	NA	S	S	S
39. Read/repeat directions to the student exactly as worded in the assessment booklet	Yes	Directions in the test booklet not normally read aloud. Permitted for college reportable ACT scores only if approved for reader or audio version of test.	Yes	S	S	S
40. Emphasis on key words in directions	Yes	Directions in the test booklet not normally read aloud. Permitted for college reportable ACT scores only if approved for reader or audio version of test. Emphasis only as marked in the printed directions; must be read verbatim without signals regarding right or wrong.	Yes	S	S	NA
41. Provide visual, auditory or physical cues to student to begin, maintain or finish task	Yes	If cues will disturb other examinees, must test individually.	Yes	S	S	NA
42. Reading aloud the reading components of the	Yes	Must be read in English . For college reportable ACT scores, must test individually if not using audio version with headset (see #61 for audio version).	Yes	S	S	S

**Spring 2009 Michigan Merit Examination (MME)
Accommodations Summary Table**

Accommodation	MME Day 1 (The ACT Plus Writing)			MME Day 2 and Day 3		
	May Request	ACT Comments	College Reportable ACT Scores ³	IEP	504	ELL
43. Reading aloud the mathematics, science and social studies components of the test	Yes	Must be read in English . For college reportable ACT scores, must test individually if not using audio version with headset (see #61 for audio version).	Yes	S	S	S
44. Reading of mathematics, social studies, and science assessment content and questions to a student <u>in the student's native language</u> , provided that the student is dominant in a native language other than English or has been determined to be at the basic or lower intermediate English language proficiency levels, and provided that the student is receiving bilingual instruction (e.g., transitional, two-way, or dual language) using the student's native language in the school setting.	Yes (State-Allowed only)	If student's reason for accommodations is English language proficiency, student must order "State-Allowed" accommodations materials.	No	S	S	S
45. Sign the mathematics, science and social studies assessments	Yes	Exact English Signing (EES) of test items may be requested and approved in specific cases for college reportable scores. Signing of items with American Sign Language (ASL) or other sign language is not ACT-Approved.	Yes – only if EES approved by ACT No – if ASL or other sign language	S	S	NA
46. Sign the English language arts assessments	Yes	Exact English Signing (EES) may be requested and approved in specific cases for college reportable scores. Signing of items with American Sign Language (ASL) or other sign language is not ACT-Approved.	Yes – only if EES approved by ACT No – if ASL or other sign language	NS	NS	NA
47. Use of a page turner	Local decision-staff must meet all ACT requirements	Because testing will be at own school rather than national test center, arrangement does not require ACT approval if no other accommodations requested. Page turner must meet same requirements as all testing staff.	Yes	S	S	NA

**Spring 2009 Michigan Merit Examination (MME)
Accommodations Summary Table**

Accommodation	MME Day 1 (The ACT Plus Writing)			MME Day 2 and Day 3		
	May Request	ACT Comments	College Reportable ACT Scores ³	IEP	504	ELL
48. Placement of teacher/proctor near student	Local decision	Because testing will be at own school rather than national test center, arrangement does not require ACT approval if no other accommodations requested.	Yes	S	S	NA
49. Use of rulers as provided by the State	NA	Items do not require rulers.	NA	S	S	S
50. Use of adapted rulers, protractors, Braille and large print rulers and protractors.	NA	Items do not require rulers or protractors.	NA	S	S	NA
51. Use of list of formulae as provided by the state	NA	No formulae allowed for ACT tests.	NA	S	S	S
52. Use of calculator/talking calculator on the noncalculator sections of the mathematics assessment	NA	There are no “noncalculator” sections of the ACT Mathematics test. See also #53.	NA	NS	NS	NA
53. Use of calculator/talking calculator on the calculator permitted sections of the mathematics assessment	Local decision-calculator must meet all ACT requirements	Calculators are permitted throughout the ACT Mathematics test (except those listed by ACT as “prohibited” in publications and on website). If talking calculator, student must test individually.	Yes	S	S	S
54. Use of a calculator on the science and social studies assessments	No	Calculators are permitted only on the ACT Mathematics, not any other tests.	NA	NA	NA	NA
55. Use of magnification devices	Yes	Provided by school or student. May require student to test individually.	Yes	S	S	NA
56. Use of auditory amplification devices or special sound systems	Submit details with request	Used only for spoken instructions. Requests considered individually based on documentation submitted. Approval and reportable status depend on detailed information about proposed devices or systems.	Depends on details	S	S	NA
57. Use of closed circuit television	Yes	Provided by school or student. Student must test individually.	Yes	S	S	S
58. Student’s use of acetate colored shield, highlighters, highlighter tape, page flags, and reading guides.	Yes	Provided by school or student. “Reading guides” are interpreted as place-keepers. May require student to test individually (e.g., highlighters).	Yes	S	S	NA

**Spring 2009 Michigan Merit Examination (MME)
Accommodations Summary Table**

Accommodation	MME Day 1 (The ACT Plus Writing)			MME Day 2 and Day 3		
	May Request	ACT Comments	College Reportable ACT Scores ³	IEP	504	ELL
59. Use of non-skid surface that will not damage the answer document or scanning equipment (DO NOT use tape or other adhesive)	Yes	Provided by school or student.	Yes	S	S	NA
60. State produced Braille and enlarged print versions of assessment	Yes		Yes	S	S	NA
61. State produced audio versions of the assessments (ELA , mathematics, science, social studies)	Yes	Must use headset if testing in a group.	Yes	S	S	S
D. Response						
62. Responding in the student's native language to the constructed response items on assessments.	NA	The only constructed response is the ACT Writing Test, and it must be written in English .	NA	NA	NA	NA
63. Oral responses	Yes	Only if tested individually, responses are in English , and responses marked on scannable document by testing staff. For college reportable ACT scores, session must be tape recorded with recording also returned to ACT.	Yes	S	S	NA
64. Use of a scribe for constructed response items (student must indicate punctuation, format and spell all key words) for ELA assessments	Yes	Applies only to ACT Writing Test. Only if tested individually. For college reportable ACT scores, session must be tape recorded with recording also returned to ACT.	Yes – only if recording of test session returned to ACT	S	S	NA
65. Use of a scribe for constructed response items for mathematics, science and/or social studies assessments	NA	No constructed response items in these subjects on ACT.	NA	S	S	S
66. Student dictates responses into a tape recorder and teacher transcribes response exactly as dictated for mathematics, science, and social studies assessments.	Yes	Only if tested individually and responses are in English . For college reportable ACT scores, tape recording must be returned to ACT.	Yes – only if recording of test session returned to ACT	S	S	NA

**Spring 2009 Michigan Merit Examination (MME)
Accommodations Summary Table**

Accommodation	MME Day 1 (The ACT Plus Writing)			MME Day 2 and Day 3		
	May Request	ACT Comments	College Reportable ACT Scores ³	IEP	504	ELL
67. Respond in sign language for English language arts	Yes	Only if tested individually and responses marked on scannable document by testing staff. For college reportable ACT scores, video documentation of test session must be returned to ACT. Sign language response to ACT Writing Test must be Exact English Signing (EES).	Yes – only if recording of test session returned to ACT and Writing Test signed EES	S	S	NA
68. Respond in sign language for mathematics, science and social studies assessments	Yes	Only if tested individually and responses marked on scannable document by testing staff. For college reportable ACT scores, video documentation of test session must be returned to ACT.	Yes – only if recording of test session returned to ACT	S	S	NA
69. Use of augmentative communication devices	Submit details with request	Requests considered individually based on documentation submitted. Approval and reportable status depend on detailed information about the devices proposed for use.	Depends on details	S	S	NA
70. Use of computer or word processor with spell check, thesaurus, and grammar check <i>disabled</i> for ELA assessment.	Yes	Applies only to ACT Writing Test.	Yes	S	S	NA
71. Use of computer or word processor with spell check, thesaurus and grammar check NOT disabled for Mathematics, Science and Social Studies.	Yes (State-Allowed only)		No	S	S	NA
72. Student points to answers or writes directly in assessment booklet (transferred to answer document by teacher)	Yes	If student points to answers, student must test individually.	Yes	S	S	S
73. Use of Braillewriter	Yes	Provided by school or student.	Yes	S	S	NA
74. Use of a scribe for constructed response items (student must indicate punctuation and spell all key words)	Yes	Applies only to ACT Writing Test (see #64). Only if tested individually. For college reportable ACT scores, session must be tape recorded with recording also returned to ACT.	Yes – only if recording of test session returned to ACT	S	S	NA
75. Adapted paper, lined or grid paper for recording answers	Yes	Provided by school. Student must test individually and responses transferred to scannable answer document by testing staff while examinee observes.	Yes	S	S	NA

**Spring 2009 Michigan Merit Examination (MME)
Accommodations Summary Table**

Accommodation	MME Day 1 (The ACT Plus Writing)			MME Day 2 and Day 3		
	May Request	ACT Comments	College Reportable ACT Scores ³	IEP	504	ELL
76. Use of computers with alternative access for an alternative response mode	Submit details with request	Requests considered individually based on documentation submitted. Approval and reportable status depend on detailed information about the proposed alternative access	Depends on details	S	S	NA
77. Use of speech to text word processor for responses for English language arts	Submit details with request	Requests considered individually based on documentation submitted. Approval and reportable status depend on detailed information about the proposed speech to text processor.	Depends on details	NS	NS	NA
78. Use of speech to text word processing for mathematics, science and social studies	Submit details with request	Requests considered individually based on documentation submitted. Approval and reportable status depend on detailed information about the proposed speech to text processor.	Depends on details	S	S	NA
79. Use of alternative writing position	Yes	If position will disturb other examinees, must test individually	Yes	S	S	NA
80. Use of special adaptive writing tools such as pencil grip or larger pencil.	Yes	Provided by school or student.	Yes	S	S	NA
81. Write directly in assessment booklet	Yes	Only if responses transcribed to scannable answer document by testing staff while examinee observes.	Yes	S	S	S

Chapter 5: Test Development Analyses

MME Components

The MME is composed of the following components for each subject. This structure is based on the results of alignment analyses between the ACT Plus Writing, WorkKeys assessments, and Michigan High School Content Standards, as well as mandates from Michigan legislation. See Table 5.1.

Table 5.1. Components of MME Test Scores

				Components Contributing to MME Scores*				
Day	Test	Subject Session	Total ELA	Reading	Writing	Mathematics	Science	Social Studies
Day 1	ACT Plus Writing	English	X		X			
		Mathematics				X		
		Reading	X	X				
		Science					X	
Day 2	WorkKeys	<i>Reading for Information</i>	X	X				
		<i>Applied Mathematics</i>				X		
		<i>Locating Information</i>				8 items		6 items
Day 3 or 4	Michigan Mathematics, Science and Social Studies	Science					X	
		Social Studies						X
		Mathematics				X		

Note that the ACT Plus Writing was given on day 1 of the assessment, the *WorkKeys* tests were given on day 2, and the Michigan components were given on the third day. For each subject (column), students needed to complete (meet the attemptedness criteria for) each section shown with an “X” to obtain a valid score on the MME.

Eight of the *WorkKeys Locating Information* items count towards MME Mathematics and six of the *WorkKeys Locating Information* items count towards MME Social Studies. This occurs because these items align well with Michigan’s high school mathematics or social studies content expectations.

The MME ELA is an average of MME Writing and MME Reading. It consists of four components, as shown in Table 5.1.

Test Specifications and Alignment Between Contributing Components

Because intact ACT Plus Writing and *WorkKeys (Reading for Information, Locating Information and Applied Mathematics)* assessments must be included in the Michigan Merit Examination (MME), the MME test specifications must start with an analysis of the combined alignment of the ACT Plus Writing and *WorkKeys* assessments. This analysis is the foundation for creating the augmentation needed to

assure sufficient alignment of the MME as a whole in each subject to Michigan's high school content standards.

To ensure that the augmented portion of the MME fulfills the requirements for alignment to Michigan's high school content standards, a yearly alignment process is undertaken. This process is described in detail below. In addition, several in-depth alignment analyses were conducted during the development of the Michigan Merit Examination. These are detailed below in the "Historical Alignment Analyses" section, and is adapted from the materials submitted to the United States Department of Education for peer review of the MME prior to the first implementation in 2007. The evidence referenced in this section is provided as addenda to this technical report.

Alignment of the 2009 MME with HSCEs: Item Selection for Day 1 and Day 1 Scoring

In July of 2008, specialists in language arts, mathematics, science, and social studies at OEAA conducted an alignment study of the ACT and *WorkKeys* test forms to be used on the Spring 2009 MME. OEAA staff reviewed secure copies of the test booklets and coded each test item to the High School Content Expectation (HSCE) the item most clearly measured. (Specialists in both mathematics and social studies reviewed the *WorkKeys Locating Information* forms.) If an item did not appear to measure an HSCE, the item was left uncoded. A tally of the assigned codes was made for each test form, by standard, to determine the breadth of standards coverage for each form. For both the ACT and *WorkKeys* tests, these tallies differed across the four administration types (initial, makeup, accommodated, and emergency). Upon reviewing the tallies OEAA decided that, rather than have all ACT and *WorkKeys* items count toward students' MME scores, only a subset of items would count. These items would be selected so that the four ACT forms would have a common distribution of MME-scored items by standard, and likewise for the four forms of each *WorkKeys* test. This would ensure that the MME-scored items taken by each student would cover the standards identically, regardless of the combination of Day 1 and Day 2 forms the student took.

In August of 2008, OEAA content specialists conducted a second review of the Day 1 and Day 2 forms to select these items. Table 5.2 shows the number of items on each Day 1 and Day 2, and how many were selected for MME scoring. Note that half of the items on each ACT English, Mathematics, and Science Test, and half of items from each *WorkKeys Reading for Information* and *Applied Mathematics* Test, were selected. Nearly all (34 out of 40) ACT Reading Test items were selected. For each *WorkKeys Locating Information* form, eight items were selected to count toward MME Mathematics scores, and six were selected to count toward MME Social Studies scores.

As mentioned above, the purpose of selecting items for MME scoring was to ensure that the MME-scored items taken by every student would cover the standards identically. Table 5.3 shows how the selected items from each Day 1 or Day 2 test were distributed across the standards for that content area.

Table 5.2. Number of items and number of selected items on each MME Day 1 or Day 2 test

Test	Number of scored items	Number of items scored for MME
ACT English	75	38
ACT Writing	one prompt	one
ACT Reading	40	34
ACT Mathematics	60	30
ACT Science	40	20
<i>WorkKeys Reading for Information</i>	30	15
<i>WorkKeys Applied Mathematics</i>	30	15
<i>WorkKeys Locating Information</i> (scored as Mathematics)	32	8
<i>WorkKeys Locating Information</i> (scored as Social Studies)	32	6

Table 5.3. Number of selected items by Day 1 and Day 2 test and by standard

Test	Standard	Number of selected items
ACT English	W1.1	30
	W1.3	8
ACT Writing	W1.3	one prompt (12 points)
ACT Reading	R2.1	18
	R2.2	11
	L3.1	5
<i>WorkKeys Reading for Information</i>	R2.1	4
	R2.2	1
	R2.3	10
ACT Mathematics	L1	1
	L2	2
	A1	10
	A2	3
	A3	2
	G1	11
	G2	1
<i>WorkKeys Applied Mathematics</i>	L2	10
	G1	4
	S1	1
<i>WorkKeys Locating Information (scored as Mathematics)</i>	L1	3
	S1	5
ACT Science	R1	20
<i>WorkKeys Locating Information (scored as Social Studies)</i>	Inquiry	6

Test Development for Michigan Components

In developing the augmentation of ACT Plus Writing and WorkKeys to produce the overall MME, it was not feasible to employ many of the procedures that the Michigan Department of Education typically employs for test development because the spring 2007 administration of the Michigan Merit Examination (MME) was the first administration of a new assessment using a new scale, and because two components of the MME are pre-designed by ACT. Therefore, there did not exist any Item Response Theory (IRT) item parameter estimates for items to be used on the spring 2007 administration (with the exception of items used to link to the pilot study of spring 2006). All analyses used to support test development had to be performed using classical test theory (CTT) statistics. However, for the spring 2008 administration, IRT parameter estimates were available for many items. The inclusion rules were, in order of decreasing importance, the following:

1. Alignment to content standards needing augmentation.

2. Positive corrected point-biserial correlations with either the MME pilot or past MEAP high school scores (preferably above 0.25, but no negatives) where statistics were available.
3. Creation of a reasonable distribution of classical item difficulty where statistics were available, meaning approximately one third of the items in each of the following ranges: 0.26-0.50, 0.51-0.75, and 0.76-1.00. Generally, we do not select items in the range of 0.00- 0.25 unless such items are absolutely needed for content alignment.
4. IRT parameter estimates were reviewed when available.

Because classical statistics were gathered from different sources (the MME pilot versus previous assessments) the distributions are not presented as the statistics do not all come from the same population.

For future cycles of the MME, more sophisticated analyses will be run for developing the assessments to ensure that they will be equitable. These include analyses of the distribution of IRT parameters, projected SEM/Information curves, projected reliability, and projected classification accuracy. The comparison with the baseline (previous year) will be included with current projections to evaluate the overall similarity of each year's assessment to the previous year.

NOTE: Item development for the augmented portion of the MME occurred during the period of the previous High School assessment (the Michigan Educational Assessment Program, or MEAP). The item development protocols and quality assurance checks are detailed in the 2005/06 final MEAP technical report.

Historical Alignment Analyses Prior to 2009 Administration

Three independent alignment studies were conducted on the ACT and *WorkKeys* against Michigan High School content standards before the pilot of the MME was created.

First, Norman L. Webb, a senior research scientist with the Wisconsin Center for Education Research and the National Institute for Science Education, conducted a preliminary alignment study of the ACT and *WorkKeys* to the Michigan content standards in December, 2004 as a first step in determining the feasibility of combining a college-entrance exam with an NCLB-compliant, standards-based exam. The evidence in these reports was used to target augmentation to the ACT and *WorkKeys* to maximize alignment to the Michigan standards in the pilot of the MME. These reports indicated that of the Michigan ELA standards that are assessable on a large scale, the ACT and *WorkKeys* combination was well aligned to Michigan's high school standards, with some minor improvements possible. The reader is referred to page 15 of *Alignment Analysis of Language Arts Standards and Assessments: Michigan Grades 9–12*. (Webb, 2005). These reports documented some areas of weakness in mathematics and science. The weaknesses in mathematics are summarized on page 13 of *Alignment Analysis of Mathematics Standards and Assessments: Michigan High School*. (Webb, 2005). The weaknesses in science are summarized on pages 15-16 of *Alignment Analysis of Science Standards and Assessments: Michigan Grades 9–12*. (Webb, 2005). Augmentation was targeted to the weak areas.

Second, John Dossey of Illinois State University evaluated the Mathematics and Science ACT Test items and *WorkKeys* items in comparison to the Michigan Mathematics and Science High School Content Standards. He identified remarkable consistency between the ACT/*WorkKeys* and the Michigan content standards, with a few areas of weakness. The weaknesses he identified were in mathematical content coverage of patterns, functions, probability and discrete mathematics, as described on page 14 of *Comparison of the ACT and WorkKeys Assessments with the Mathematics and Science Content Expectations in the Michigan Curriculum Framework*. (Dossey, 2005). Although science was well covered, identified weaknesses in life, physical, and earth science are summarized on page 20 of the same document (Dossey, 2005). Augmentation was targeted to maximize alignment on these areas.

Third, Timothy Shanahan of University of Illinois at Chicago evaluated the ACT and *WorkKeys* items in comparison to the Michigan English Language Arts (ELA) content standards. In summary, Shanahan clearly states on page 7 of *Review of ACT Coverage of Michigan Language Arts Standards* (Shanahan, 2005) that the ACT English and Reading assessments are strongly aligned with the Michigan ELA content standards. Although the alignment study suggested no need to further augment the ELA portion of the assessment, OEAA chose to augment the Writing portion. Specifically, in order to resolve an issue with balance of representation, a score for Social Studies Decision Making (constructed response item) was added to the Writing total score. This addition offset the large number of English Multiple Choice points that were being counted as part of the Writing score.

Post-Hoc Alignment Studies of the Pilot Michigan Merit Exam

Norm Webb from the University of Wisconsin led another alignment study for the Michigan Merit Examination pilot in May, 2006, involving curriculum, instruction and assessment experts from within and outside of the State. For the English Language Arts (ELA) and mathematics portions of the MME, alignment was considered in regard to both the current (2004) Michigan Curriculum Framework Standards and Benchmarks and the soon-to-be-implemented (2006) High School Content Expectations.

Below, findings are presented only from the alignment with the 2004 Michigan Curriculum Framework Standards and Benchmarks.

Members of the alignment teams were solicited from a diverse group of educators who had not previously taken part in developing the assessment instruments, in order to ensure the objectivity of the study.

The alignment studies indicated the following for the individual content areas:

- For ELA, seven of the twelve (2004) standards were reasonably addressed by an on-demand assessment, as stated on page 10 of *Alignment Analysis of Reading and Language Arts Standards and Michigan Merit Exam: Michigan High School* (Webb, 2006). The MME demonstrated Categorical Concurrence for all seven standards (see page 9). Five standards showed Depth-of-Knowledge Consistency and Range of Knowledge, and all but one had an appropriate Balance of Representation.
- For mathematics, there were six (2004) standards, all of which were addressed in an on-demand assessment. As described in *Alignment Analysis of Mathematics Standards and Michigan Merit Exam: Michigan High School* (Webb, 2006), the MME demonstrated Categorical Concurrence on all six standards. Four standards showed Depth-of-Knowledge Consistency, two had an acceptable Range of Knowledge, and all but one had an appropriate Balance of Representation.
- For science, the panel concluded that the alignment is reasonable if only the benchmarks that are more suitably assessed by an on-demand assessment are considered. These analyses are described in *Alignment Analysis of Science Standards and Michigan Merit Exam: Michigan High School* (Webb, 2006). Of the five 2004 standards, all but “Reflecting on Scientific Knowledge” demonstrated Categorical Concurrence. This was corrected beginning with the Spring 2007 MME by adding six items assessing Reflecting on Scientific Knowledge. These items were selected to also address depth of knowledge, range of knowledge, and balance of representation. Of the remaining standards, all showed Depth-of-Knowledge Consistency, three had an acceptable Range of Knowledge, and all had an appropriate Balance of Representation.

The new Michigan Merit Examination (MME) is based on the ACT Plus Writing and three *WorkKeys* assessments (*Reading for Information*, *Applied Mathematics* and *Locating Information*), with Michigan-developed augmented portions designed to address standards not covered by the ACT tests and the *WorkKeys* assessments. In assembling the Michigan-developed component for MME, the post-hoc alignment studies were used to indicate areas where the ACT and *WorkKeys* tests need to be augmented.

From the results of the post-hoc alignment studies, it appears that the targeted augmentations of the Mathematics and Science assessments were effective.

Chapter 6: Erasure Analyses

Description and Purpose

Erasure analysis (also known as mark darkness analysis) is the study of the degree to which certain groups of students tend to mark and then erase those marks on multiple choice items. The purpose is to identify unusually low or unusually high rates of answer-changing behavior as circumstantial evidence to support investigations in situations where allegations of widespread cheating have been received and to identify plausible targets for on-site monitoring.

Data and Methods

The data captured to analyze erasure patterns is described here. In a data file with one row per student per subject, the following data are captured:

- DistrictCode (NULL for state rollup)
- BuildingCode (NULL for district rollup)
- Grade (NULL for all grades rollup)
- Subject (NULL for all subjects rollup)
- NW2W (Number of wrong to wrong erasures)
- NW2R (Number of wrong to right erasures)
- NR2W (Number of right to wrong erasures)

Based on the form of the assessment and upon the data already in the file, the following two fields are added to the student-level file:

- Nerase (Total number of erasures, or $NW2W + NW2R + NR2W$)
- Ntotal (Total number of MC items responses)

From these data, summary data files are created with one row for each district/school/grade/subject combination. Each row of the file contains the following data:

- DistrictCode
- BuildingCode (NULL for district rollups)
- Grade
- Subject
- DistrictCode (NULL for state rollup)
- BuildingCode (NULL for district rollup)
- Grade (NULL for all grades rollup)
- Subject (NULL for all subjects rollup)
- NW2W (sum of wrong to wrong erasures over all students)
- NW2R (Number of wrong to right erasures over all students)
- NR2W (Number of right to wrong erasures over all students)
- Nerase (Total number of erasures, or $NW2W + NW2R + NR2W$)
- Ntotal (Total number of MC items responses)

From the data in the summary file, two additional fields are created for each row as follows:

R1 (ratio of all erasures to all responses in the combination, or $N_{\text{erase}}/N_{\text{total}}$)

R2 (ratio of wrong-to-right erasures to all erasures in the combination, or $NW2R/N_{\text{erase}}$)

Based upon the data in this file, four threshold values are calculated for each statistic and each subject at the district level and at the school level. These thresholds are based on the distributions of the ratio statistics at the district and school level. These thresholds may change based on their usefulness in operation, but current plans are that they will be:

1. 3SDlow (3 standard deviations below the mean or zero, whichever is greater)
2. Prcntlow (The 5th percentile)
3. 3SDhigh (3 standard deviations above the mean)
4. Prcnhigh (The 95th percentile)

The following flags are applied in the summary data files, based on the thresholds above:

- R1LowSD (1 if less than 3SDlow, 0 otherwise for R1)
- R1LowPct (1 if less than Prcntlow, 0 otherwise for R1)
- R1HighSD (1 if greater than 3SDhigh, 0 otherwise for R1)
- R1HighPct (1 if greater than Prcnhigh, 0 otherwise for R1)
- R2LowSD (1 if less than 3SDlow, 0 otherwise for R2)
- R2LowPct (1 if less than Prcntlow, 0 otherwise for R2)
- R2HighSD (1 if greater than 3SDhigh, 0 otherwise for R2)
- R2HighPct (1 if greater than Prcnhigh, 0 otherwise for R2)

Based on these flags, district/school/grade/subject combinations with unusually low or unusually high ratios are identified. The criteria for identifying individual combinations will need to be determined through more experience with operational data.

However, there will be at least two uses of the data. First, these data will be used as evidence in investigations following up on allegations of unethical behavior. Second, these data will be used to target individual schools and/or districts for on-site monitoring by MDE and/or contractor staff during the next assessment cycle. It is expected that the erasure data will also be useful in research on erasure patterns as related to item characteristics.

Because the behaviors of these summary statistics are not well known, either in a univariate or bivariate fashion, summary statistics will also be presented to inform OEAA understanding. These summaries will display both graphically and numerically the univariate and bivariate distributions of the ratio statistics, thresholds, and flags where the displays are reasonable. These displays will aid in future construction of erasure analysis indices.

Day 1 and 2 Analysis

Overview

The Day 1 and Day 2 systems employed by ACT each generated an item response file for OEAA's use in erasure analysis. The files include one record, representing one answer document, for each Select record provided.

Input

Three types of files are input to the process:

Mark Intensity Files

These contain data from the original scanning of the answer documents, including mark intensities for the test items. There is a mark intensity value for each possible test item response (oval), using a 16 intensity level scan. The lowest mark intensity values (0–3) are not used.

Mark Intensity Files do not identify erasures per se; they identify scan intensity values. Scan intensity is based on a number of factors such as area completed and darkness (light reflection). There is no automated way to distinguish a light mark from an erasure or dark erasure from a true mark. The scanner can only measure that one item response has more intensity (or the same intensity) as another. For building of the erasure patterns, the process assumes the item response with the least intensity is an erasure within an item.

Select File(s)

These contain student records, including the item responses (one response for each item) that were used for scoring. For each item, the scored response is the one for which the darkest mark was scanned (unless there is a double grid, defined below).

Scoring Keys

To identify the correct response to each item.

Output

Match to mark intensity files

For each Select record, the matching mark intensity record is found using the batch and PAS or UIN (a unique identifier for the answer document) for Day 1 and the batch and UIN for Day 2.

Create item response record

An item response record, containing response analysis values, is created for each Select record.

Response analysis values and item response file layout are shown in Tables 6.1 and 6.2 below.

Table 6.1. Day 1 and Day 2 Response Analysis Values for Erasure Analysis

Categories	Proposed Response Analysis Value	How determined
“No item response or no erasure detected (normal mark)”	0	There is no more than one mark intensity value for the item.
“Incorrect response changed to correct response”	1	The scored response is correct, and there are one or more marks with a lesser intensity value for the item.
“Correct response changed to an incorrect response”	2	The scored response is incorrect, and there is a lesser mark intensity value for the correct response (but not a double grid).
“Incorrect response changed to another incorrect response”	3	The scored response is incorrect (and is not a double grid), and there are one or more marks with a lesser intensity value for the item that are also incorrect and the correct response has not been marked (i.e., erased).
“Double grid”	4	The two highest mark intensity values for the item are equal intensity or side-by-side on the intensity scale.

Table 6.2. Day 1 and Day 2 Erasure Data Item response file layout

Pipe-Delimited Field #	Fixed Field Start	Fixed Field End	Field Name	ACT (Day 1)		WorkKeys (Day 2)	
				Length	Content	Length	Content
1	1	1	Program	1	“1”	1	“2”
2	2	11	Student Barcode Number	10	Select file pos. 412–421	10	Select file pos. 458–467
3	12	19	Student Batch/Process Number	4	Select file pos. 1016–1019	8	Select file pos. 475–482
4	20	25	Student PAS	6	Select file pos. 1029–1034	0	n/a for Day 2 as a separate field (Day 2 PAS is included in field #5 below.)
5	26	45	Student UIN	20	Select file pos. 686–705	6	Select file pos. 469–474
6	46	120	Test 1 Response Analysis	75	Calculated from Select file pos. 436–510	33	Calculated from Select file pos. 192–224
7	126	185	Test 2 Response Analysis	60	Calculated from Select file pos. 511–570	33	Calculated from Select file pos. 279–311
8	186	225	Test 3 Response Analysis	40	Calculated from Select file pos. 571–610	38	Calculated from Select file pos. 366–403
9	226	265	Test 4 Response Analysis	40	Calculated from Select file pos. 611–650	0	n/a for Day 2
10	266	271	Test Site ACT Code	6	Select file pos. 204-209	6	Select file pos. 136-141

Day 3 Erasure Analysis

Erasure analysis was performed for the Day 3 components by Measurement Incorporated on all operational multiple-choice responses once all scanning, data correction, and multiple-choice scoring was completed. Data for each student multiple-choice response was programmatically analyzed to determine if the response contained a mark that exceeded the mark threshold and if the lighter marks were potential erasures. Statistics were captured and aggregated at a school and district level to determine whether the school/district data was outside the state norm. Final results were provided to OEAA for review and analysis.

A program processed a JPEG grayscale image and assigned a Hex value for each multiple-choice bubble. The Hex range was 0 – 15; where Hex 0 was the lightest and represented no shading contained in the bubble, and Hex 15 was the darkest and represented a dark, filled bubble. A student selected response was captured when the Hex value for the bubble was Hex 12 (definite mark threshold) or above. A bubble detected in the range of 9 – 11 was captured as the student response, if no other bubble

for the multiple-choice question was above an 8. A bubble was considered an erasure if the Hex value for the bubble was greater than 5 and less than 12 and not identified as the student response. The following diagram demonstrates the student response and erasure identification process.

Figure 6.1. Mark identification examples for Day 3 erasure analysis.

Mark Identification Examples

	Minimum Hex value	Maximum Hex Value					Examples		
Valid Student Selected Response	9	15		A	B	C	D	Mark	Erasure
			Hex	15	2	3	4	A	No
			Hex	3	11	3	3	B	No
			Hex	9	2	2	4	A	No
			Hex	3	3	1	13	D	No
Multiple Student Selected Response				A	B	C	D	Mark	Erasure
			Hex	2	9	9	3	BC/Multiple	No
			Hex	1	13	2	15	BD/Multiple	No
			Hex	9	1	10	3	AC/Multiple	No
			Hex	2	11	13	3	BC/Multiple	No
Student Selected Response with Erasures				A	B	C	D	Mark	Erasure
			Hex	9	13	3	3	B	A
			Hex	3	6	3	9	D	B
			Hex	7	11	13	3	BC/Multiple	A
			Hex	1	13	7	6	B	CD
			Hex	1	1	14	6	C	D

The answer key for each test was used to compare the student selected response, the correct answer, and the erased bubble to determine multiple-choice erasure results. There were three results for an erased multiple-choice question: wrong answer to correct answer; correct answer to wrong answer; or wrong answer to wrong answer. A result flag was set for each erased multiple-choice case.

Using the image processed Hex value for each bubble in a multiple-choice question, each Hex value was analyzed to determine if an erasure was present. A flag was set for each bubble that was detected as an

erasure. The iErasureA flag was set if the A bubble was erased; iErasureB was set if the B bubble was erased and so on.

Tabulated Data Format

All multiple-choice erasure information was tabulated for aggregation into various result sets. The tabulated data was stored in the following format:

Table 6.3. Day 3 Erasure Analysis Data Format

Data Field	Type	Description
District Number	5 Byte Text	Unique numeric district identifier (e.g., 73903)
School Number	5 Byte Text	Unique numeric school identifier (e.g., 08294)
Grade	2 Byte Text	Numeric grade value; padded with 0 (e.g., 08)
Subject	1 Byte Text	(M) for math, (S) for science,(X) for social studies
Class Group Number	4 Byte Text	Captured from the answer document
Student Litho	8 Byte Text	Unique student document identifier
Item Position	2 Byte Text	Item position within form (e.g., 01, 02)
Erasure A	Bit	Indicates bubble contains an erasure
Erasure B	Bit	Indicates bubble contains an erasure
Erasure C	Bit	Indicates bubble contains an erasure
Erasure D	Bit	Indicates bubble contains an erasure
Wrong to Right	Bit	Indicates response changed from wrong to right
Right to Wrong	Bit	Indicates response changed from right to wrong
Wrong to Wrong	Bit	Indicates response changed from wrong to wrong

Determining the erasure results was a two-step process. The first step was to analyze the tabulated wrong-to-right data and calculate the state average and standard deviation for each subject at each grade level. The second step was to identify student tests containing wrong-to-right erasures that exceeded the state average by more than four standard deviations.

Student Data File

A data file was generated containing only students exceeding the four standard deviations criterion. The file was formatted in the following layout:

Table 6.4. Day 3 Student Data File Layout for Erasure Analysis

Data Field	Type	Description
District Number	5 Byte Text	Unique numeric district identifier (e.g., 73903)
School Number	5 Byte Text	Unique numeric school identifier (e.g., 08294)
Grade	2 Byte Text	Numeric grade value; padded with 0 (e.g., 08)
Subject	1 Byte Text	(M) for math, (S) for science,(X) for social studies
Class Group Number	4 Byte Text	Captured from the answer document
Student Barcode	10 Byte Text	Captured from the answer document
Student Lithocode	8 Byte Text	Unique student document identifier
Total Erasures	Integer	Number of erasures on the student test
Total Wrong-to-Right	Integer	Number of responses that had a correct response and one or more erasures

Psychometric Data File

A data file was generated containing a list of all students. The file was formatted in the following layout:

Table 6.5. Day 3 Psychometric Data File Layout for Erasure Analysis

Data Field	Type	Description
District Number	5 Byte Text	Unique numeric district identifier (e.g., 73903)
School Number	5 Byte Text	Unique numeric school identifier (e.g., 08294)
Grade	2 Byte Text	Numeric grade value; padded with 0 (e.g., 08)
Subject	1 Byte Text	(M) for math, (S) for science,(X) for social studies
Class Group Number	4 Byte Text	Captured from the answer document
Student Barcode	10 Byte Text	Captured from the answer document
Student Lithocode	8 Byte Text	Unique student document identifier
Total Erasures	Integer	Number of erasures on the student test
Total Wrong-to-Right	Integer	Number of responses that had a correct response and one or more erasures

Chapter 7: ACT Writing Training and Scoring

Results of Constructed Response Scoring Procedures

The MME assessment includes the written essay component of the ACT Writing Test. The procedure for scoring ACT Writing responses is outlined below. This is the scoring process that Pearson Educational Measurement Performance Scoring Center follows.

Rangefinding

The goal of the rangefinding sessions is to identify a sufficient pool of student responses which illustrate the full range of student performance in response to the prompt, and for which consensus scores can be resolved. This pool of responses includes borderline responses—ones that do not fit neatly into one of the score levels and that, therefore, represent some of the decision-making problems that scorers may face—as well as drawing a line between two score points.

All contracted scorers are trained and qualify to score ACT Writing Test responses using the Baseline Prompt training. The Baseline prompt is chosen from a retired operational prompt that performed well in operational scoring. The Baseline prompt and training materials are selected to represent the range and types of responses scorers will see during prompt-specific operational scoring.

Papers are chosen for the Baseline Anchor from operational student responses. The Baseline Anchor Set consists of three papers at each score point, for a total of eighteen papers. Baseline training also consists of four Practice sets, each with ten papers. Contract scorers must then pass two of three ten-paper Qualification sets. Rangefinding sessions for the Baseline Anchor Set are held biennially.

In addition to training and qualifying on the Baseline Anchor set, contracted scorers undergo prompt-specific training on each operational prompt. Prompt-specific training consists of an Anchor Set of nine papers and a Practice Set of four papers.

Rangefinding sessions for prompt-specific training sets are held annually in a separate session from Baseline Training rangefinding.

Prior to all Baseline and Prompt-Specific rangefinding sessions, the Contractor compiles rangefinding papers for prompts into proposed training sets, including writing annotations for all Anchor and Practice papers. ACT staff attends all rangefinding sessions and has final approval of scores assigned to all rangefinding papers.

Rater Training

Thorough training is vital to the consistent application of the scoring rubric and, therefore, accurate scores. The primary goal of training is to convey to the contract scorers the decisions made during training paper selection about what type(s) of responses correspond to each score point and to help scorers internalize the scoring protocol so that they may effectively apply those decisions.

Scorers are better able to comprehend the scoring guidelines in context, so the rubric is presented in conjunction with the anchor papers. Anchor papers are the primary points of reference for scorers as they internalize the rubric. Trainers draw scorers' attention to the score point description from the rubric,

as well as the illustrative anchor papers encouraging scorers to immediately connect the language of the rubric with actual student performance. Each anchor paper is also annotated with a scoring explanation that describes why the paper earned the given score. Annotations are meant to further illustrate the connection between the rubric descriptors and the elements present in a given essay.

After presentation and discussion of the anchor papers, each scorer is shown a practice set. Practice papers represent each score point and are used during training to help scorers become familiar with applying the rubric. Some papers clearly represent the score point. Others are selected because they represent borderline responses. Use of these practice sets provides guidance to scorers in defining the line between score points.

Training is a continuous process, and scorers are consistently given feedback as they score. With the help of the reliability reports, the scoring lead staff can closely monitor each scorer's performance.

Scoring

All responses are blind-scored by two scorers using a 6-point holistic scale. If the scores between the two scorers differ by more than 1, the paper is routed for resolution scoring. The resolution scorer will assign a holistic score using a scale of 1-6 inclusive of 0.5, representing adjacent scores. Resolution scoring is non-blind.

Comment Codes

Essay comments, derived from the scoring rubric, are selected by contract scorers to help student writers understand the strengths and weaknesses of their essays. ACT has developed five comment codes per each whole- and half-point score points. During operational scoring, one of the two contract scorers and the resolution scorer, if resolution is required, must assign at least one and not more than four comment codes to each response. Comment code training occurs on the Baseline prompt.

Contractor will identify validity responses for ACT approval.

Rater Monitoring

Pearson Educational Measurement (the contractor) is responsible for the management and overall monitoring of the operational rangefinding and scoring, but ACT has ongoing access to performance reports.

Rater Validity Checks

An additional set of data, known as validity scoring, are collected daily to check for reader drift and reader consistency in scoring to the established criteria. When scoring supervisors identify ideal student responses, they route these to the scoring directors for preview. Scoring directors review the responses and choose appropriate papers for validity scoring. Validity responses are usually solid score point responses. ACT approves all validity responses and has access to ongoing calibration responses and annotations.

Readers score a validity response approximately every 30 responses for ACT Writing. Validity scoring is blind; because image based scoring is seamless, scorers do not know when they are scoring a validity response. Results of validity scoring are analyzed regularly by scoring directors. The contractor

provides scorers who perform below the standard validity percentage with constructive feedback, close monitoring, and/or recalibration in the form of calibration papers. Calibration papers are used to correct scorer drift or to illustrate differences between problematic score points for struggling scorers. Appropriate intervention measures are initiated as needed, including the retraining or releasing of scorers who continue to perform below project standards.

Inter-Rater Reliability

Inter-rater agreement is expressed in terms of exact agreement (Reader Number One's score equals Reader Number Two's score) plus adjacent agreement (+/- 1 point difference). The Contractor must obtain a cumulative inter-rater reliability ("IRR") level of 0.60 at the conclusion of each scoring window.

In addition, the Contractor must obtain a perfect plus adjacent agreement of 0.95 at the conclusion of each scoring window—that is, 5% or less of resolution scoring.

Contractor staff monitors the accuracy of scoring to maintain the agreed upon inter-rater reliability through back reading, validity, and calibration papers. The validity percentage is 3%.

Chapter 8: Model Fit

The MME Writing, Mathematics, Reading, and Science assessments were scaled and equated using PARSCALE (Muraki & Bock, 1997) and a three parameter logistic IRT/generalized partial credit model for item calibration. (The methods used for estimating examinee scores are discussed later in the chapter on scaling and equating.) The MME Social Studies assessment was scaled with the Rasch credit model using WINSTEPS.

The MME calibration runs for Writing, Mathematics, Reading and Science were conducted using PARSCALE under the generalized partial credit model for constructed response items and the three parameter logistic model for dichotomous items. Two model fit indices were used for the dichotomous and polytomous items. They are the Chi-square (χ^2) statistics provided in PARSCALE phase 2 output generated from the calibration runs, and Orlando & Thissen's (2000) S-X² statistics. To compute the Chi-square index, the number of ability groups defined was 10, which coincides with the MME item analysis practice of using 10 deciles. Tables 8.1 to 8.4 contain the item fit statistics of all MME scored items on the initial forms for the test subjects of Writing, Reading, Mathematics and Science, respectively.

To test the goodness of fit for each item, a significance level (α) of .05 was used. If the observed p-value associated with the fit indices for an item was lower than .05, the item was considered a "poorly" fitting item. The χ^2 tests of item fit are, however, extremely sensitive to sample size, which is very large for MME. Based on the S-X² statistics, approximately 22%, 12%, 33% and 32% of the scored items for MME Mathematics, Science, Reading and Writing, respectively, were found to be significant. For all subjects, the Pearson χ^2 statistics tended to be significant.

One plausible reason for the observed misfit is the degree of multidimensionality in the assessments that occurs because of the lack of state control over portions of the assessment. A consequence of multidimensionality is that it is more difficult to obtain assessment results that load heavily on the first principal component. Given more complete control over test design and development, it is possible to construct a more unidimensional test that would have better goodness of fit indices for each item."

However, this does not invalidate these measures. This simply indicates that beyond the strong overall achievement measured by the MME subject tests, there are also some minor dimensions of achievement that impact the individual item scores of individual students. That the overall dimensions (or principal components) measured by each subject assessment are very strong is demonstrated by both (1) strong Cronbach's alpha internal consistency reliabilities (a Classical Test Theory index of measurement precision of the overall dimension), and (2) strong empirical IRT-model-based reliabilities (a measure of measurement precision of the overall dimension derived from the IRT model). For these measures of reliability, see Chapter 10 where all internal consistency and empirical IRT reliabilities are reported.

In addition, Yen and Fitzpatrick (2006) indicate that item misfit is typically caused by using an underspecified psychometric model (such as the Rasch or 2-PL model when items provide differing levels of information about the principal component, or when guessing is prevalent).

Yen and Fitzpatrick (2006) describe additional causes of item misfit, including differential item functioning, small sample sizes, poorly estimated item parameters, item stem quality, item miskeys, and item distractor quality. All of these potential causes were carefully investigated and rectified through both ACT and Michigan processes. Therefore, we are confident that these are not contributing factors in the fit statistics presented above.

Given that other possible sources of item misfit have been carefully addressed, and given that the Generalized Partial Credit Model is the most highly specified psychometric model that has been validated for use in large-scale assessment, the use of that model for MME is the best possible choice available to increase item fit.

Finally, the matrix plots of item characteristic curves resulting from PARSCALE calibration runs are presented in Figures 8.1 to 8.4. In these plots, there are some item characteristic curves (ICCs) that have flat ICCs.

For MME Social Studies, the mean square fit (MNSQ) statistics obtained from WINSTEPS were used to determine whether items were functioning in a way that is congruent with the assumptions of the Rasch mathematical model. Two types of MNSQ values are presented, OUTFIT and INFIT. MNSQ OUTFIT values are sensitive to outlying observations. MNSQ INFIT values are sensitive to behaviors that affect students' performance on items near their ability estimates. According to the item analysis specification, the model is considered to be moderately misfit if the values are between 1.5 and 2.0, and highly misfit if the values are greater than 2.0. These fit indices are presented in Table 8.5. Based on the MNSQ INFIT statistics, zero percent of items was flagged as moderately or highly misfit. Based on MNSQ OUTFIT statistics, about 6 percent of the items were considered to be moderately misfit but none was highly misfit.

This chapter examines individual item fit; the subsequent chapters examine the functioning of the MME assessments as a whole.

Table 8.1. Item Fit Statistics – Writing for Spring 2009

ITEM	SX2	df_SX2	p_SX2	X2	df	p
AE01	54.02	59	0.66	116.00	10	0.00
AE03	55.55	54	0.42	100.78	9	0.00
AE04	140.20	54	0.00	28.93	9	0.00
AE05	81.82	58	0.02	42.02	9	0.00
AE06	61.37	58	0.36	87.07	9	0.00
AE07	62.74	60	0.38	110.74	10	0.00
AE09	61.35	57	0.32	141.99	9	0.00
AE12	61.54	62	0.49	65.02	10	0.00
AE14	85.45	62	0.03	74.67	10	0.00
AE15	58.05	59	0.51	33.87	10	0.00
AE18	65.89	61	0.31	46.10	10	0.00
AE22	70.61	62	0.21	47.11	10	0.00
AE24	45.71	59	0.90	56.31	10	0.00
AE25	63.29	60	0.36	143.07	10	0.00
AE28	64.46	58	0.26	20.17	10	0.03
AE30	82.55	61	0.03	44.19	10	0.00
AE33	62.73	53	0.17	52.95	8	0.00
AE34	65.57	61	0.32	84.02	10	0.00
AE36	74.13	57	0.06	81.14	9	0.00
AE37	88.72	58	0.01	16.60	9	0.06
AE38	96.37	61	0.00	20.80	10	0.02
AE39	55.12	56	0.51	70.76	9	0.00
AE41	59.05	56	0.36	150.35	9	0.00
AE42	33.21	54	0.99	115.67	9	0.00
AE43	74.76	56	0.05	91.82	9	0.00
AE45	86.82	58	0.01	43.14	10	0.00
AE49	84.88	58	0.01	31.96	10	0.00
AE50	76.57	60	0.07	41.63	10	0.00
AE51	72.76	56	0.07	24.35	9	0.00
AE54	74.30	61	0.12	10.79	10	0.37
AE59	52.28	57	0.65	130.34	9	0.00
AE60	52.96	53	0.48	278.83	8	0.00
AE62	53.75	57	0.60	145.25	9	0.00
AE63	91.89	59	0.00	281.74	10	0.00
AE64	75.66	56	0.04	228.37	9	0.00
AE68	49.32	61	0.86	80.64	10	0.00
AE73	87.04	58	0.01	217.56	10	0.00
AE75	76.76	60	0.07	175.07	10	0.00
AW01	139.89	151	0.73	240.65	40	0.00
AW02	120.32	150	0.96	211.98	40	0.00

Table 8.2. Item Fit Statistics – Reading for Spring 2009

ITEM	SX2	df_SX2	p_SX2	X2	df	p
AR01	52.27	46	0.24	108.49	10	0.00
AR03	30.91	43	0.92	416.97	10	0.00
AR04	30.52	41	0.88	738.06	10	0.00
AR05	63.51	46	0.04	296.27	10	0.00
AR06	35.54	45	0.84	650.59	10	0.00
AR07	45.85	44	0.40	539.92	10	0.00
AR08	61.92	43	0.03	832.74	10	0.00
AR10	55.81	44	0.11	197.93	10	0.00
AR11	46.87	44	0.36	278.64	10	0.00
AR12	41.47	45	0.62	437.55	10	0.00
AR13	44.36	42	0.37	254.95	10	0.00
AR14	70.25	44	0.01	358.25	10	0.00
AR16	50.44	44	0.23	183.12	10	0.00
AR17	47.10	44	0.35	605.61	10	0.00
AR18	50.50	44	0.23	666.20	10	0.00
AR19	60.33	46	0.08	463.13	10	0.00
AR20	47.99	43	0.28	369.52	10	0.00
AR21	29.63	41	0.91	590.11	10	0.00
AR22	57.95	45	0.09	275.64	10	0.00
AR23	52.59	45	0.20	642.24	10	0.00
AR25	43.51	43	0.45	1183.39	10	0.00
AR27	57.21	45	0.10	559.46	10	0.00
AR28	65.61	45	0.02	1061.66	10	0.00
AR30	32.42	46	0.94	145.91	10	0.00
AR31	40.55	45	0.66	569.98	10	0.00
AR32	100.76	43	0.00	2421.06	9	0.00
AR33	61.46	45	0.05	412.20	10	0.00
AR34	105.45	43	0.00	1655.60	9	0.00
AR35	41.25	45	0.63	858.91	10	0.00
AR36	77.18	44	0.00	1822.54	9	0.00
AR37	119.69	43	0.00	2567.12	9	0.00
AR38	115.13	43	0.00	1731.77	9	0.00
AR39	81.20	44	0.00	1419.20	10	0.00
AR40	63.27	44	0.03	940.37	10	0.00
WK02	32.13	29	0.31	76.57	8	0.00
WK03	18.80	27	0.88	96.66	8	0.00
WK05	99.80	33	0.00	2398.73	9	0.00
WK07	101.82	39	0.00	166.94	10	0.00
WK09	42.74	42	0.44	351.80	10	0.00
WK11	29.93	30	0.47	168.29	9	0.00
WK15	245.79	42	0.00	1568.47	10	0.00
WK16	64.19	41	0.01	2001.68	10	0.00
WK18	36.22	40	0.64	564.08	10	0.00
WK20	37.52	46	0.81	213.96	10	0.00
WK21	45.19	44	0.42	711.51	10	0.00

WK24	43.32	45	0.54	311.80	10	0.00
WK28	51.76	46	0.26	370.39	10	0.00
WK29	51.02	45	0.25	577.21	10	0.00
WK30	23.47	45	1.00	417.39	10	0.00

Table 8.3. Item Fit Statistics – Mathematics for Spring 2009

ITEM	SX2	df_SX2	p_SX2	X2	df	p
AM01	64.08	71	0.71	431.50	10	0.00
AM02	69.78	63	0.26	344.13	9	0.00
AM03	65.44	71	0.66	493.16	10	0.00
AM04	107.25	69	0.00	3277.15	9	0.00
AM05	105.27	72	0.01	1304.29	10	0.00
AM06	99.3	72	0.02	1930.07	10	0.00
AM10	65.23	72	0.70	2876.79	9	0.00
AM12	73.35	76	0.56	783.77	10	0.00
AM13	53.73	67	0.88	3092.00	9	0.00
AM17	82.09	78	0.35	828.95	10	0.00
AM21	68.28	76	0.72	822.70	10	0.00
AM22	48.55	72	0.98	3673.04	9	0.00
AM23	100.68	75	0.03	3583.18	9	0.00
AM25	91.8	74	0.08	816.86	10	0.00
AM26	106.39	71	0.00	2090.97	10	0.00
AM27	91.16	78	0.15	850.06	10	0.00
AM28	73.02	75	0.54	3308.26	9	0.00
AM29	84.51	72	0.15	2630.90	10	0.00
AM31	63.52	75	0.82	1192.08	10	0.00
AM35	61.15	75	0.88	1579.79	10	0.00
AM37	79.71	79	0.46	2104.22	10	0.00
AM39	75.36	75	0.47	2511.65	10	0.00
AM42	84.23	77	0.27	3520.97	10	0.00
AM47	103.98	76	0.02	4082.26	10	0.00
AM49	92.11	73	0.06	3369.58	10	0.00
AM51	76.55	79	0.56	3010.06	10	0.00
AM52	110.62	75	0.00	3402.64	10	0.00
AM54	82.54	77	0.31	2051.78	10	0.00
AM57	65.03	77	0.83	4371.25	10	0.00
AM59	88.57	77	0.17	1025.83	10	0.00
WM01	98.67	56	0.00	1789.63	10	0.00
WM04	16.18	48	1.00	84.69	7	0.00
WM08	160.06	58	0.00	1474.42	8	0.00
WM12	48.68	55	0.71	233.24	9	0.00
WM15	352.72	64	0.00	9183.98	8	0.00
WM16	109.44	66	0.00	1771.82	9	0.00
WM19	92.64	68	0.03	3563.20	8	0.00
WM21	52.43	68	0.92	1927.70	9	0.00
WM23	56.9	70	0.87	1535.47	9	0.00
WM24	68.34	70	0.53	1924.31	9	0.00
WM25	84.49	80	0.34	1776.16	10	0.00
WM26	81.33	76	0.32	1583.45	10	0.00
WM27	107.94	77	0.01	8352.48	9	0.00
WM28	83.91	77	0.28	2936.25	10	0.00
WM29	184.82	76	0.00	9759.35	9	0.00

WL01	25.42	53	1.00	69.54	9	0.00
WL02	71.94	75	0.58	249.22	10	0.00
WL03	39.16	67	1.00	221.26	10	0.00
WL04	60.36	72	0.83	581.72	10	0.00
WL05	56.96	73	0.92	243.28	10	0.00
WL06	48.73	71	0.98	1050.58	9	0.00
WL07	82.73	76	0.28	1856.45	10	0.00
WL08	63.48	73	0.78	2189.92	10	0.00
MI01	78.17	73	0.32	999.88	10	0.00
MI02	66.4	73	0.69	1830.38	10	0.00
MI03	61.39	73	0.83	1079.95	9	0.00
MI04	58.05	74	0.91	2775.87	10	0.00
MI05	76.84	75	0.42	1110.16	10	0.00
MI06	56.68	73	0.92	1973.63	10	0.00
MI07	81.57	77	0.34	271.22	10	0.00
MI08	59.68	73	0.87	52.45	9	0.00
MI09	544.75	73	0.00	3260.32	10	0.00
MI10	87.64	77	0.19	3037.35	10	0.00
MI11	62.61	72	0.78	1432.93	9	0.00
MI12	96.06	76	0.06	1735.54	10	0.00
MI13	96.28	77	0.07	535.98	9	0.00
MI14	77.58	78	0.49	1236.00	9	0.00
MI15	62.28	71	0.76	60.81	9	0.00
MI16	48.05	69	0.97	258.90	8	0.00
MI17	46.79	69	0.98	1142.62	9	0.00
MI18	97.33	80	0.09	91.49	10	0.00
MI19	56.01	73	0.93	45.62	9	0.00
MI20	299.16	80	0.00	240.31	10	0.00

Table 8.4. Item Fit Statistics – Science for Spring 2009

ITEM	SX2	df_SX2	p_SX2	X2	df	p
AS01	26.50	50	1.00	436.68	10	0.00
AS02	63.07	51	0.12	1491.06	10	0.00
AS05	48.72	55	0.71	354.69	10	0.00
AS07	47.74	52	0.64	1174.74	10	0.00
AS08	46.66	57	0.83	411.28	10	0.00
AS09	95.18	58	0.00	65.92	10	0.00
AS10	51.25	57	0.69	1200.53	10	0.00
AS12	58.66	53	0.28	482.98	10	0.00
AS13	58.90	56	0.37	198.00	10	0.00
AS14	49.87	54	0.63	576.02	10	0.00
AS15	64.28	56	0.21	282.52	10	0.00
AS16	57.58	56	0.42	473.51	10	0.00
AS17	64.75	59	0.28	434.46	10	0.00
AS18	52.71	57	0.64	912.24	10	0.00
AS23	46.48	57	0.84	541.71	10	0.00
AS24	66.53	58	0.21	723.97	10	0.00
AS30	44.95	55	0.83	579.40	10	0.00
AS31	57.81	57	0.45	670.48	10	0.00
AS33	60.05	57	0.37	1289.61	10	0.00
AS38	93.94	60	0.00	926.59	10	0.00
MI01	74.36	59	0.09	111.87	10	0.00
MI02	57.92	58	0.48	111.69	10	0.00
MI04	52.53	54	0.53	588.18	10	0.00
MI05	50.63	56	0.68	54.12	10	0.00
MI07	55.05	58	0.59	197.49	10	0.00
MI08	90.08	60	0.01	77.09	10	0.00
MI10	84.73	52	0.00	815.60	10	0.00
MI11	44.29	55	0.85	146.00	10	0.00
MI13	68.91	58	0.15	305.63	10	0.00
MI14	49.64	57	0.74	106.03	10	0.00
MI16	38.13	54	0.95	832.97	10	0.00
MI17	47.79	58	0.83	83.92	10	0.00
MI19	50.41	58	0.75	882.42	10	0.00
MI20	52.71	55	0.56	118.54	10	0.00
MI22	61.41	56	0.29	357.71	10	0.00
MI23	38.02	54	0.95	120.35	9	0.00
MI25	62.04	58	0.33	322.82	10	0.00
MI26	45.13	55	0.83	171.90	10	0.00
MI28	22.04	44	1.00	349.43	9	0.00

MI29	42.37	53	0.85	120.97	9	0.00
MI31	62.68	59	0.35	166.11	10	0.00
MI32	64.15	58	0.27	390.44	10	0.00
MI34	55.74	55	0.45	396.22	10	0.00
MI35	56.10	58	0.55	240.79	10	0.00
MI37	67.35	60	0.24	163.01	10	0.00
MI38	56.15	55	0.43	68.34	10	0.00
MI40	63.10	56	0.24	465.15	10	0.00
MI41	63.85	57	0.25	72.68	10	0.00
MI43	39.17	57	0.97	636.55	10	0.00
MI44	61.53	60	0.42	33.66	10	0.00
MI46	311.32	54	0.00	1253.64	10	0.00
MI47	132.40	59	0.00	284.91	10	0.00

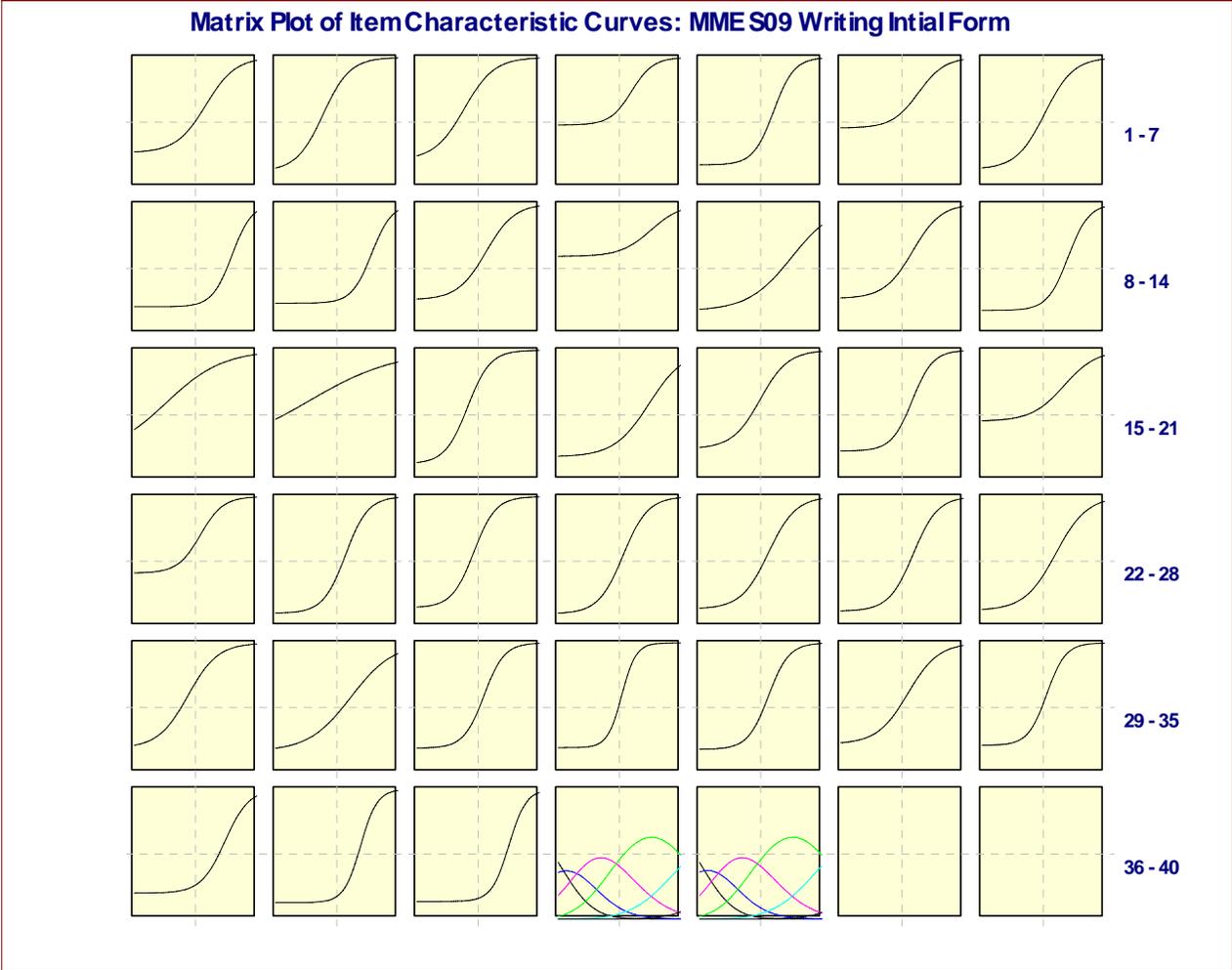


Figure 8.1. Item characteristic curves – Writing Spring 2009: 38 selected ACT English MC items plus one ACT CR item.

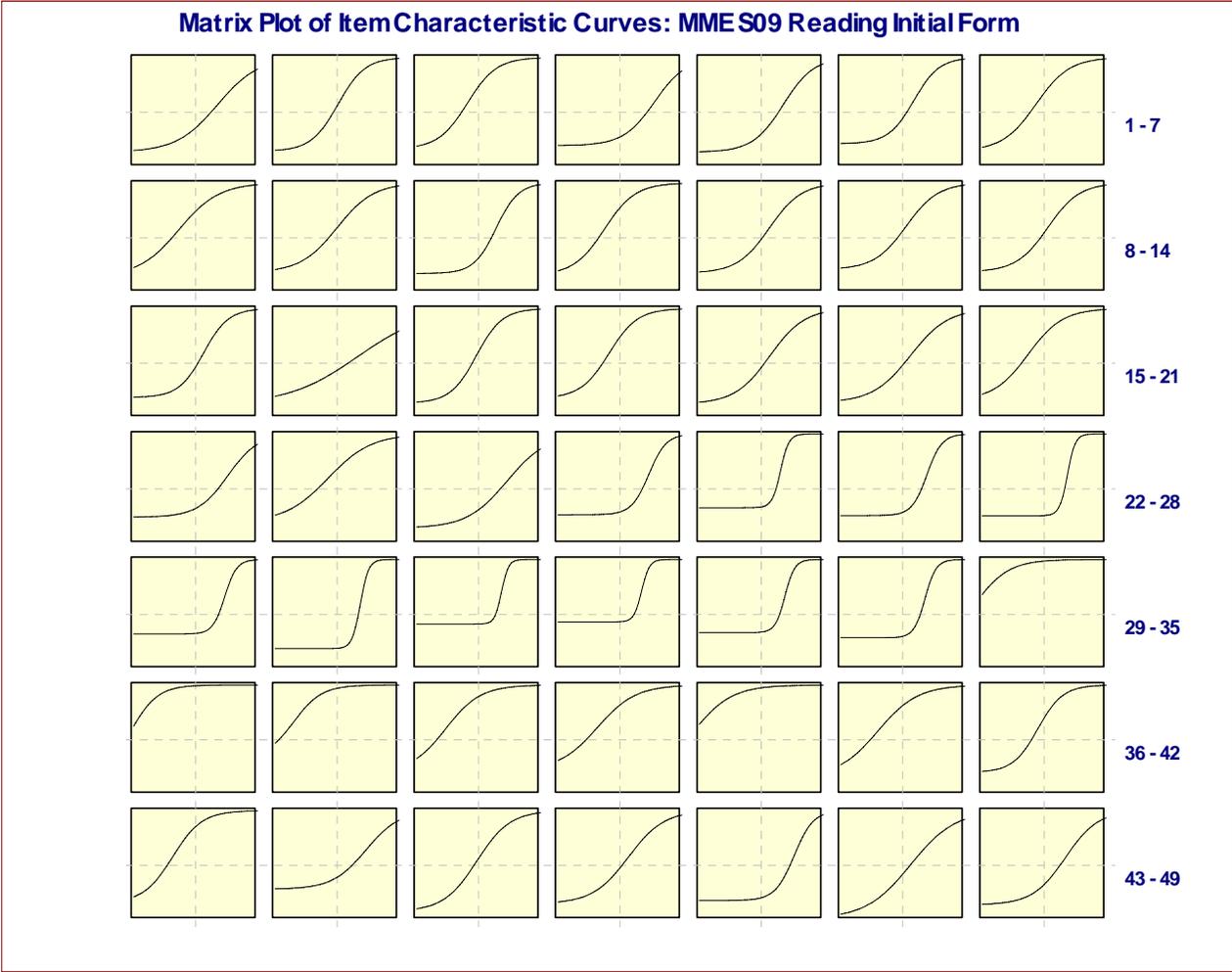


Figure 8.2. Item characteristic curves – Reading Spring 2009: 34 selected ACT Reading items plus 15 *WorkKeys Reading for Information* items.

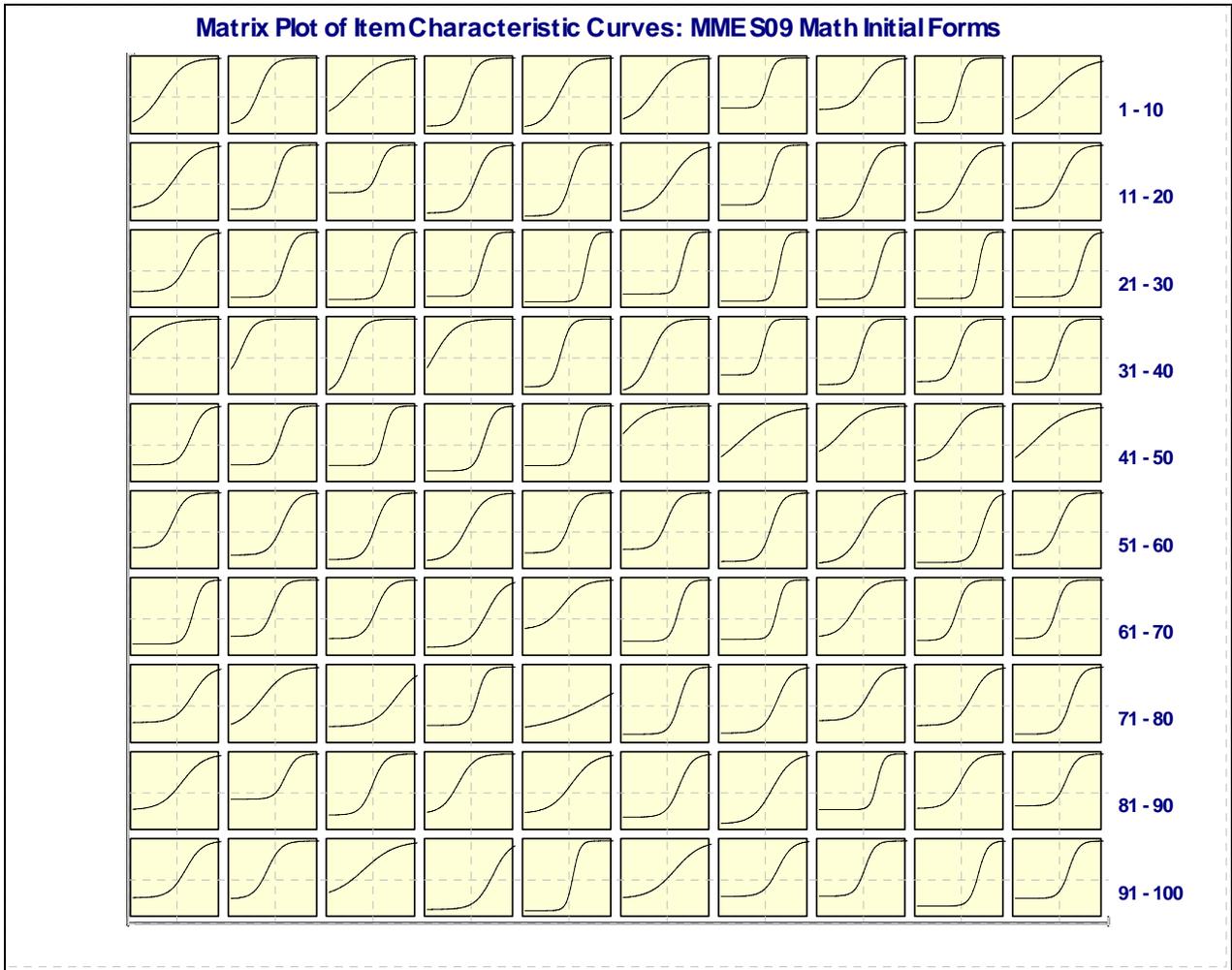


Figure 8.3. Item characteristic curves – Mathematics Spring 2009: 30 selected ACT

Mathematics items plus 15 selected *WorkKeys Applied Mathematics* items plus eight selected *WorkKeys Locating Information* items plus 75 unique Michigan-developed mathematics items.

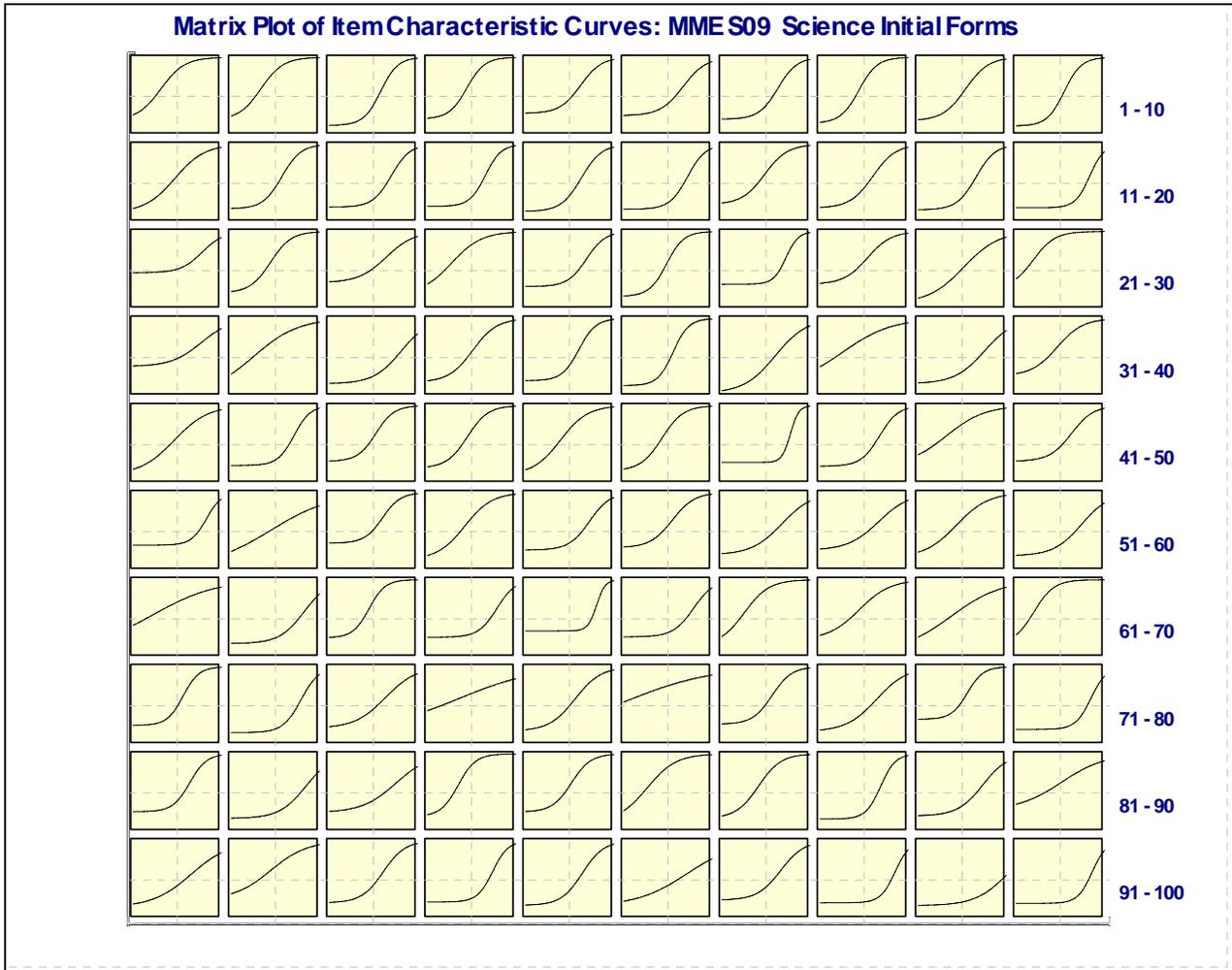


Figure 8.4. Item characteristic curves – Science Spring 2009: 20 selected ACT Science Items plus 108 unique Michigan-developed science items.

Table 8.5. Item Fit Statistics – Social Studies for Spring 2009

Item	INFIT MNSQ	OUTFIT MNSQ
WKLI01	0.99	1.05
WKLI02	1.06	1.62
WKLI03	1.13	1.38
WKLI04	0.99	1
WKLI05	1.09	1.3
WKLI06	0.97	0.94
SocS01	1.03	1.02
SocS02	1.09	1.12
SocS03	1.02	1.02
SocS04	0.95	0.85
SocS05	0.76	0.64
SocS06	0.96	0.94
SocS07	0.91	0.87
SocS08	0.79	0.72
SocS09	0.95	0.93
SocS10	1.11	1.13
SocS11	1.01	1.02
SocS12	0.88	0.8
SocS13	0.92	0.87
SocS14	1.08	1.11
SocS15	1.06	1.11
SocS16	1.24	1.32
SocS17	0.84	0.72
SocS18	0.9	0.85
SocS19	1.12	1.16
SocS20	1.03	1.04
SocS21	1.21	1.54
SocS22	0.74	0.67
SocS23	0.95	0.94
SocS24	0.99	0.99
SocS25	0.89	0.81
SocS26	1.15	1.19
SocS27	0.79	0.65
SocS28	1.05	1.06

Chapter 9: Scaling and Equating

Quality Control Protocols for MME Calibrations

The following quality control (QC) tasks were implemented for MME calibrations. For the MME test subjects of Writing, Mathematics, Reading and Science, the MME calibration runs were conducted using PARSCALE (Muraki & Bock, 1997) under the three parameter logistic model (3PLM) for dichotomously scored multiple choice (MC) items and the generalized partial credit model (GPCM) for constructed response (CR) items. For calibrating MME Social Studies, the Rasch credit model was employed.

A thorough review of the test maps for Michigan-developed tests and *WorkKeys* was conducted including the following activities:

- Cross-checks on fields/variables regarding items (such as item code and item key) provided on the test map.
- Cross-reference of test positions for scrambled versions.
- Checks on field test items (e.g., test positions, same field test items occurring on multiple forms).

Each updated test map for Michigan-developed tests provided on the Measurement Inc. /ACT ftp site was reviewed.

The linking items were also reviewed and verified. Specifically, based on the information regarding linking items from the test maps, the new and old test booklets were compared word by word to ensure that there were no differences in linking items from one form to the next.

Files containing the item parameter estimates of ACT, *WorkKeys*, and Michigan linking items were prepared for review. The file naming conventions for such files were developed in advance. The values of the item parameter estimates and the test positions on the new and old forms were checked by test subject and form.

To facilitate creation of the PARSCALE and WINSTEPS control files, the 0/1 score data layout was created in advance. The positions for the 0/1 scores in the calibration data files were double-checked.

As a preliminary check on the calibration data file, SAS analyses were implemented to produce N-counts, classical item statistics, as well as frequency distributions on form codes, total raw scores, and scores for CR items. These analyses were examined for strange results, outliers, and so forth.

To review the calibration results, the following tasks were implemented:

- Check convergence for each calibration run.
- Compare classical item statistics produced by PARSCALE runs with those produced from SAS calculations, for an exact match.
- Check the discrimination parameter estimates. There should be no negative values.
- Compute correlation coefficients between p-value and b parameter estimates for reasonableness. The p-values and b parameter estimates should be negatively correlated. Examine the scatter plot of p-values versus b parameter estimates for outliers.
- Check c parameter estimates for unusually large values, with the understanding that c-parameters interact with a- and b-parameters such that there may be some well-performing items with relatively large c-parameters where the empirical ICCs match the parameterized ICC well.

- Review ICC plots produced by PARSCALE.
- Check that fixed item parameter estimates have the correct values.
- Compare p-values for ACT items with those from the history to check that they look reasonably similar.
- Compare p-values for *WorkKeys* linking items with those from the history to check that they look reasonably similar.
- Compare p-values for Michigan linking items with those from the history to check that they look reasonably similar.

For constructed response items, compare the item parameter estimates for the two raters to check that they look reasonable. Because the raters are randomly assigned, no difficulty, discrimination, or step parameters should differ by more than 0.01 across raters.

Equating for ACT

Several new forms of each of the ACT tests are developed each year. Even though each form is constructed to adhere to the same content and statistical specifications, the forms may differ slightly in difficulty. To control for these differences, subsequent forms are equated, and the scores reported to examinees are scale scores that have the same meaning regardless of the particular form administered to examinees. Thus, scale scores are comparable across test forms and test dates.

A carefully selected sample of examinees from one of the five national test dates each year is used as an equating sample. The examinees in this sample are administered a spiraled set of “n” forms—the new forms (“n – 1” of them) and one anchor form that has already been equated to previous forms. (The base form is the form used initially to establish the score scale.) The use of randomly equivalent groups is an important feature of the equating procedure and provides a basis for confidence in the continuity of scales. More than 2,000 examinees take each form.

Scores on the new forms are equated to the score scale using an equipercentile equating methodology. In equipercentile equating, a score on Form X of a test and a score on Form Y are considered to be equivalent if they have the same percentile rank in a given group of examinees. The equipercentile equating results are subsequently smoothed using an analytic method described by Kolen (1984) to establish a smooth curve, and the equivalents are rounded to integers. The conversion tables that result from this process are used to transform raw scores on the new forms to scale scores on the base form scale.

The equipercentile equating technique is applied to the raw scores of each of the four tests for each form separately. The composite score is not directly equated across forms. It is, instead, a rounded arithmetic average of the scale scores for the four equated tests. The subscores are also separately equated using the equipercentile method. Note, in particular, that the equating procedure does *not* lead to a given reported test score being equal to some prespecified arithmetic combination of subscores. As specified in the *Standards for Educational and Psychological Testing* (AERA et al., 1999), ACT conducts periodic checks on the stability of the ACT scores. The results appear reasonably stable to date.

Equating for *WorkKeys*

New forms of the *WorkKeys* tests are developed as needed. Though each form is constructed to adhere to the same content and statistical specifications, the forms may be slightly different in difficulty. To control for these differences, scores on all forms are equated so that when they are reported to test takers (as either

Level Scores or Scale Scores), equated scores have the same meaning regardless of the particular form administered. Thus, Level Scores and Scale Scores are comparable across test forms and test dates. However, they are not comparable across tests. For example, a Level Score of 3 or a Scale Score of 73 in *Reading for Information* does not have the same meaning as a Level Score of 3 or a Scale Score of 73 on any other *WorkKeys* test (e.g., *Applied Mathematics*). Two common equating designs are used with the *WorkKeys* tests (Kolen & Brennan, 2004).

In a randomly equivalent groups design, new test forms are administered along with an anchor form that has already been equated to previous forms. A spiraling process is used to distribute test forms to test takers. For example, in each testing room the first person receives Form 1, the next Form 2, and the next Form 3. This pattern is repeated so that each form is given to one-third of the test takers and the forms are given to randomly equivalent groups. When this design is used, the difference in total-group performance on the new and anchor forms is considered a direct indication of the difference in difficulty between the forms. Scores on the new forms are placed to the score scale using various equating methodologies including linear and equipercentile procedures (e.g., see Kolen & Brennan, 2004). When the Level Score and Scale Score conversions are chosen for each form, the equating functions are examined, as are the resulting distributions of the scores and their means, standard deviations, skewnesses, and kurtoses.

A common-item nonequivalent groups design has been used when a spiraling technique cannot be implemented in a test administration, when only a single form can be administered per test date, or when some items are changed in a revised form. In a common-item nonequivalent groups design, the new form and base form have a set of items in common. These common item sets (anchors) are chosen to represent the content and statistical characteristics of the test and are usually interspersed among the other items in the new test form. The different forms are then administered to different groups of test takers. In this design, the groups are not assumed to be equivalent. Observed differences of performances between groups can result from a combination of (a) test-taker group ability differences and (b) test form difficulty differences. The common items are used to control for group differences, so that adjustments can be made for form differences. Strong statistical assumptions are required to separate these group and form differences.

The various equating methods under the common-item nonequivalent groups design are distinguished in terms of their statistical assumptions (Kolen & Brennan, 2004). Observed-score equating methods are typically used in equating *WorkKeys* test forms. For each form, the equating functions are examined, as are the resulting distributions of scale scores and the mean, standard deviation, skewness, and kurtosis of the scale scores. The set of equating conversions chosen for each form is the one that results in scale score distributions and scale score moments that are judged to be reasonable based on the sample sizes, the magnitudes of the form differences and group differences, and the historical statistics for the test.

Equating for MME Social Studies

Social Studies is the only MME subject using the Rasch model to derive MME scale scores. The model provides a one-to-one relationship between the derived (i.e., scale) and the raw scores. The item calibration and proficiency estimates are obtained using the Rasch model and procedures implemented in WINSTEPS version 3.63. The statistical elements of the calibration/scaling process are referred to as Rasch Calibration/Scaling as described in the WINSTEPS manual.

In spring 2009, the MME Social Studies included selected *WorkKeys Locating Information* items. These items were calibrated concurrently with other MME Social Studies items. The item scores for selected *WorkKeys Locating Information* items and MME Social Studies items were summed to obtain a MME

Social Studies raw score. The MME Social Studies raw scores were then converted to MME Social Studies scale scores.

Following calibration, operational items are “fixed” when the field test items are calibrated. Each year, new test forms are built based on the test blueprint and available statistical information obtained from previous field testing. New field test items are embedded in test forms for building and replenishing the item pool. These forms are spiraled in the administration. This procedure puts field test item parameters on the scale of the operational items.

Specific Steps for Equating of Social Studies:

- Review test maps and obtain item parameters from the MME item pool for anchored items
- Create data sets for item calibration and equating
- Check the parameter stability of anchored items
- Run operational item calibration with fixed anchored items using WINSTEPS (version 3.63)
- Review calibration results
- Create a raw-to-scale score conversion table for scoring
- Run field test item calibration using WINSTEPS
- Review field test item calibration results for future form construction and linking

Equating for MME Writing, Reading, Mathematics, Science

Depending on the MME test subject (Writing, Reading, Mathematics and Science), an MME test can consist of up to four components across three days of testing: items from the ACT tests (from Day 1), one or more of the three *WorkKeys* tests (*Reading for Information*, *Applied Mathematics* or *Locating Information*, from Day 2), and Michigan-developed tests (Mathematics, Science, or Social Studies, from Day 3). To develop the MME scale, an MME base form was administered in the spring 2006 Baseline Study. A fixed-parameter calibration approach is employed for equating MME forms, and putting new form scores on the base form scale.

The MME equating plan is exhibited in Figure 9.1. The shaded areas in Figure 9.1 indicate *WorkKeys* and Michigan-developed common items that link between forms (e.g., *WorkKeys* form W1 and W2). The common items have parameter estimates from previous MME administrations. These item parameter estimates are placed on the MME scale and fixed for equating new MME forms. For instance, as illustrated in Figure 9.1, for equating MME form 2, items that are fixed in MME calibration runs include, depending on the MME testing subject, *WorkKeys* common items with item parameters existing from MME form 1, Michigan-developed common items with item parameters existing from MME form 1, and all ACT items on form C1 which have been placed on the MME scale. The equating for MME ACT forms is discussed in the following section.

Figure for MME Linking/Equating Plan

MME Form 1	ACT Form B2		WorkKeys Form W1		Michigan Form M1	
MME Form 2		ACT Form C1		WorkKeys Form W2		Michigan Form M2

The shaded areas indicate common items between forms.

Figure 9.1. MME linking/equating plan.

The item parameter estimates for all ACT forms administered in MME are separately calibrated under the 3PLM using the ACT national samples discussed previously, and then placed on the MME scale using the Stocking-Lord characteristic curve method (Stocking & Lord, 1983). Figure 9.2 below exhibits the ACT linking studies. Within the same ACT linking study, the randomly equivalent groups design is employed to ensure that form groups are equivalent. For instance, in study 2 as shown in Figure 9.2, form B1 and forms A1 through A4 are administered to randomly equivalent groups. Across ACT equating studies, the Stocking-Lord transformation is employed. For example, study 1 and study 2 are linked through form B1, and forms A1 through A4 can then be placed on the study 1 scale accordingly.

ACT Linking Studies

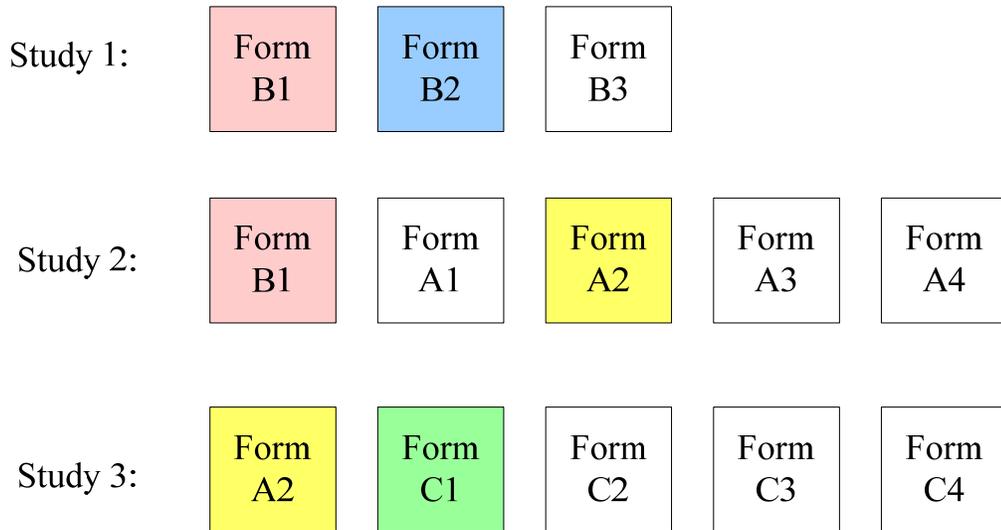


Figure 9.2. ACT linking studies.

In light of the ACT linking studies, any ACT form can be placed on the MME scale. Figure 9.3 depicts the linkage for MME ACT forms. For example, as shown in Figure 9.3, forms A3, B2 and C1 exist in the MME pool. In the ACT linking study comprising forms B1 and B2 that are administered to randomly equivalent groups, form B2 can be equated to the MME pool via the Stocking-Lord procedure. Form B1 can then be equated to the MME pool.

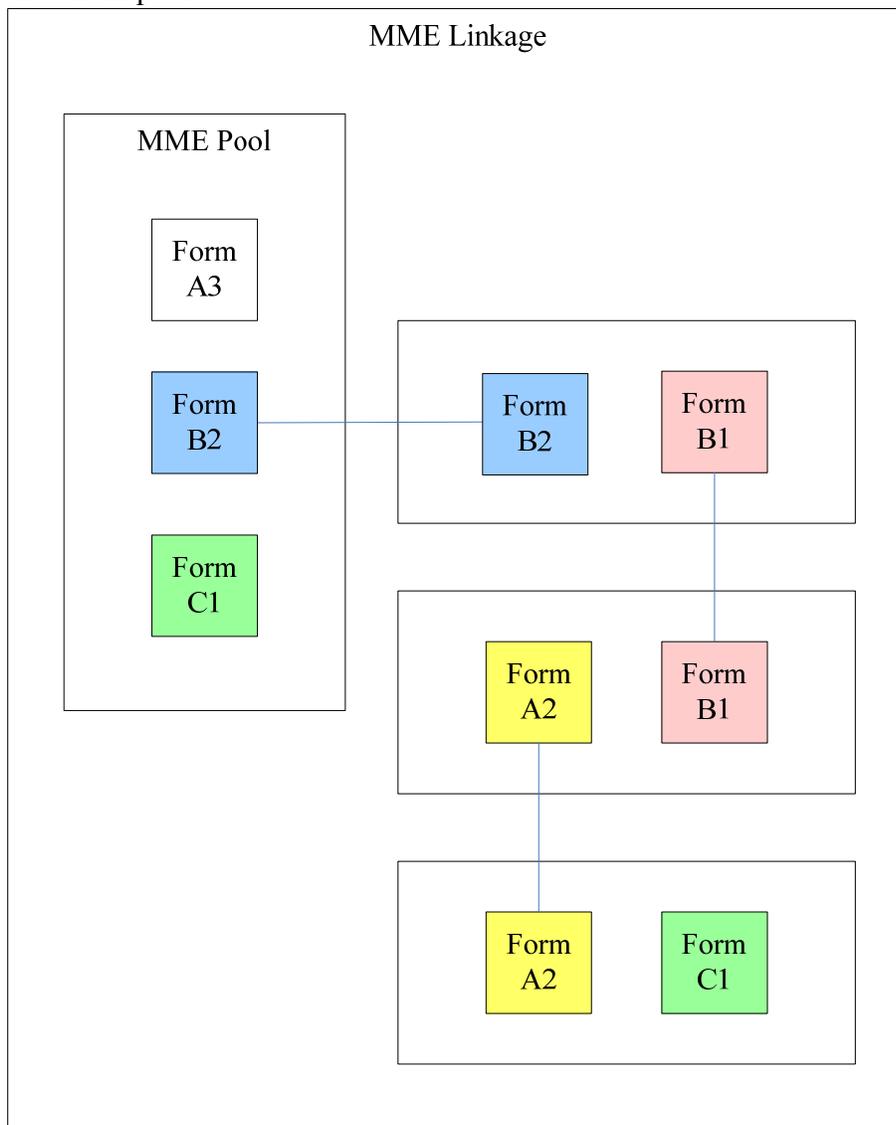


Figure 9.3. Diagram for MME linkage

To link the *WorkKeys* and Michigan-developed test forms, respectively, to the MME base form, a set of anchor items is employed and calibrated using MME sample data as shown in Figure 9.1.

For MME calibrations, the 0/1 scores of 11th grade students who meet MME attemptedness criteria were used. The MME calibration runs were conducted using PARSCALE version 4.1 (Muraki & Bock, 1997) under the GPCM for CR items and the three 3PLM for dichotomous items. These models are given as follows:

Under the GPCM, the probability that an examinee j scores z with $z = 0, 1, \dots, Z_i$ on item i with $Z_i + 1$ response categories is modeled by

$$P(z | \theta_j, \alpha_i, \beta_i, \tau_{ci}) = \frac{\exp \sum_{c=0}^z \alpha_i [\theta_j - (\beta_i - \tau_{ci})]}{\sum_{y=0}^{Z_i} \exp \sum_{c=0}^y \alpha_i [\theta_j - (\beta_i - \tau_{ci})]}$$

where α_i is the discrimination of item i , β_i denotes the difficulty of item i , and τ_{ci} represents the location parameter for a category on item i . For model identification, it needs to set $\tau_{0i} = 0$, $\sum_{c=1}^{Z_i} \tau_{ci} = 0$ and

$\exp \sum_{c=0}^0 \alpha_i [\theta_j - (\beta_i - \tau_{ci})] = 1$ in the model above (Muraki, 1992, 1996).

Under the 3PLM, the probability that an examinee j scores z with $z = 0$ or 1 on item i is modeled by

$$P(z | \theta_j, \alpha_i, \beta_i, c_i) = c_i + (1 - c_i) \frac{1}{1 + \exp[-1.7\alpha_i(\theta_j - \beta_i)]}$$

where α_i , β_i and c_i denote the discrimination, difficulty and pseudo-guessing parameter estimates, respectively.

For the MME administration, a concurrent calibration run for the various components is implemented with fixed item parameter estimates for all ACT items, fixed item parameter estimates for the *WorkKeys* anchor items, and fixed item parameter estimates for the Michigan-developed items; with all other items being placed on the MME scale by the calibration run. Michigan-developed operational items that were administered as field test items in previous MME administrations were recalibrated.

For scrambled versions of the Michigan-developed forms that are used in different testing situations, (i.e., initial, makeup and accommodated), the item parameter estimates for Michigan-developed anchor items are obtained from a master initial calibration run using the data for the initial forms for all of the various MME components. These calibration analyses are based on the assumption that the sample size for the master initial run is the largest, and the IRT assumption that item location does not affect item parameters. Under the IRT property of group invariance, these item parameters were fixed for the calibration runs for other form combinations. Also, for calibrating Michigan-developed field test items, item parameters of ACT items, *WorkKeys* items and Michigan-developed operational items were fixed. Field test items with point biserials less than .10 were excluded from the field test item calibrations as per OEAA's direction.

For MME scoring, only the selected ACT items, selected *WorkKeys* items and Michigan-developed items were employed.

Specific steps for equating MME Writing, Mathematics, Reading and Science are as follows:

1. Review test maps.
2. Obtain item parameter estimates from the pool for anchor items.
 - For forms with small N-counts (e.g., Braille or makeup), item parameter estimates obtained from master initial calibration runs are employed if available.
 - For forms that are a scrambled version of the initial form, item parameter estimates of the initial form are used.
3. Create data sets for calibration and equating.
4. Check anchor item parameter stability.
5. Conduct fixed-parameter calibration runs using PARSCALE without field test items.

6. Evaluate calibration results of operational items and pass item parameter estimates for MME scoring. (Only the selected ACT items, selected *WorkKeys* items and Michigan-developed items are employed for MME scoring.)
7. Run PARSCALE to calibrate field test items with item parameter estimates of all operational items fixed. Field test items with point biserials less than .10 are excluded from the field test item calibrations as per OEAA's direction.
8. Review calibration results of field test items for future form construction considerations and linking.

Equating for MME ELA

MME ELA is not separately equated; it is the average of two separately equated components, MME Writing and MME Reading.

Calibration Summary Reports

Calibration summary reports that discussed N-counts, calibration convergence, and ACT's suggestions for MME scoring were presented to OEAA for their review.

IRT Model Fit and Plots

Matrix plots of item characteristic curves resulting from PARSCALE calibration runs were presented to OEAA for their review. The plots of SE/information curves produced by PARSCALE with the MME cut scores imposed for the testing subjects of Writing, Reading, Mathematics and Science, respectively, were also created and presented to OEAA for their review.

For MME Social Studies, the mean square fit (MNSQ) statistics obtained from WINSTEPS are used to determine whether items were functioning in a way that is congruent with the assumptions of the Rasch mathematical model. Two types of MNSQ values are presented, OUTFIT and INFIT. MNSQ OUTFIT values are sensitive to outlying observations. MNSQ INFIT values are sensitive to behaviors that affect students' performance on items near their ability estimates. According to the item analysis specification, the model is considered to be moderately misfit if the values are between 1.5 and 2.0 and highly misfit if the values are greater than 2.0.

The MME calibration runs for Writing, Mathematics, Reading and Science are conducted using PARSCALE (Muraki & Bock, 1997) under the GPCM for CR items and the 3PLM for dichotomous items. Two model fit indices are used for the dichotomous and polytomous items. They are the Chi-square (χ^2) statistics provided in PARSCALE phase 2 output generated from the calibration runs, and Orlando and Thissen's (2000) S-X² statistics. To compute the Chi-square index, ten ability groups are used. To test the goodness of fit for each item, a significance level (α) of .05 is used. If the observed p-value associated with a fit index for an item is lower than .05, the item is considered a "poorly" fitting item. The χ^2 tests of item fit are, however, extremely sensitive to sample size, which is very large for MME. The item fit statistics are reported in Tables 8.1 through 8.4.

Item Analysis

After the MME administration, the Measurement Research Department (MRD) at ACT receives matched data files. MRD computes classical item statistics as specified by OEAA for the Michigan-developed

operational and field-test MC items and creates a SAS dataset containing these item statistics that it sends to OEAA for their review.

MRD computes IRT based item statistics as specified by OEAA for the Michigan-developed operational and field-test MC items and adds these statistics to the classical item statistics SAS dataset. MRD also adds the item parameters to this file. It then sends this combined item analysis file to OEAA for their review.

Theta Generation

ACT developed an ACT-written program, SCOREST, to compute MME thetas (θ s), and uses independent checks on the thetas. The SCOREST program is written in C++ and developed by the Corporate Assessment Infrastructure (CAI) team in conjunction with IT. For the purpose of independent checks on MME theta scores, MRD staff developed two FORTRAN programs for estimating θ s: MULTEST for multiple choice data and MIXEDEST for mixed format data. These programs produce theta estimates and standard errors of (SE) theta.

IRT Models

Two IRT models are employed in the scoring programs: the 3PLM and the GPCM. The 3PLM for dichotomous MC items with $z = 0$ or 1 is given as follows:

$$P(z | \theta_j, \alpha_i, \beta_i, c_i) = c_i + (1 - c_i) \frac{1}{1 + \exp[-D\alpha_i(\theta_j - \beta_i)]},$$

where $D = 1.7$, α_i is the discrimination of item i , β_i denotes the difficulty of item i , and c_i is the pseudo-guessing parameter of item i . Under the GPCM, the probability that an examinee j scores z with $z = 0, 1, \dots, Z_i$ on item i with $Z_i + 1$ response categories is modeled by the following:

$$P(z | \theta_j, \alpha_i, \beta_i, \tau_{ci}) = \frac{\exp \sum_{c=0}^z D\alpha_i [\theta_j - (\beta_i - \tau_{ci})]}{\sum_{y=0}^{Z_i} \exp \sum_{c=0}^y D\alpha_i [\theta_j - (\beta_i - \tau_{ci})]},$$

where α_i is the discrimination of item i , β_i denotes the difficulty of item i , and τ_{ci} represents the location parameter for a category on item i . For model identification, set $\tau_{0i} = 0$, $\sum_{c=1}^{Z_i} \tau_{ci} = 0$ and

$\exp \sum_{c=0}^0 \alpha_i [\theta_j - (\beta_i - \tau_{ci})] = 1$. For both 3PLM and GPCM, item parameter estimates are computed using PARSCALE.

Algorithms for the Scoring Programs

The MULTEST and MIXEDEST programs use the grid search algorithm to estimate the maximum likelihood estimate (MLE) of θ values. Under the grid search algorithm, the θ space ranging from -6.0 to $+6.0$ is divided into grids with a width of $.001$, and magnitudes of the log-likelihood are computed for all grid points under the appropriate IRT model(s). The theta score with the highest log-likelihood value is selected and denoted $\hat{\theta}^*$. A finer search with a grid width of $.0001$ is then conducted in the neighborhood of $\hat{\theta}^*$. The MLE theta score $\hat{\theta}$ is then given by the theta score that yields the highest log-likelihood value in the

finer search, and SE for $\hat{\theta}$ is computed accordingly. One advantage of the grid search algorithm is that the non-convergence for cases with irregular log-likelihood curves (e.g., flat, monotonically increasing, monotonically decreasing, or multi-modal) under the Newton-Raphson algorithm is avoided.

The algorithm employed by SCOREST for computing MLE $\hat{\theta}$ is a modified grid search, using the appropriate psychometric model(s) for each item. First, the theta score denoted θ_1 which maximizes log-likelihood over 121 equally spaced thetas between -6 and 6 (spaced by .1) is selected. Then the lower bound is set to be $\theta_1 - .1$ and the upper bound is set to be $\theta_1 + .1$, and the theta score denoted θ_2 that yields the maximum log-likelihood over 121 theta values spaced by .01 is computed. Similarly, the lower bound is set to be $\theta_2 - .01$ and the upper bound is set to be $\theta_2 + .01$, and the theta score denoted θ_3 that yields the maximum likelihood over 121 theta values spaced by .001 is computed. This procedure is repeated until the spacing between theta values is less than .00001.

After the MLE $\hat{\theta}$ is computed, the SE for $\hat{\theta}$ is computed using the following algorithm. The test information at $\hat{\theta}$ is evaluated by summing item information functions over operational items administered. The calculation of item information depends on whether the item is a multiple choice item (and satisfies a 3PLM) or is a constructed response item (and satisfies a GPCM). The SE for $\hat{\theta}$ is then computed as the square root of the inverse of the test information.

Results of Test Runs

Comparisons between Pearson's ISE, PARSCALE, MULTEST (or MIXEDEST), and SCOREST results were conducted to produce some initial information on how well results from these programs match on writing, reading, mathematics and science thetas for the 2007 spring administration. For this comparison study, ISE, MULLTEST, MIXEDEST and SCOREST used the 2007 spring initial form test samples from the final match file, and PARSCALE used the 2007 initial test calibration sample datasets. N-counts for these samples were over 100,000. Note that student records with missing item scores or $\hat{\theta}$ in the match file were excluded from this study. Also, student records for which PARSCALE did not produce a theta estimate (i.e., $\hat{\theta} = 999$ reported by PARSCALE) were excluded from the analysis. The study results demonstrated that SCOREST, MULTEST and MIXEDEST yielded acceptable thetas in comparison to PARSCALE and PEM's ISE. All the absolute differences among the methods were within .001.

Scoring Procedures for the Spring 2009 MME Administration

Upon OEAA's approval of item parameter estimates for MME forms (i.e., the initial, makeup or accommodation forms) for MME writing, mathematics, reading and science and raw-to-scale conversions for MME social studies, a score file using the matched file record layout was produced by ACT. This score file contained IDs, MME attemptedness flag, MLE theta estimates computed using SCOREST, MME scale scores and other scores reported for MME. This file was passed to ACT's Measurement Research Department (MRD) for QC checks.

Using the score file, for students who did not meet MME attemptedness criteria by MME subjects, MRD checked that no MME scores (e.g., MLE thetas and MME scale scores) were computed. For students who met MME attemptedness criteria by MME subjects, MRD independently computed MLE thetas, SE of thetas, MME scale scores and SE of MME scale scores using MULTEST or MIXEDEST, and these scores

were checked against those in the score file. Other scores in the score file that were checked by MRD include MME raw scores, MME performance levels, all raw scores for all MME standards along with their percent correct and possible points. Also, the high and low MME scale score values for each student were checked. Finally, a SAS task was implemented to check the MME ELA scores. Upon all the scores passing MRD's QC checks, a file was created and passed to ACT's IT team. For the spring 2009 MME administration, all score files delivered to OEAA passed MRD's QC checks.

Chapter 10: Score Precision

Reliability of a test can be regarded as a measure of consistency of the observed test scores. Since consistency entails repeated realization of an entity, reliability of a test is hard to estimate, particularly with a single administration of a test. If observed test scores are not consistent, or unreliable, the inconsistency is due to measurement error, more specifically, random error of measurement. Statistically, reliability is thus defined as the ratio of the true score (i.e., score without random error) variance to the observed score variance. Several methods of estimating reliability coefficients have been presented. Among them, coefficient alpha is popular because it is based on internal consistency from a single administration. When the measurement is used for classification decisions, classification consistency may function as a reliability measure.

Internal Consistency Reliability

For the spring 2009 administration, over 45,000 examinees of the initial test samples were included in the reliability analysis dataset, depending on the content area. Table 10.1 exhibits the alpha coefficients (Cronbach's alpha) for the 2009 spring MME administration.

Table 10.1. Cronbach's alpha for Spring 2009

Assessment	Cronbach's alpha
Writing	0.89
Reading	0.88
Mathematics	0.92
Science	0.87
Social Studies	0.85

For the constructed response items, Table 10.2 presents the percentage of agreement between two raters on the constructed response items.

Table 10.2. Rater Validity Percent of Agreement for Spring 2009

Absolute Score Difference Between Two Raters	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	94271	75.07	94271	75.07
1	30792	24.52	125063	99.59
2	500	0.40	125563	99.99
3	14	0.01	125577	100.00

Further Evidence of Reliability on ACT Writing

Data from a special study (ACT, 2009) was used to estimate alternate forms reliability of the ACT writing test, where approximately 6,000 high school students took two forms of the essay. Counter-balancing was used so that each form was administered in both the morning/first session and the afternoon/second session. Approximately 30 different essay forms were used in this study and essays were assigned randomly to both

students and a pool of raters. A “test-retest” correlation was computed for each pair of essays by comparing the scores an examinee received on each of his/her two prompts, and the average of these reliability coefficients yielded the value of 0.67. This value includes variability due to different essay forms and different essay raters.

Reliability analyses were also conducted using data from a field test study in which new prompts were administered to students prior to operational use, to examine how well they worked. OEAA requested that ACT provide additional reliability analyses for constructed response items, so ACT conducted the following study to comply with that request. Each examinee responded to two prompts on successive days. The prompts were spiraled to control for sampling error, and administered in counterbalanced order to control for order effects. To carry out these reliability analyses, several prompts were scored in a students x prompts x raters, utilizing a completely crossed design. There were six prompts, each administered to 20 examinees, and scored by two raters on a 1-6 scale. The prompts and examinees were chosen randomly from those in the field test study. Generalizability Theory analyses produced G-coefficients (internal consistency indices of score consistency) for each prompt pair. The median G-coefficient for the writing test was .70 over the six prompt pairs. Prompts and raters contributed negligible amounts to the total variance, which means the level of student achievement, not the particular prompt asked or the particular raters doing the scoring, is what most strongly determines the scores. Lastly, it was found that the median inter-rater reliability was .94 over the 12 prompts in the Generalizability study.

Empirical IRT Reliability

For the IRT methods, the conditional standard error of measurement (CSEM) is computed as part of the item parameter estimation process, via the test information function. Once the mean squared CSEM over examinees is computed, the equation below can be used to compute the reliability given in Table 10.3. In reference to this equation, $\bar{\sigma}^2(E)$ is the mean squared CSEM and $\sigma^2(S)$ is the observed variance of scale scores for the test taken over examinees.

$$rel = 1 - \frac{\bar{\sigma}^2(E)}{\sigma^2(S)}$$

Table 10.3. Empirical IRT reliability for Spring 2009

Assessment	Empirical IRT reliability
Writing	0.87
Reading	0.88
Mathematics	0.90
Science	0.83
Social Studies	0.81

MME Scale Score Reliability

Because the MME scale is a linear function of theta, MME scale score reliabilities are the same as the theta reliabilities. Therefore, the reliabilities in Table 10.3 are also the reliabilities of the MME scale scores.

SEM/Information Curves with Cuts Scores (Imposed)

Appendix A exhibits the plots of SEM/information curves produced by PARSCALE with the MME cut scores imposed for the testing subjects of Writing, Reading, Mathematics and Science, respectively. The vertical lines represent the performance level cut scores. For spring 2009, the performance levels were Not Proficient, Partially Proficient, Proficient, and Advanced.

Classification Consistency and Classification Accuracy

Classification consistency indices quantify the reliability of categorizing examinees into mastery or achievement levels, with respect to specific standards. Several model-based approaches have been developed for estimating classification consistency for a single test administration because repeated testing data are seldom available. An IRT model-based approach (Lee, Hanson, & Brennan, 2002) is used in this technical report to calculate the agreement index, P .

Assuming the two raw score random variables X_1 and X_2 from two administrations of a test are independent and identically distributed, the conditional joint distribution of X_1 and X_2 is given by $f(x_1, x_2 | \theta) = f(x_1 | \theta)f(x_2 | \theta)$, where θ denotes true examinee ability. Then, the marginal joint distribution of X_1 and X_2 can be obtained by integrating the conditional probabilities over the distribution of θ as

$$f(x_1, x_2) = \int f(x_1, x_2 | \theta)g(\theta)d\theta .$$

A consistent classification is made if both x_1 and x_2 for an examinee belong to the same category I_h ($h=1, 2, \dots, H$). The conditional probability of falling in the same category on the two testing occasions is

$$Pr(X_1 \in I_h, X_2 \in I_h | \theta) = \left[\sum_{x_1=c_{(h-1)}}^{c_h-1} f(x_1 | \theta) \right]^2 ,$$

where $c_1, c_2, \dots, c_{(H-1)}$ are raw cutoff scores, c_0 is the lowest raw score, and c_H is a perfect test score. Then, the agreement index P conditional on θ is obtained by

$$P(\theta) = \sum_{h=1}^H Pr(X_1 \in I_h, X_2 \in I_h | \theta),$$

and the marginal values of agreement index can be computed by

$$P = \int P(\theta)g(\theta)d\theta .$$

For each MME assessment, there are three cutoff score points and four categories at the scale-score level. Since there are four categories, examinees are classified into one of the four mutually exclusive categories based on their scale scores and the cutoff points on the MME assessment. To estimate classification consistency, however, 4×4 contingency tables for the MME assessment are created using the psychometric model, with the columns and rows showing the four classification categories. The elements of the 4×4 tables indicate the joint probabilities of examinees being classified in the pairs of the column and row categories; for example, being classified in the Basic level on one occasion (column) and in the Proficient Standards level on the other (row). The sums of the diagonal elements of the 4×4 tables are the indices of classification consistency.

The data used to compute classification consistency reported in Table 10.4 were obtained from the MME tests administered in spring 2009. The three parameter logistic model, the generalized partial credit model and the Rasch model are used to estimate the classification index. The basic role of these IRT models is to estimate the theta distribution and predict the observed score distribution. Once these distributions are estimated, 4×4 contingency tables can be created, which, in turn, are used as a basis for computing the classification index. Table 10.4 shows the 4×4 contingency tables and indices of classification consistency for the MME assessments.

Table 10.4. The 4 × 4 contingency table and classification consistency for the MME assessments for Spring 2009

MME Writing

	Not Proficient	Partially Proficient	Proficient	Advanced
Not Proficient	0.03192	0.03842	0.00014	0.00000
Partially Proficient	0.03842	0.32287	0.06618	0.00000
Proficient	0.00014	0.06618	0.36809	0.01940
Advanced	0.00000	0.00000	0.01940	0.02882

MME Reading

	Not Proficient	Partially Proficient	Proficient	Advanced
Not Proficient	0.05905	0.04319	0.00730	0.00000
Partially Proficient	0.04319	0.11619	0.07355	0.00000
Proficient	0.00730	0.07355	0.52035	0.01705
Advanced	0.00000	0.00000	0.01705	0.02223

MME Mathematics

	Not Proficient	Partially Proficient	Proficient	Advanced
Not Proficient	0.20298	0.04735	0.00932	0.00000
Partially Proficient	0.04735	0.06813	0.04804	0.00001
Proficient	0.00932	0.04804	0.31093	0.03263
Advanced	0.00000	0.00001	0.03263	0.14325

MME Science

	Not Proficient	Partially Proficient	Proficient	Advanced
Not Proficient	0.15328	0.04427	0.01958	0.00000
Partially Proficient	0.04427	0.04640	0.05236	0.00000
Proficient	0.01958	0.05236	0.44933	0.02882
Advanced	0.00000	0.00000	0.02882	0.06090

MME Social Studies

	Not Proficient	Partially Proficient	Proficient	Advanced
Not Proficient	0.07096	0.01738	0.00817	0.00001
Partially Proficient	0.01738	0.02002	0.02462	0.00022
Proficient	0.00817	0.02462	0.16989	0.04612
Advanced	0.00001	0.00022	0.04612	0.54610

Table 10.5 provides classification accuracy indices for the MME scales using an index based on estimated thetas and conditional standard errors. Classification accuracy evaluates the degree of accuracy of classifying examinees into score categories based upon observed scores. An expected classification accuracy index (Martineau, 2007) using measurement error is employed in this report. Let κ denote the vector of $H+1$ cut scores that divide the theta score scale into H categories, or $\kappa = [\kappa_1, \kappa_2, \dots, \kappa_{H+1}]$ where $\kappa_1 < \kappa_2 < \dots < \kappa_{H+1}$ and $\kappa_1 = -\infty, \kappa_{H+1} = \infty$. For an examinee i with observed theta score $\hat{\theta}_i$ and standard error $SE_{\hat{\theta}_i}$, an expected probability that the student falling into the h_i performance level under the assumption of conditional normality of measurement error is defined as the area from κ_h to κ_{h+1} under the normal curve with mean $\hat{\theta}_i$ and standard deviation $SE_{\hat{\theta}_i}$. Let $p_{ih_i} = \phi(\kappa_{h_i}, \kappa_{h_i+1}, \hat{\theta}_i, SE_{\hat{\theta}_i})$ represent this expected probability. Then, the expected classification accuracy index, based on measurement error, is equal to $\tau = \sum_{i=1}^N \phi(\kappa_{h_i}, \kappa_{h_i+1}, \hat{\theta}_i, SE_{\hat{\theta}_i}) / N$ where N is the number of examinees. This index ranges from 0 to 1, with 0 indicating no accuracy in examinee classifications, with 0.5 indicating random accuracy, and 1 indicating perfect expected accuracy in examinee classification.

Table 10.5. Classification accuracy for the MME assessments using four classification categories Spring 2009

Assessment	Index Value
Writing	0.84
Reading	0.82
Mathematics	0.82
Science	0.79
Social Studies	0.79

Chapter 11: Validity

Validity refers to the extent to which scores reflect what the test is intended to measure. As stated in the *Standards for Educational and Psychological Testing* (1999), validity refers to the “degree to which evidence and theory support the interpretations of test scores entailed by the proposed uses of tests.” This statement shows that validation is an ongoing process, which begins the moment that work on a test begins and continues throughout the life of the test. Validation is the process of continually accumulating and reviewing evidence from various sources to refine the utility of a test for making recommended interpretations consistent with the intended uses of the test scores.

Construct Validity Evidence from Content and Curricular Validity

Content validity involves the systematic examination of test content to determine whether it covers the curricular standards to be measured. As stated in Chapter 3, the MME augmentation is used to measure content which Michigan educators believe all students should know and be able to achieve in the content areas that are not measured by the ACT and *WorkKeys* assessments. Assessment results quantify how Michigan students and schools perform when compared with standards established by the State Board of Education. The MME is based on an extensive definition of the content the test is intended to assess and its match to the content standards. Therefore, the MME assessments are content-based and aligned directly to the statewide content standards.

Relation to Statewide Content Standards

Prior to the development and implementation of the MME, a committee of educators, item development experts, assessment experts, and OEAA staff met annually to review new and field-tested items for use on the MEAP (the old high school assessment). These stakeholders now meet to review new and field-tested items for use in augmenting the MME. The OEAA has established a sequential review process, as illustrated in Figure 11.1. This process continues to provide many opportunities for these professionals to offer suggestions for improving or eliminating items and to offer insights into the interpretation of the statewide content standards. These review committees participate in this process to ensure test content validity.

In addition to providing information on the difficulty, appropriateness, and fairness of these items, committee members provide a necessary check on the alignment between the items and the content standards they are intended to measure. When items are judged to be relevant (i.e., representative of the content defined by the standards), this provides evidence to support the validity of inferences made with MME results regarding knowledge of this content. When items are judged to be inappropriate for any reason, the committee can either suggest revisions (e.g., reclassification or rewording) or elect to eliminate the item from the field-test item pool. Items that are approved by the content review committee are later embedded in live MME forms to allow for the collection of performance data. In essence, these committees review and verify the alignment of the test items with the objectives and measurement specifications to ensure that the items measure appropriate content. The nature and specificity of these review procedures provide strong evidence for the content validity of the MME.

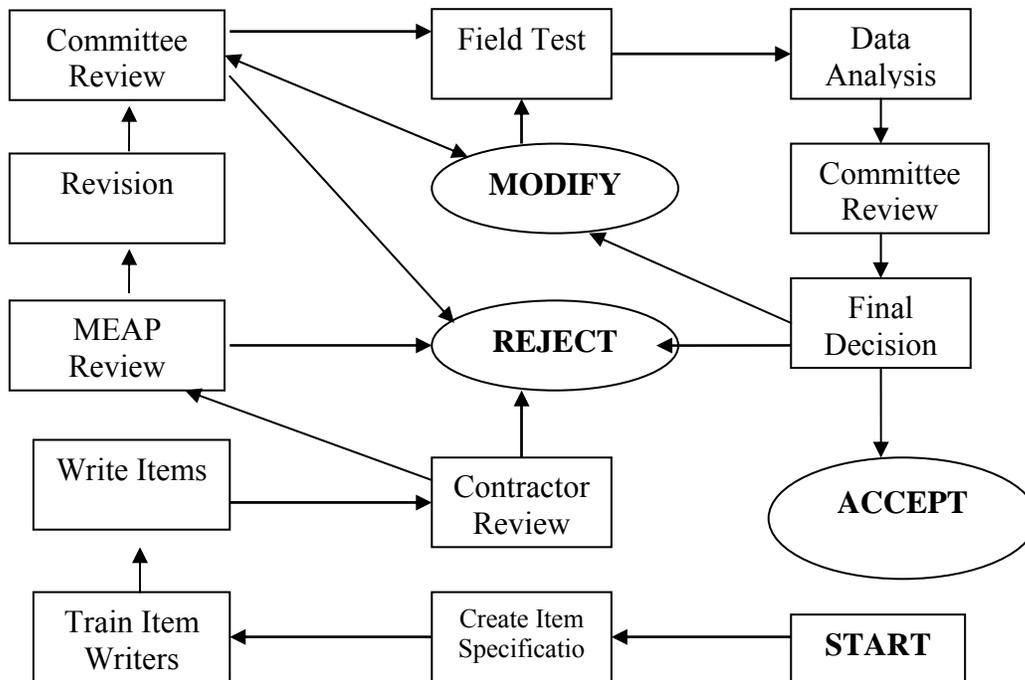


Figure 11.1. Item development/review cycle.

MME Alignment Studies

As detailed in Chapter 5, “Test Development Analyses,” two alignment studies have been performed for the MME, documenting alignment of the overall set of items from the ACT, *WorkKeys*, and Michigan-developed augmentation to Michigan’s content standards. These independent alignment studies provide validity evidence which is complementary to the input provided during content reviews. Along with the reliability analyses and other technical analyses, these alignment studies provide strong evidence of the validity of MME.

Educator Input

Michigan educators provide valued input on the MME content and the match between the items and the statewide content standards. In addition, many current and former Michigan educators and some educators from other states work as independent contractors to write items specifically to measure the objectives and specifications of the content standards for the MME. Using a varied supply of item writers provides a system of checks and balances for item development and review that reduces single source bias. Because many people with various backgrounds write the items, it is less likely that items will suffer from a bias that might occur if items were written by a single author. This direct input from educators, many of whom serve on the aforementioned committees, offers evidence regarding the content validity of the MME.

Construct-related Validity Evidence from Criterion Validity Analyses

Criterion validity refers to the degree to which a test correlates with other external outcome criteria. Criterion validity addresses how accurately criterion performance can be predicted from test scores. The key to criterion-related evidence is the degree of relationship between the assessment and the outcome criterion. The criterion should be relevant to the assessment and reliable. As the ACT and *WorkKeys* are administered intact as a part of the MME, and there is substantial evidence concerning their reliability and validity, there is a built in relevance of these criteria to the MME.

There is a large body of evidence that the ACT successfully predicts success in college, and that *WorkKeys* successfully predicts workplace success. As a criterion, a strong correlation of MME with *WorkKeys* and the ACT would indicate that the MME also can be used to predict college and workplace success.

The correlations among the old high school MEAP, the MME, the ACT, and *WorkKeys* from the Spring 2006 pilot are presented in Table 11.1. The cells reported in bold are the correlations between the ACT and the MME scores and the *WorkKeys* and MME scores. These correlations are very high, and indicate that the MME should be approximately as effective in predicting workplace and college success as the ACT and *WorkKeys* assessments.

In addition, the correlations among the MME and old high school MEAP are strong, indicating that as expected, the assessments measure similar constructs.

For the MME Spring 2009 administration, the correlations among the MME and the ACT and *WorkKeys* scale scores were as follows. The sample sizes employed for computing these correlations were over 100,000.

- MME Writing and ACT English: .84
- MME Reading and ACT Reading: .86
- MME Reading and *WorkKeys Reading for Information*: .84
- MME Mathematics and ACT Mathematics: .80
- MME Mathematics and *WorkKeys Applied Mathematics*: .86
- MME Science and ACT Science: .79

Table 11.1. Correlations between MME and other related measures for the Spring 2006 pilot.

Correlations (based on 3306 students who had valid scores on all MME subjects)																				
Subject		ELA								Mathematics				Science			Social Studies			
		English	Writing			Reading														
		ACT	MME	ACT	MEAP	MME	ACT	WK	MEAP	MME	ACT	WK	MEAP	MME	ACT	MEAP	MME	MEAP		
ELA	English	ACT	1.00	0.96	0.47	0.51	0.76	0.75	0.62	0.60	0.72	0.72	0.59	0.68	0.75	0.71	0.67	0.67	0.67	
		MME	0.96	1.00	0.59	0.57	0.78	0.74	0.63	0.62	0.73	0.71	0.59	0.69	0.75	0.71	0.67	0.67	0.67	0.67
	Writing	ACT	0.47	0.59	1.00	0.52	0.44	0.42	0.34	0.39	0.40	0.39	0.29	0.38	0.39	0.41	0.34	0.35	0.35	0.35
		MEAP	0.51	0.57	0.52	1.00	0.47	0.44	0.38	0.46	0.43	0.40	0.34	0.44	0.43	0.41	0.40	0.41	0.41	0.41
	Reading	MME	0.76	0.78	0.44	0.47	1.00	0.89	0.82	0.60	0.69	0.64	0.60	0.62	0.74	0.68	0.66	0.68	0.68	0.68
		ACT	0.75	0.74	0.42	0.44	0.89	1.00	0.59	0.56	0.61	0.61	0.51	0.57	0.69	0.65	0.62	0.64	0.64	0.64
		WK	0.62	0.63	0.34	0.38	0.82	0.59	1.00	0.51	0.63	0.57	0.58	0.57	0.65	0.59	0.58	0.58	0.58	0.58
		MEAP	0.60	0.62	0.39	0.46	0.60	0.56	0.51	1.00	0.52	0.49	0.43	0.52	0.58	0.51	0.56	0.59	0.59	0.59
Mathematics	MME	0.72	0.73	0.40	0.43	0.69	0.61	0.63	0.52	1.00	0.90	0.88	0.84	0.81	0.77	0.71	0.66	0.66	0.66	
	ACT	0.72	0.71	0.39	0.40	0.64	0.61	0.57	0.49	0.90	1.00	0.74	0.82	0.77	0.74	0.69	0.63	0.63	0.63	
	WK	0.59	0.59	0.29	0.34	0.60	0.51	0.58	0.43	0.88	0.74	1.00	0.72	0.70	0.65	0.63	0.58	0.58	0.58	
	MEAP	0.68	0.69	0.38	0.44	0.62	0.57	0.57	0.52	0.84	0.82	0.72	1.00	0.76	0.70	0.72	0.66	0.66	0.66	
Science	MME	0.75	0.75	0.39	0.43	0.74	0.69	0.65	0.58	0.81	0.77	0.70	0.76	1.00	0.89	0.88	0.76	0.76	0.76	
	ACT	0.71	0.71	0.41	0.41	0.68	0.65	0.59	0.51	0.77	0.74	0.65	0.70	0.89	1.00	0.67	0.65	0.65	0.65	
	MEAP	0.67	0.67	0.34	0.40	0.66	0.62	0.58	0.56	0.71	0.69	0.63	0.72	0.88	0.67	1.00	0.73	0.73	0.73	
Social Studies	MME	0.67	0.67	0.35	0.41	0.68	0.64	0.58	0.59	0.66	0.63	0.58	0.66	0.76	0.65	0.73	1.00	1.00	1.00	
	MEAP	0.67	0.67	0.35	0.41	0.68	0.64	0.58	0.59	0.66	0.63	0.58	0.66	0.76	0.65	0.73	1.00	1.00	1.00	

Criterion-related Validity Evidence for MME Science

Science standards underwent significant revisions prior to the 2009 MME administration. In order to compile additional evidence for criterion-related validity of the MME science scale scores, additional analyses were conducted. These analyses examine the criterion related validity of MME science scale scores. Using spring 2009 data, three external criterion variables were selected: 1) science course grades, 2) number of semesters students have taken science courses, and 3) whether students have taken advanced science courses.

Average MME science scale scores, grouped by each of the criterion variables are presented in Table 11.2, 11.3, and 11.4 respectively. As shown, the average MME science score increases as the course grade increases for the subjects of General Science, Biology, Chemistry and Physics. Students tend to have higher MME scores if they have taken science courses for a longer period of time, and students who have taken advanced science courses score higher than students who haven't. The criterion related validity of MME science is supported by this evidence.

Table 11.2. Average MME Science Scale Scores, by Course Grade of Science Courses

General Science	MME	Biology	MME	Chemistry	MME	Physics	MME
F	1070	F	1071	F	1079	F	1082
D	1077	D	1080	D	1088	D	1087
C	1085	C	1089	C	1098	C	1096
B	1098	B	1103	B	1112	B	1113
A	1118	A	1122	A	1128	A	1130

Table 11.3. Average MME Science Scale Scores, by Semesters of Science

Number of Semesters of Science	Mean MME Science Score
1	1061
2	1073
3	1079
4	1090
5	1090
6	1101
7	1101
8	1119

Table 11.4. Average MME Science Scale Scores, by Students with Advanced Courses in Natural Sciences

AP, Accelerated, or Honors Courses in Natural Sciences	Mean MME Science Score
Yes	1118
No	1110

DIF Analyses of the Spring 2009 MME Data

For the DIF analyses, only students who took all three days of the MME in the same administration mode—all initial, all makeup, or all accommodated—were considered. Of those students, only those who had valid flags of “Y” in the MME match file for that subject area (i.e., students who met attemptedness in the subject area, did not have nonstandard accommodations, did not have prohibited behavior, and were not involved in a misadministration) were included in the data sets for the analyses.

Two focal/reference group comparisons were conducted: females compared with males, and African Americans (Blacks) compared with Whites. For each multiple-choice item and each comparison, several statistics were computed: the Mantel-Haenszel delta statistic (MH-D), the value of the associated chi-square statistic (MH-CHISQ), the probability (P) of this chi-square value under the null hypothesis of no DIF, and the ETS A/B/C category for the item based on the values of MH-D and P. Table 11.5 presents the criterion for the A/B/C category. For a further description of these statistics and the A/B/C categories see, for example, Holland and Wainer, 1993. A positive MH-D denotes an item that favors the focal group, while a negative value indicates an item that favors the reference group.

Table 11.5. Criterion for the A/B/C category

Category	Description	Criterion
A	Negligible DIF	Nonsignificant MH-CHISQ ($P > 0.05$) or $ MH-D < 1.0$
B	Moderate DIF	Significant MH-CHISQ ($P \leq 0.05$) and $1.0 \leq MH-D < 1.5$
C	Large DIF	Significant MH-CHISQ ($P \leq 0.05$) and $ MH-D \geq 1.5$

For the polytomously-scored ACT Writing Test, in place of the MH-D statistic, the standardized mean difference (SMD) index, the standard deviation in Writing Test scores (SD) for the focal and reference groups combined, and the resulting effect size ($ES = |SMD/SD|$) are computed, as are the AA/BB/CC classifications resulting from the values of ES and P. Table 11.6 presents the criterion for the AA/BB/CC classifications. For a further description of these statistics and the AA/BB/CC categories see, for example, Dorans and Schmitt, 1991. A positive SMD index denotes an item that favors the focal group, while a negative value indicates an item that favors the reference group.

Table 11.6. Criterion for the AA/BB/CC classifications

Category	Description	Criterion
AA	Negligible DIF	Nonsignificant MH-CHISQ ($P > 0.05$) <u>or</u> Significant MH-CHISQ ($P \leq 0.05$) and $ES \leq 0.17$
BB	Moderate DIF	Significant MH-CHISQ ($P \leq 0.05$) and $0.17 < ES \leq 0.25$
CC	Large DIF	Significant MH-CHISQ ($P \leq 0.05$) and $ES > 0.25$

Matching Criterion

The matching criterion for each comparison was the total raw score for the subject area to which the item belongs. These raw scores are described below:

1. Writing: This is the sum of the ACT English Test raw score and the ACT Writing Test score. This sum ranged from 0 to 87.
2. Reading: This is the sum of the ACT Reading Test raw score and the *WorkKeys Reading for Information* Test raw score. This sum ranged from 0 to 70.
3. Mathematics: This is the sum of the ACT Mathematics Test raw score, the *WorkKeys Applied Mathematics* Test raw score, the raw score on the eight *WorkKeys Locating Information* Test items that counted toward a student's MME Mathematics score, and the Day 3 Mathematics Test raw score (operational items only). This sum ranged from 0 to 118.
4. Science: This is the sum of the ACT Science Test raw score and the Day 3 Science Test raw score (operational items only). This sum ranged from 0 to 72.
5. Social Studies: This is the sum of the Day 3 SS Test raw score (operational items only) and the raw score on the six *WorkKeys Locating Information* Test items that counted toward a student's MME SS score. This sum ranged from 0 to 34.

For Days 1 and 2, there was just one operational initial form of each test. All students who took the initial ACT form, for example, responded to the same operational items. For Day 3, there were 10 versions of each initial form: Forms 0901-0910. For the Mathematics and Science forms, the set of operational items varied from one version to the next, with a number of items appearing in more than one version. (The operational items for Social Studies were the same across all 10 versions; the versions differed only in field-test items. The Mathematics and Science forms differed in field-test items, as well.) Because of this, the Mathematics or Science raw scores attained on any version of the Day 3 initial form are not directly comparable to those attained on any other version, and likewise the matching criteria described in (3) or (4) above are not comparable across the 10 versions. In order to provide the cleanest, most comprehensive results possible, separate DIF analyses were conducted on the students who took each of the 10 Day 3 Mathematics initial forms, and on the students who took each of the 10 Day 3 Science initial forms.

Tables 11.7 through 11.9 present the number of "A", "B," and "C" items, by subject area, for the initial, makeup and accommodated testing. Tables 11.10 through 11.12 break down the "B" and "C" items by the favored group (i.e., males or females, blacks or whites) for the initial, makeup and accommodated testing. Table 11.13 gives the SMD results for the ACT Writing Test.

Table 11.7. Summary of Mantel-Haenszel results by focal/reference groups and subject, Initial testing

	Number of items in Category			Total
	A	B	C	
Females/Males				
Writing	75			75
Reading	70			70
Mathematics	1097	67	16	1180
Science	683	35	2	720
Social Studies	31	3		34
<i>Locating Information</i>	20	2		22
Total	1976	107	18	2101
Blacks/Whites				
Writing	68	5	2	75
Reading	70			70
Mathematics	1136	38	6	1180
Science	694	16	10	720
Social Studies	32	2		34
<i>Locating Information</i>	21	1		22
Total	2021	62	18	2101

Table 11.8. Summary of Mantel-Haenszel results by focal/reference groups and subject, Makeup testing

	Number of items in Category			Total
	A	B	C	
Females/Males				
Writing	74	1		75
Reading	68	2		70
Mathematics	106	7	5	118
Science	70	2		72
Social Studies	33	1		34
<i>Locating Information</i>	21	1		22
Total	372	14	5	391
Blacks/Whites				
Writing	70	4	1	75
Reading	67	3		70
Mathematics	108	7	3	118
Science	70	2		72
Social Studies	32	1	1	34
<i>Locating Information</i>	21	1		22
Total	368	18	5	391

Table 11.9. Summary of Mantel-Haenszel results by focal/reference groups and subject, Accommodated testing

	Number of items in Category			Total
	A	B	C	
Females/Males				
Writing	74	1		75
Reading	69	1		70
Mathematics	108	9	1	118
Science	72			72
Social Studies	30	4		34
<i>Locating Information</i>	22	1		23
Total	375	16	1	392
Blacks/Whites				
Writing	74	1		75
Reading	67	3		70
Mathematics	116	2		118
Science	71		1	72
Social Studies	33		1	34
<i>Locating Information</i>	23			23
Total	384	6	2	392

Table 11.10. Numbers of Category “B” and “C” items, by favored group, Initial testing

	B		C		Total
	Females	Males	Females	Males	
Mathematics	45	22	5	11	83
Science	2	33		2	37
Social Studies	1	2			3
<i>Locating Information</i>	2				2
Total	50	57	5	13	125
	Blacks	Whites	Blacks	Whites	Total
Writing	2	3	1	1	7
Mathematics	7	31	2	4	44
Science	4	12		10	26
Social Studies	2				2
<i>Locating Information</i>		1			1
Total	15	47	3	15	80

Table 11.11. Numbers of Category “B” and “C” items, by favored group, Makeup testing

	B		C		Total
	Females	Males	Females	Males	
Writing	1				1
Reading	1	1			2
Mathematics	3	4	4	1	12
Science		2			2
Social Studies		1			1
<i>Locating Information</i>	1				1
Total	6	8	4	1	19

	B		C		Total
	Blacks	Whites	Blacks	Whites	
Writing	1	3		1	5
Reading	1	2			3
Mathematics	3	4	3		10
Science	1	1			2
Social Studies		1		1	2
<i>Locating Information</i>		1			1
Total	6	12	3	2	23

Table 11.12. Numbers of Category “B” and “C” items, by favored group, Accommodated testing

	B		C		Total
	Females	Males	Females	Males	
Writing		1			1
Reading	1				1
Mathematics	6	3	1		10
Social Studies	3	1			4
<i>Locating Information</i>		1			1
Total	10	6	1		17

	B		C		Total
	Blacks	Whites	Blacks	Whites	
Writing		1			1
Reading		3			3
Mathematics		2			2
Science				1	1
Social Studies				1	1
Total		6		2	8

Table 11.13. Summary of SMD results, by focal/reference groups, all testings

Testing	Females/Males		Blacks/Whites	
	Category	Group Favored	Category	Group Favored
Initial	BB	Females	AA	
Makeup	CC	Females	BB	Blacks
Accommodated	CC	Females	BB	Blacks

Validity Evidence for the Day 1 Stand Alone Component: ACT Assessment

Validity is often categorized into several types such as content validity, construct validity, and criterion-related validity. More fundamentally, validity can be defined as “the appropriateness, meaningfulness, and usefulness of the specific inferences made from test scores” (AERA et al., 1999, p.9). Since ACT scores can be used for diverse purposes, ACT scores have been thoroughly studied for common interpretations and uses, such as measuring educational achievement, making admissions decisions, making course placement decisions, evaluating the effectiveness of high school college-preparatory programs, and evaluating students’ probable success in the first year of college and beyond (ACT, 2007b). The following is a brief summary of the validity evidence of the ACT for its various uses. For the technical details such as descriptive and inferential statistics, see ACT (2007b).

Measuring Educational Achievement

Among the validity types, content validity is particularly important for the use of ACT to measure educational achievement. The ACT tests are designed to measure students’ problem-solving skills and knowledge in particular subject domains and are closely reviewed to ensure that the test content represents current high school and university curricula. This content validation process is standardized so that the ACT test scores can have the same meaning for all students, test forms, and test dates. Statistical analyses were also conducted. For example, ACT test results were compared with high school grades, and a strong relationship was found between them. Also, longitudinal growth was investigated using ACT, PLAN, and EXPLORE: the three testing programs of ACT’s Educational Planning and Assessment System (EPAS). The large intertest correlations and the increases in the average scores indicate EPAS is measuring educational achievement as students progress through the grades.

Making Admissions Decisions

Appropriate admissions decisions are important for students, parents, and postsecondary institutions alike. For this use of ACT tests, validity can be measured in relationship with first-year college grades and GPAs. Validity of ACT test scores and high school grades are also important since they can serve as multiple measures for making college admission decisions. Research studies conclude that the ACT Composite scores provide greater differentiation across levels of academic achievement during the first year of college than do high school GPAs, in terms of probable success during the first year of college.

Course Placement Decisions

The ACT tests were also designed to facilitate placement of first-year college students to appropriate-level courses such as “standard,” “remedial,” or “advanced.” Helping with placement decisions is accomplished by the close connection of the ACT test battery with subject matter content. The content specifications of the ACT tests are based on the recommendations of nationally representative panels of secondary and postsecondary educators. Statistical relationships of ACT scores with course grades and high school GPAs have also been investigated by subjects and for subgroups. It was found that a typical institution using the ACT optimal cutoff score from their data could expect that at least 64% of the placement decisions would be correct decisions.

Indicators of Educational Effectiveness

The ACT tests can be used to evaluate college-preparatory programs since they have been developed to measure academic skills and knowledge that are obtained in high school and are necessary for academic success in the first year of college. However, a content review should be conducted to determine the extent to which the tests represent important outcomes the college-preparatory programs wish to measure.

Evaluating Probable College Success

The use of the ACT tests to evaluate probable college success is closely connected with the other uses of the ACT. According to recent studies, the ACT College Readiness Benchmarks show that students who are college-ready are more likely to immediately enroll in college, and once they enroll, tend to be more successful during their first year. Also, the ACT College Readiness Benchmarks can assist in determining who will succeed in college, even into the second year.

Validity Evidence for the Day 2 Stand Alone Component: the WorkKeys Assessments

WorkKeys assessments are designed for use in both business and educational settings. To support these uses, ACT has adopted a multi-faceted approach to validation of *WorkKeys* Assessments: *Reading for Information* (RFI), *Applied Mathematics* (AM), and *Locating Information* (LI). Three types of validity evidence have been collected to justify the use of *WorkKeys* assessment scores, including content-related evidence, criterion-related evidence, and construct related evidence. To accumulate such evidence, ACT has conducted validity studies or worked with organizations to collect data on students and employees. The results are reported in *WorkKeys* Assessment Technical Manuals (ACT, 2008a, 2008b, and 2008c).

Content-Related Evidence

To support the content-related validity of the three test scores, ACT uses two job analysis procedures—*WorkKeys* Job Profiling and the SkillMap Job Inventory—to link the *Reading for Information*, *Applied Mathematics*, and *Locating Information* Skill Levels, to relevant job behaviors. *WorkKeys* Job Profiling and SkillMap are both designed to meet federal standards and other industry guidelines for content validation of employment tests used for high-stakes decisions such as hiring and promotion. Both procedures can be used to define critical job tasks, determine which *WorkKeys* skills are relevant to performing the tasks, and identify the level of skill required for performing them.

Criterion-Related Evidence

To support the criterion-related validity of the three test scores, ACT has gathered data from various organizations on the correlation between the test scores used to select job applicants and their subsequent job performance ratings. While sample sizes and correlations vary from study to study, all of the correlations have been positive, ranging from 0.12 to 0.86, which compares favorably with the correlations typically found in the general research literature on criterion-related validity of employment tests. ACT has also conducted classification consistency studies, comparing the employees' job performance classification to their classification by *WorkKeys* Assessment Skill Levels. In these studies, the percentage of employees classified the same way by both measures ranged from 30 percent to 100 percent depending on sample size, skill level, and participants.

Construct-Related Evidence

To support the construct-related validity of the three test scores, ACT examined the relationship between *WorkKeys* assessments and other tests measuring somewhat similar skills and found a moderate correlation between the test scores. In addition, ACT examined the relationship among the three fundamental skill assessments. Initial results suggest that (1) each assessment measures unique job-related skills as well as some general skills, and (2) each assessment has a strong unidimensional structure.

Gender and Race/Ethnicity Analyses

Because *WorkKeys* assessments can be used for high-stakes employment decisions, ACT has analyzed Skill Level scores for evidence of adverse impact by gender and racial/ethnic groups. Evidence of adverse impact has been found to be consistent with existing research on the validity of employment test scores used for high-stakes selection decisions. In this context, such findings reinforce the need to clearly link use of *WorkKeys* test scores to the critical tasks and skills required for the job.

Fairness Review

Fairness Review for the ACT

According to the *Code of Fair Testing Practices in Education*, test developers should provide “tests that are fair to all test takers regardless of age, gender, disability, race, ethnicity, national origin, religion, sexual orientation, linguistic background, or other personal characteristics.” As a testing organization, ACT endorses the *Code* and makes every effort to see that all ACT tests are fair to the populations for which the tests are intended. The work of ensuring test fairness takes place during every stage of the test development process, including item (test question) writing and review, item pretesting, forms construction, and forms review. ACT is committed to ensuring that each of its testing programs upholds the *Code*’s standards for appropriate test development practice.

Item Writing, Review, and Pretesting

Most of the individuals who write items for ACT’s tests are actively engaged in teaching at the high school or the university level. ACT makes every attempt to include item writers who represent the diversity of the population of the United States with respect to ethnic background, gender, and geographic location. Item writers work closely with ACT test development associates in producing items and passages of high quality that are designed to meet the test specifications, represent diversity, and be fair to all examinees. Item writers are provided with detailed guidelines to assist them in developing test materials, including specific information on fairness concerns. Among these is the way in which various groups of the population are portrayed, and the degree to which representatives of various groups are depicted in active versus passive circumstances, as exhibiting stereotypic mental or physical characteristics or tendencies, or as engaged in particular occupations or roles. In addition, the construction of fair test items requires sensitivity to the changing circumstances of our society: increased variation in family structures; the multiethnic composition of the population; and a wide range of socioeconomic and urban, suburban, and rural lifestyles.

Every new test item and passage is comprehensively reviewed for fairness, interest, and appropriateness to the grade level for which the test is designed; adherence to specifications; soundness and defensibility; grammatical accuracy; and sound measurement characteristics. The items are first reviewed by ACT’s test development associates and editorial staff. Any problems found in this review are corrected immediately. The items are then sent to two groups of external reviewers: content experts (including classroom teachers, college faculty, and curriculum specialists, representing diversity as to geographic region, ethnicity, and gender), who focus on content accuracy, item classifications, skill levels, and grade-level appropriateness;

and fairness reviewers, who are of diverse ethnicity, gender, and geographic background and are sensitive to issues of test and item fairness. The fairness reviewers carefully examine all items and stimulus materials to make sure that they do not contain any language, roles, situations, or contexts that could be considered offensive or demeaning to any population group.

ACT selects fairness reviewers from among African American, Asian American, Latino, American Indian, and female consultants. ACT communicated with prominent, nationally recognized advocacy groups to obtain nominees who could review the ACT test materials. (See the ACT publication, *Fairness Report for the ACT Tests 2007-2008*, for a list of these organizations.) From the recommendations of these groups, ACT selects fairness reviewers who have a history of active participation in promoting the concerns of the group within educational settings and beyond.

Items that pass the content and fairness reviews are pretested on a representative sample of the ACT examinee population in a national administration of the ACT. The purpose of pretesting items is to determine whether the items are technically sound and at the appropriate level of difficulty for the ACT examinee population. Statistical indices of item difficulty and discrimination, among other statistics, are compiled on the basis of pretest results. Items are evaluated according to their performance in the pretest. Those that perform acceptably on all criteria are included in the item pool from which preliminary forms of the ACT tests are constructed.

Operational Forms Construction

Preliminary forms of the ACT tests are constructed using the items that survive the pretest. Items are selected to match the requirements of both the content and statistical specifications for the tests. The distributions of the items in each test form are also examined for fairness, variety, diversity, and balance. Each test form is balanced with regard to multicultural and gender representation. While it is impossible, given the constraint of the limited amount of material in each test form, to represent every group in every form, a good-faith effort to represent diversity should be discernible in every final form. Two strategies ACT uses to attain this diversity are ensuring the inclusion of culturally diverse passages within each form and ensuring that all passages depict universal themes applicable to all groups.

Preliminary versions of ACT test forms are subjected to the same comprehensive reviews for content and fairness issues that items undergo. ACT's test development associates, editorial staff, and measurement staff conduct the initial forms reviews. The forms are then sent to external content experts and fairness reviewers (not the same individuals who conducted the reviews prior to pretest). The comments made by all reviewers are collated, and the items/passages identified as problematic are replaced as necessary. ACT's test development associates work with the fairness and content reviewers to prepare the final forms of the test for printing. In all, at least sixteen independent reviews are made of each test item before its appearance on a national form of the ACT, primarily to ensure that each student's level of achievement is accurately and fairly evaluated.

Differential Item Functioning (DIF)

Fairness in the content of the items does not necessarily prevent items from functioning statistically in different ways for different population subgroups. Differential item functioning (DIF) can be described as a statistical difference between the probability of a specific population group (the "focal" group) getting the item right and a comparison population group (the "base" group) getting the item right *given that both groups have the same level of expertise* with respect to the content being tested.

To detect the existence of differential item functioning (DIF) for items in each test form, ACT analyzes the response data from actual national and state administrations of each of the forms. After each national or

state administration of a test form, large random samples representing the examinee groups of interest are selected from the total number of examinees taking the test. The groups compared are African Americans/Caucasians, Mexican Americans/Caucasians, Hispanics (other than Mexican origin)/Caucasians, Asian Americans or Pacific Islanders/ Caucasians, American Indians or Alaska Natives/Caucasians, and females/males. The statistics ACT uses for detecting DIF are the standardized difference in proportion-correct (STD) and the Mantel-Haenszel common odds-ratio (MH). The samples of examinees' responses to each item are analyzed using the STD and MH procedures. All items with MH and/or STD values exceeding a preestablished statistical tolerance level are flagged for further review. The flagged items are reviewed by ACT's test development associates for possible explanations of the unusual STD or MH results. In the event that a problem is found with an item, actions are taken as necessary to eliminate or minimize the influence of the problem.

Fairness review for WorkKeys

Fairness review is an important step in developing and pretesting news items for the three *WorkKeys* assessments. Participants in fairness review include ACT test development staff and external business and education experts from diverse cultural and ethnic backgrounds. Fairness reviewers representing gender, cultural, and ethnic/racial subgroups work to ensure that no item was unfair to any minority group members. ACT gives the reviewers written guidelines and requires them to write an evaluation of each item. ACT reviews the evaluations and responds to any concerns the reviewers raise. Any item rejected by the reviewers is removed from the operational pool. Items that pass reviews and meet specifications are left intact to preserve the accuracy of the pretest item data. Such items constitute the pool from which subsequent operational forms are drawn. Please see more details in *WorkKeys* Assessment Technical Manuals (ACT, 2008a, 2008b, and 2008c).

Conclusion

The evidence from the methods used for item development, item review, augmentation, alignment, and correlation with related measures provide validity-related evidences for the interpretation of MME scores. Given the desired interpretation of scores as described in this chapter, the validity-related evidence strongly support the interpretability of the MME scores.

Chapter 12: Item Analysis

Post-Field-Test Item Review

After field-test administration, item analyses were conducted to prepare data for two more rounds of reviews: bias/sensitivity review and content review. For the 2009 MME, the Rasch model was used for item analysis for the social studies portion of the exam. The three parameter logistic item response theory model was used for all other subjects on the exam. This section describes data based on Rasch model analysis for these two post-field-test reviews. A section on item field testing is also in Chapter 3, and the reader may refer to that section for a presentation that is complementary to this one.

Data

All field-test items were embedded in the live test forms for each test. After the calibration of live test forms, field-test items were calibrated and put onto the same scale as the live operational items. Appendix B lists all the statistics created for the field-tested items. The statistics for each field-test item can be summarized into nine categories.

1. General test information: test name, subject, grade, level;
2. Administration related information: year cycle, administration year, released position;
3. Specific item information: item ID, CID, item type, answer key, maximal score, maturity, item function, character code, number of forms the item appears on, form numbers, test position, n-count (total, male, female, white, and black students), percent for each comment code, percent for each condition code;
4. Content-related information: strand, benchmark, grade level expectation, depth of knowledge, domain, scenario;
5. Option analysis: percent for each option and each score point (total, male, female, white, and black students), p-value or item mean (total, male, female, white, and black students), adjusted p-value, difficulty flag, item standard deviation, item-total correlation, biserial/polyserial correlation, corrected point-serial correlation, item-total correlation flag, option point-biserial correlation, flag for potential miskeying;
6. DIF analysis: Mantel Chi-square, Mantel-Haenszel Delta and its standard error, signed and unsigned SMD, SMD signed effect size, DIF category, and favored group for male versus female comparison and white versus black comparison;
7. IRT parameters: b-parameter and its SE, step parameters and their respective SE, item information at cut points;
8. Fit statistics: mean-square infit, mean-square outfit, mean-square fit flag, misfit level;
9. Data for creating plots: conditional item mean for decile 1 to 10 for each student group (total, male, female, white, and black students) for creating conditional mean plots, 5th, 25th, 50th, 75th, 95th percentile for creating Box & Whisker plot for each student group (total, male, female, white, and black students) for each option and each score point.

These statistics were reviewed by Pearson and OEAA for creating item labels for bias/sensitivity review and content review.

Statistics and Graphs Prepared for Review Committees

Statistics from item analyses for field-test items were used to create item labels for the post-field-test reviews. Different sets of statistics were prepared for MC and CR items for review committee. Table 12.1 displays all the statistics prepared for MC items for the review committee. These include six categories.

1. General administration information: test name, grade, subject, and administration time;
2. Item general information: CID, maturity, forms and positions;
3. Item specific information: item type, key, p-value, n-count, Rasch difficulty, difficulty flag, point-biserial correlation, point-biserial correlation flag, fit flag, option quality flag;
4. Breakout group descriptives and optional analysis: percent of students selecting each option and omit, option point-biserial correlations, and n-count for all and subgroups: male, female, white, and black students;
5. Differential Item Functioning: flag, and favored group for male versus female and white versus black;
6. Review decision;

Table 12.2 displays all the statistics prepared for CR items for the review committee. These include seven categories.

1. General administration information: test name, grade, subject, and administration time;
2. Item general information: CID, maturity, forms and positions;
3. Item specific information: item type, maximal score point, adjusted p-value, item mean, n-count, Rasch difficulty, difficulty flag, item-total correlation, item-total correlation flag, fit flag, score point distribution flag;
4. Breakout group descriptives and score point distribution: percent of students obtaining each score point and omit and n-count for all and subgroups: male, female, white, and black students, omit point-biserial correlation;
5. Invalid code distributions: total invalid scores, frequency of students at each invalid code;
6. Differential Item Functioning: flag, and favored group for male versus female and white versus black;
7. Review decision;

All statistics prepared for the review committee for MC and CR items are explained in Appendix C. When the p-value for an MC item, adjusted p-value for a CR item, or Rasch difficulty was out of the desired range, a difficulty flag was shown. When a point-biserial correlation for an MC item or item-total correlation for a CR item was out of range, the appropriate flag was shown. If the mean square infit or outfit was out of desired range, an infit or outfit flag was presented. Similarly, if DIF or improperly functioning options (distracters) were detected, the corresponding flag was activated for the item. The criteria used for flagging an MC or CR item are presented in Table 12.3.

For further psychometric reference, conditional mean plots and Box & Whisker plot for two student group comparison, male versus female and white versus black were prepared for the flagged items for the two post-field-test reviews. See Figure 12.1a (for MC items) and 12.1b (for CR items) for conditional mean plots and Figure 12.2a (for MC items) and 12.2b (for CR items) for Box & Whisker plots.

Members of the bias review and content review committees were given specific training in analyzing item quality. Some of the supporting materials for the training sessions are provided in Appendix D (for bias review) and Appendix E (for content review).

Table 12.1. Item Label for a MC Item

MME Grade: 11 Subject: Social Science Admin: Fall 2006

CID: 6688999

GLCE: C.2.h.1

Form: 2

Position: 46

Passage:

- Accept as is
- Reject
- Accept with revision

Table 1. Item Information

Type: MC	P-value: 0.37	Rasch Difficulty: 0.15	Difficulty Flag:
Key: B	N-count: 860	PB Correlation: 0.24	PB Correlation Flag: CL
	Maturity: FT	Fit Flag:	Option Quality Flag: P

Table 2. Breakout Group Descriptives and Option Analysis

		N-count	Percent of Students Selected Option				
			A	B	C	D	Omit
Group	All	860	20	37*	21	20	2
	Male	447	21	35	21	20	3
	Female	413	18	40	20	21	1
	White	587	21	35	20	22	2
	Black	207	15	46	20	14	3
Option PB Correlations			-0.13	0.24	-0.14	0.04	

Table 3. Differential Item Functioning

Reference/ Focal Group	Male/ Female	White/ Black
Flag		C
Favored Group		Black

Explanation of DIF Flags
 Blank - No or negligible DIF
 B - Moderate DIF
 C - Large DIF

Table 12.2. Item Label for a CR Item

MME **Grade:** **11** **Subject:** **Social Science** **Admin:** **Fall 2006**

ID: 6666666 **Maturity:** FT
Form: 2 5
Position: 27 27
Passage: Government Health Care

- Accept as is
- Reject
- Accept with revision

Table 1. Item Information

Type: CR	Adj. P value: 0.34	Rasch Difficulty: 0.22	Difficulty Flag:
Max: 5	Item Mean: 1.71	Item-Total Corr: 0.55	Item-Total Corr Flag:
	N-count: 1574	Fit Flag:	Score Point Dist. Flag:

Table 2. Breakout Group Descriptives and Score Point Distributions

		N-count	Item Mean	Percent of Students at Each Score Point								
				0	1	2	3	4	5	6	Omit	
Group	All	1574	1.71	17	34	29	13	7				
	Male	811	1.54	22	36	25	10	7				
	Female	763	1.90	11	32	32	17	8				
	White	1028	1.77	16	33	29	13	9				
	Black	371	1.58	18	34	28	15	5				
Omit PB Correlation												

Table 3. Condition Code Distributions

Frequency of Students at Each Condition Code				
A	B	C	D	E
1		8		

Table 4. Differential Item Functioning

Reference/ Focal Group	Male/ Female	White/ Black
Flag	C	
Favored Group	Female	

Explanation of DIF Flags
 Blank - No or negligible DIF
 B - Moderate DIF
 C - Large DIF

Table 12.3. Flagging Criteria

Statistic	Flag	Flag Definition	Flag Field
PVAL PVAL ADJPVAL	PL PH BL	For MC 4 options, if p-value LT .3 (PL) or GT .9 (PH) For CR items, if adj. p-value LT .10 (PL) or GT .9 (PH)	DIFFICFL
BPAR	BH	If b-parameter LT -2.5 (BL) or GT 2.5 (BH)	
ITOT	CL	If item-total correlation or point biserial correlation LT 0.25 (CL)	ITOTFL
MSQIN MSQOUT	MH MM TP	If msqin or msqout GT 2 (MH) If msqin 1.5 through 2 and msqout LE 2 (MM) If msqout 1.5 through 2 and msqin LE 2 (MM) If msqin LT 0.5 and msqout LT 1.5 (TP) If msqout LT 0.5 and msqin LT 1.5 (TP)	MSQINFL MSQOUTFL
DIF_MF DIF_WB	A B C AA BB CC	For MC items: A: If either MH Delta is not significantly GT 0 ($p < 0.05$, using either MH-Chi-Sq or standard error of MH Delta) or if the MH Delta is LT 1 B: If MH Delta is significantly GT 0 and is either GE 1 and LE 1.5 or is GE 1 but not significantly GT 1 ($p < 0.05$, using standard error of MH Delta) C: If MH Delta is both GT 1.5 and significantly GT 1 ($p < 0.05$, using standard error of MH Delta) For CR items: AA: If the Mantel Chi-Sq is not significant ($p > 0.05$) or the Effect Size (ES) of SMD LE 0.17 BB: If the Mantel Chi-Sq is significant ($p < 0.05$) and the ES is GT 0.17 but LE 0.25 CC: If the Mantel Chi-Sq is significant ($p < 0.05$) and the ES is GT 0.25	DIF_MF DIF_WB Categories A and AA are not displayed in flag field
A, B, C, D M, S5, S6, O APB BPB CPB DPB OPB	H L P O N B	For MC items: If the keyed option is not the highest percentage (H) If any option LE 2% (L) If any non-keyed option pb-corr GT 0 (P), or if omit pb-corr GT 0.03 (O) If the keyed option pb-corr LT 0 (N) For CR items: For CR, if omit pb-corr GT 0.03 (O) For CR, if any score point LT 0.5% (L) For CR, if omit GT 20% (B)	MISKFL

Meaning of Flags:

- **PL ... p-value low**
- **PH ... p-value high**
- **BL ... b-parameter low**
- **BH ... b-parameter high**
- **CL ... correlation low between item and total**
- **MH ... misfit high**
- **MM ... misfit moderate**
- **TP ... too predictable**
- **A or AA ... no or negligible DIF**
- **B or BB ... moderate DIF**
- **C or CC ... substantial DIF**
- **H ... highest percentage is not a keyed option**
- **L ... low percentage of any option**
- **P ... positive pb-correlation for any non-keyed option**
- **N ... negative pb-correlation for the keyed option**
- **O ... omit has a positive pb-correlation**
- **B ... blanks are over 20%**

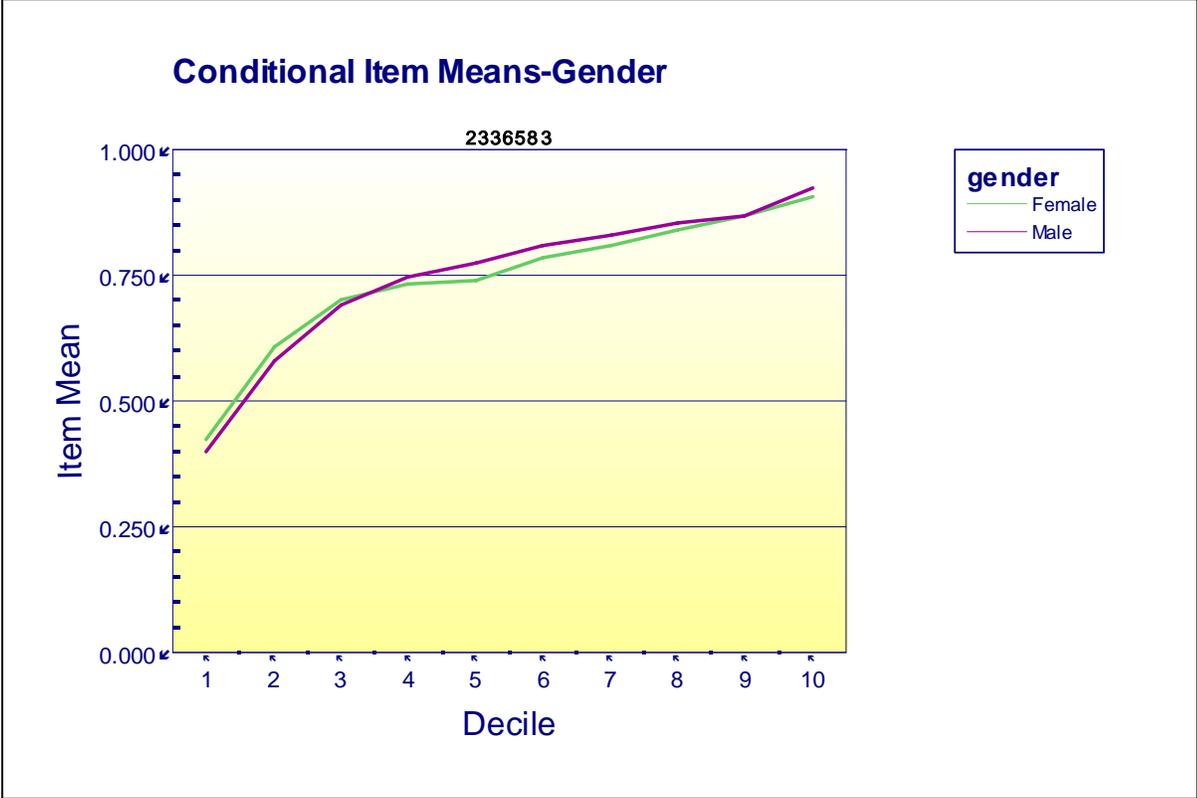
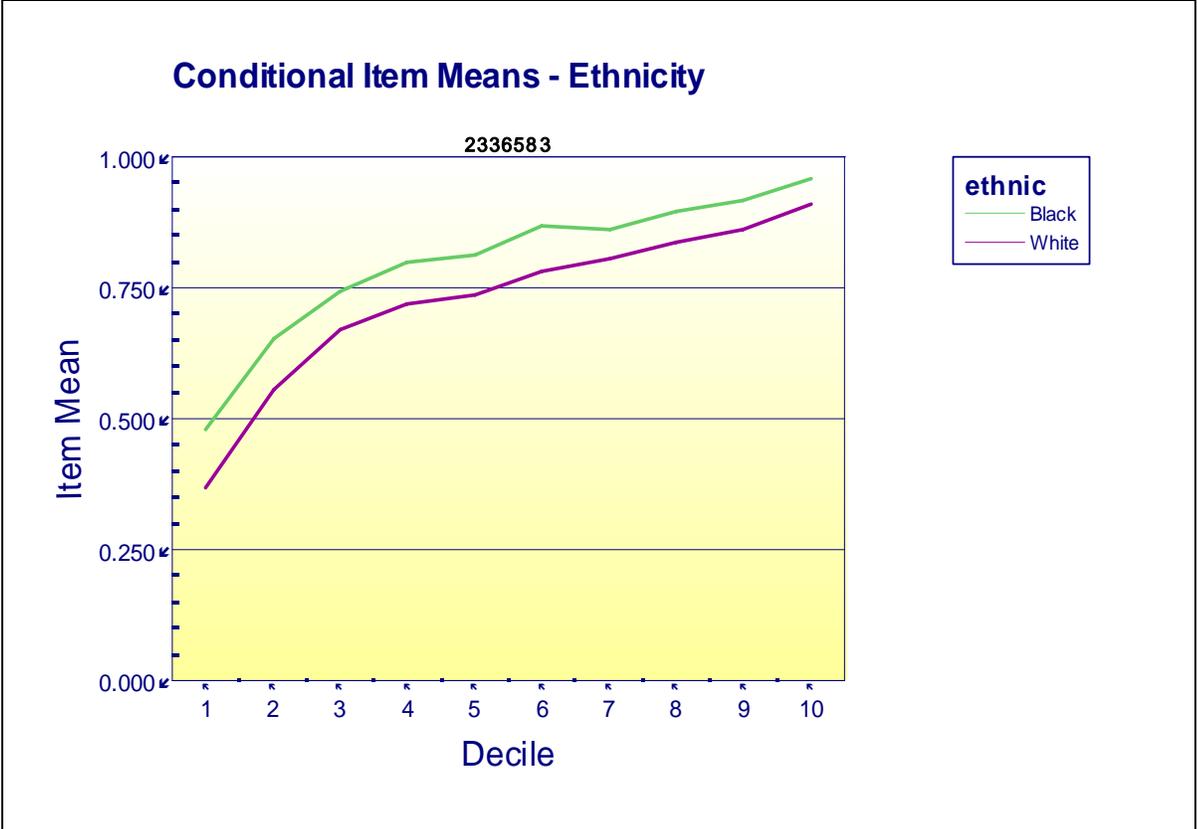


Figure 12.1a. Conditional item mean plots for ethnicity and gender for MC items.



Figure 12.1b. Conditional item mean plots for ethnicity and gender for CR items.

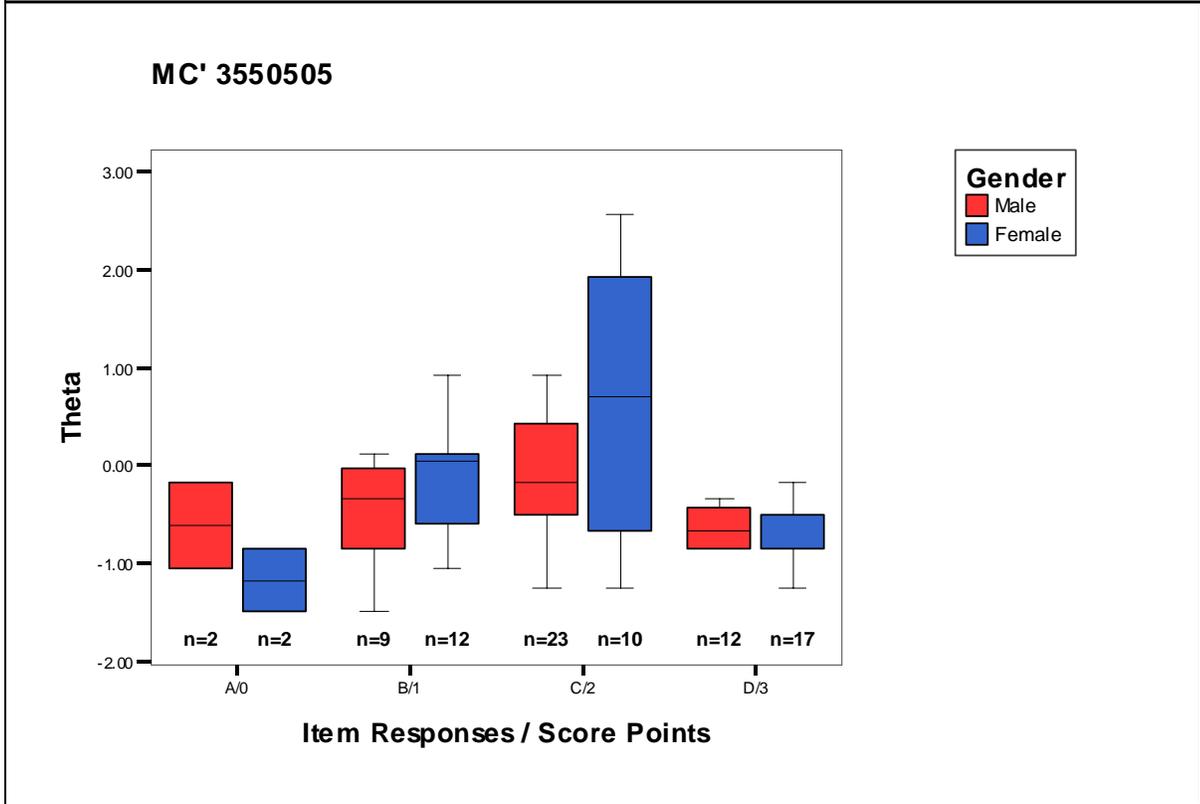
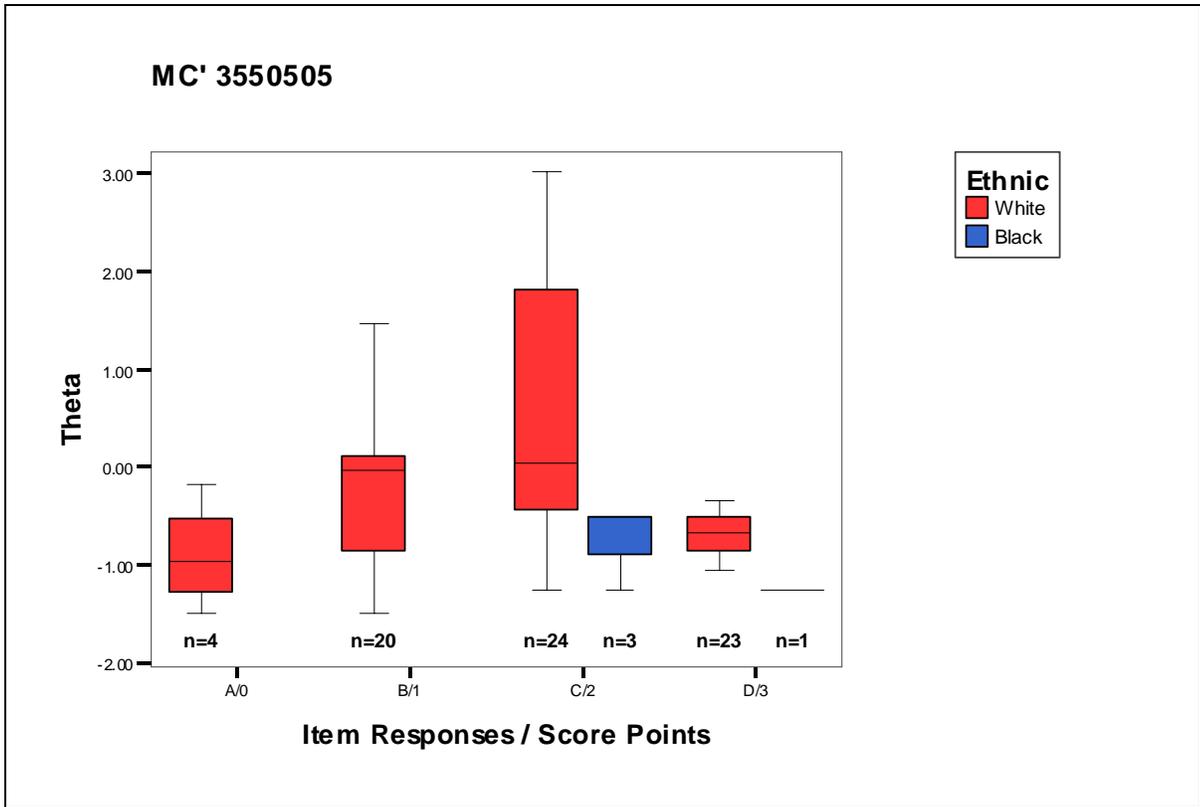


Figure 12.2a. Box & Whisker plots for ethnicity and gender for MC items.

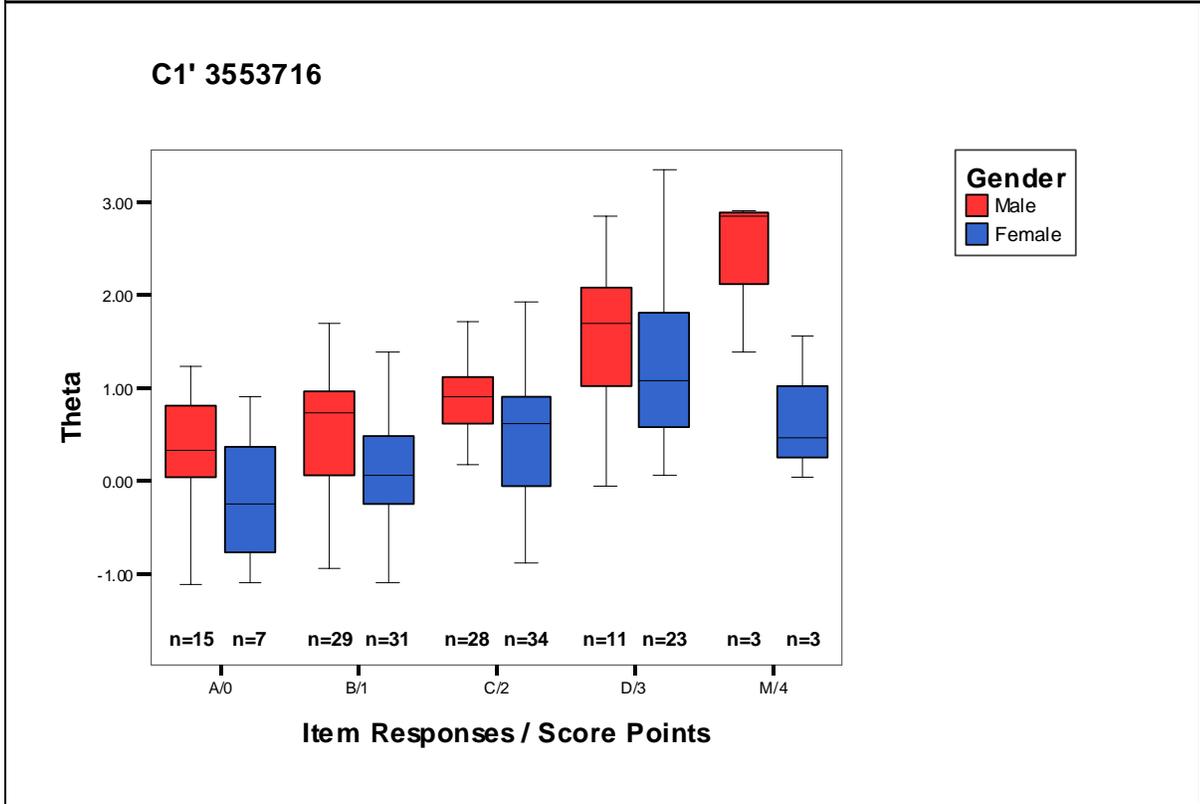
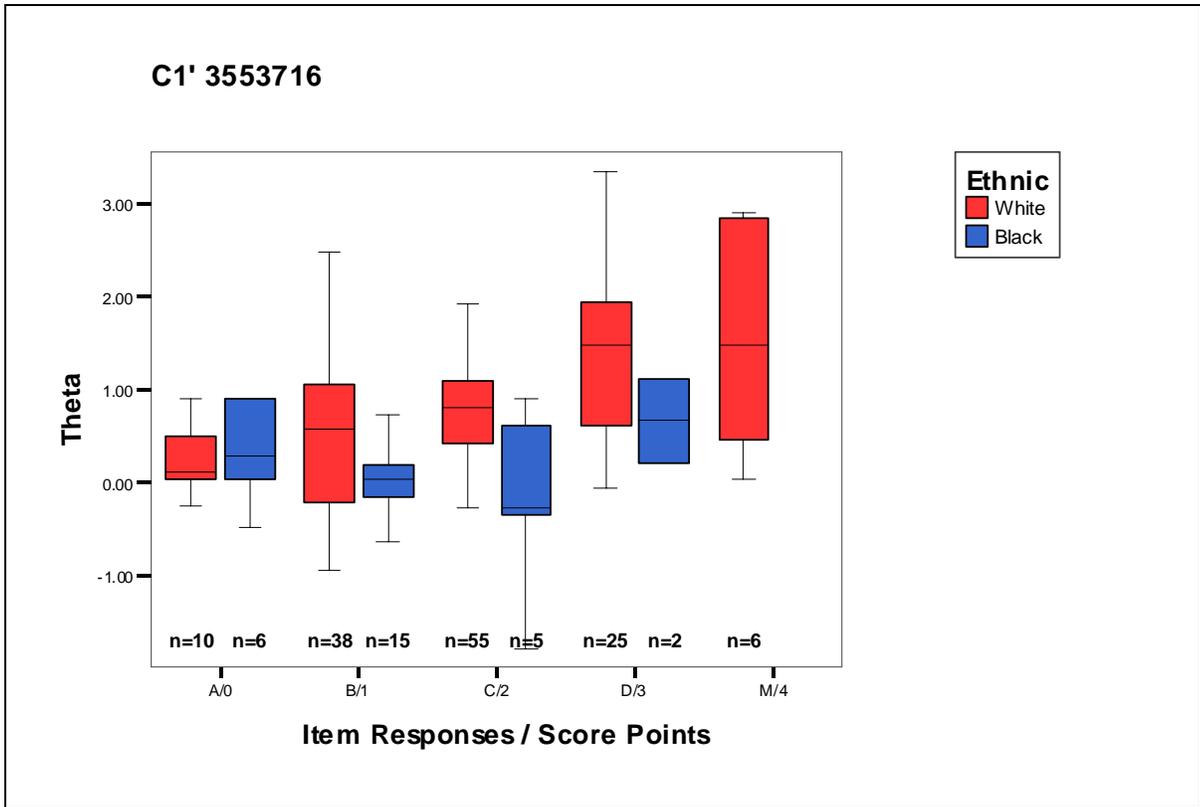


Figure 12.2b. Box & Whisker plots for ethnicity and gender for CR items.

Chapter 13: Standard Setting

Standard setting is a complex and detailed procedure that requires extensive documentation, particularly given the high-stakes nature of standard setting for state-administered assessments. In order to provide the most complete information possible regarding standard setting, those interested in learning more about the standard setting process are asked to reference the complete standard setting document, found at http://www.michigan.gov/documents/mde/Item_BB_177386_7.pdf. Below is a brief overview of the main activities involved in standard setting for the MME. Standard setting was conducted for the MME in 2006 using the procedures outlined below.

The plan for establishing cut scores for the performance levels is contained in the *Standard Setting Plan* (Assessment and Examination Service, 2006). This document describes the data collection, methodology (the Bookmark or Item Mapping method) and agenda for conducting the standard setting studies for the MME.

The results of a modified item mapping procedure are described in the *Standard Setting Report* (Assessment and Examination Service, 2006). The following modified item mapping method was used: “In the ordered item booklet, three items were flagged as reference items, one for each performance standard (Partially Proficient, Proficient, Advanced). If selected, these items would produce cut-scores such that the percentage of students in each of the four categories would be the same as the results of the Spring 2006 Grade 11 assessments.” The data for the MME standard setting were obtained from a group of panelists who reviewed items ordered with respect to a 2006 field test of the Michigan Merit Examination in Reading, Writing, Mathematics, and Science. The *Standard Setting Report* recommended three cut scores to delineate the four performance levels: Not Proficient, Partially Proficient, Proficient, or Advanced. The MME cut scores are reported in Table 13.1. These cut scores were placed on the MME scale.

A *Michigan Department of Education Memorandum* in October 2006 described four possible sets of cut scores for the performance levels, and recommended one. A second *Michigan Department of Education Memorandum* (November 2006) revised the recommendation to a different set of cut scores, and provided a justification based on a change in content specifications. The revised recommendation was to adopt MME cut scores based on a linkage to the MEAP.

The formal adoption of MME cut scores is detailed on page 5 of the minutes of the November 2006 State Board of Education meeting (*Minutes of the State Board of Education* November 14, 2006).

Table 13.1. Proficiency level cut scores by subject.

	Performance Standard Cut Score		
	Not Proficient/ Partially Proficient	Partially Proficient/ Proficient	Proficient/ Advanced
Writing	1051	1100	1146
Mathematics	1089	1100	1128
Reading	1078	1100	1158
Science	1087	1100	1143
Social Studies	1086	1100	1129
ELA	1065	1100	1152

Chapter 14: Adequate Yearly Progress and Education YES

The major policy-based uses of assessment data from the MME, MEAP and MI-Access are for public reporting and school accountability decisions.

Legislative Grounding

- The federal No Child Left Behind Act (NCLB) requires that Adequate Yearly Progress (AYP) be calculated for all public schools, for each school district, and for the state.
- Michigan statute (section 1280 of the Revised School Code) requires the State Board of Education to accredit public elementary and secondary schools. The State Board approved *Education YES – A Yardstick for Excellent Schools!* in 2002, and accepted the report of the Accreditation Advisory Committee in 2003.

NCLB requires that AYP be determined for all public schools, for each school district, and for the state. The school or district must attain the target achievement goal in reading and mathematics or reduce the percentage of students in the non-proficient category (Partially Proficient and Not Proficient) of achievement by 10% (“safe harbor”). A school or district must also test at least 95% of its students enrolled in the grade level tested for the school as a whole and for each required subgroup. In addition, the school and district must meet or exceed the other academic indicators set by the state: graduation rate for high schools and attendance rate for elementary and middle schools. These achievement goals must be reached for each subgroup that has a measurable group of students.

Education YES! uses several components that are interlinked to present a complete picture of performance at the school level. *Education YES!* is a broad set of measures that looks at school performance and student achievement in multiple ways. Measures of student achievement in Michigan’s school accreditation system include:

- Achievement status to measure how well a school is doing in educating its students.
- Achievement change to measure whether student achievement is improving or declining.
- Achievement growth (delayed until 2007-2008) to measure whether students are demonstrating at least one year of academic growth for each year of instruction.

In addition, the Indicators of School Performance measure investments that schools are making in improved student achievement, based on indicators that come from research and best practice.

Procedures for Using Assessment Data for Accountability

The school or district must attain the target achievement goal in English language arts (reading and writing) and mathematics or reduce the percentage of students in the non-proficient category (Partially Proficient and Not Proficient) of achievement by 10% (“safe harbor”). A school or district must also assess at least 95% of its students enrolled in the grade level tested for the school as a whole and for each required subgroup. In addition, the school must meet or exceed the other academic indicators set by the state: graduation rate for high schools of 80%, and attendance rate for elementary and middle schools of 85%. These achievement goals must be reached for each subgroup that has at least the minimum number of students in the group. The group size is the same for the school, school district, and the state as a whole. The subgroups are:

- Major Racial/Ethnic Groups
 - Black or African American
 - American Indian or Alaska Native
 - Asian American, Native Hawaiian or other Pacific Islander
 - Hispanic or Latino
 - White
 - Multiracial
- Students with Disabilities
- Limited English Proficient
- Economically Disadvantaged

Michigan’s minimum subgroup size is 30 students. For a district or school that enrolls more than 3,000 students, the minimum subgroup size will be 1% of enrollment, up to 200 students. An AYP determination will be made for all subgroups of 200 or more students.

It is the policy of the Michigan State Board of Education that all students participate in the state assessment program. The student’s status, in terms of enrollment for a full academic year, is not relevant to whether the student should be assessed. The federal No Child Left Behind Act requires that at least 95% of enrolled students be assessed. The number of students to be assessed is determined from the Michigan Student Data System (MSDS—formerly the SRSD), collected by the Center for Educational Performance and Information (CEPI). This is taken from the Fall (September) collection for grades 3-8 and from the Spring (February) collection for high schools.

The State Board of Education in Michigan has determined the AYP state targets (Annual Measurable Objectives) for the determination of AYP. The initial targets are based on assessment data from the 2001-02 administration of the MEAP tests and represent the percentage of proficient students in a public school at the 20th percentile of the State’s total enrollment among all schools ranked by the percentage of students at the proficient level. Michigan’s AYP targets for 2008-09 were:

- 65% - Elementary Mathematics
- 59% - Elementary English Language Arts
- 54% - Middle School Mathematics
- 54% - Middle School English Language Arts
- 55% - High School Mathematics
- 61% - High School English Language Arts

Because NCLB is currently focused on mathematics and English scores, the scores of all tested students must be used in the AYP determination for those subjects. Michigan has extended the grade range targets with separate targets for each grade, and by basing a school’s target on a weighted average of the statewide targets for the grades tested at the school. This procedure accounts for differences in performance standards across grade levels. The method also permits a single AYP determination for the school, through a comparison between student achievement and the school’s target.

Proficiency for AYP is based on the weighted sum of a proficiency index that is computed at each grade (3-11) counted for AYP at the school. Michigan did not change the approved AYP targets that were set previously. A set of grade level targets applicable to the 2008-09 school year has been developed and incorporated into the calculation of a Proficiency Index. The Proficiency Index is used to determine if a school, district, or student group meets the state AYP target.

A school, school district, or subgroup meets the state objective if the proficiency index is equal to or greater than zero (0). MDE will not determine or report AYP by grade. The grade level targets will be used to compute the proficiency index, which is aggregated across grades based on the school's configuration.

It is generally accepted that the standard error of measurement (SEM) varies across the range of student proficiencies and that individual score levels on any particular test could potentially have different degrees of measurement error associated with them. For this reason, it is generally useful to report not only a test level SEM estimate, but individual score level estimate as well. Individual score level estimates of error are commonly referred to as conditional standard errors of measurement (CSEM). The CSEM provides an estimate of error variability, conditional on the proficiency estimate (θ). In other words, it provides an error estimate, at each score point. According to the IRT model, there is typically more information in the middle of the θ score distribution, so the CSEM is usually smallest in this range. Michigan began use of the conditional standard errors of measurement in 2005-06 for its state assessments. Conditional standard errors of measurement are used to improve the accuracy of AYP determinations.

In addition the Indicators of School Performance measure investments that schools are making in improved student achievement, based on indicators that come from research and best practice. Scores on all three components of *Education YES!* have been converted to a common 100 point scale where: 90-100 A; 80-89 B; 70-79 C; 60-69 D; and 50-59 F. Grades of D and F are not used for the school's composite grade, where the labels D/Alert and Unaccredited are used.

Achievement Status

Achievement status is measured in English Language Arts and Mathematics at the elementary level. It includes Science and Social Studies at the middle school and high school levels. Achievement Status uses up to three years of comparable data from the Michigan Educational Assessment Program, the Michigan Merit Examination, or the MI-Access Assessments.

The method of computing achievement status uses students' scale scores on the Michigan assessments, as weighted by the performance level or category (1,2,3, or 4) assigned to each student's score. Scale score values at the chance level are substituted for values below the chance level because values below that point do not have valid information about the student's performance. A template is provided so that a school can paste in their assessment data to see how the values are derived. The weighted index is computed by following these steps:

1. Multiply each student's scale score by the performance level (i.e., 1100×2);
2. Sum of the resulting values resulting in the sum of the index values;
3. Sum of the performance levels or weights;
4. Divide the sum of the index values by the sum of the weights.

The intent of the weighted index is to encourage schools to place priority on improving the achievement of students that attain the lowest scores on the Michigan assessments.

Cut scores for the score ranges in achievement status were set by representative panels that assigned grades to selected schools. The cut scores were reviewed by the Accreditation Advisory Committee and approved by the State Board of Education. The Accreditation Advisory Committee, a group of five national experts, was appointed by the State Board of Education to advise the Board on the implementation of the *Education YES!* school accreditation.

Achievement Change

Achievement change uses up to five years of comparable assessment data to determine if student achievement in a school is improving at a rate fast enough to attain the goal of 100% proficiency in school year 2013-14, as required by the No Child Left Behind Act (NCLB). The change grade is derived from the average of up to three calculations of improvement rates (slopes) using the school's assessment data. Scores from assessments that are not comparable will not be placed on the same trend line. Achievement Change is based on the goal of 100% percent proficient in 2013-14, as set in NCLB. Achievement Change is computed by dividing the computed slope by the target slope, determining the percent of the target that the school has attained.

The linear regression methodology previously used to calculate Achievement Change was not used in 2006-07 for the elementary and middle school levels because scores from assessments that are not comparable cannot be placed on the same slope line. Multiple linear regression was used to predict each school's 2008-09 score based on the school's scores from 2007 and 2008. A prediction was made for each content area and grade level that was tested in previous years. The prediction was compared to the school's actual 2008-09 percent proficient. The Difference is computed as the (Actual – Predicted). The school's status score for each content area and grade range is adjusted as follows:

- Schools where the actual score exceeds the prediction plus 1.5 times the standard error of the estimate had a 15 point adjustment added to the achievement score for that content area;
- Schools where the actual score exceeds the prediction plus the standard error of the estimate had a 10 point adjustment added to the achievement score for that content area;
- Schools where the actual score is less than the prediction minus 1.5 times the standard error of the estimate had a 15 point deduction applied to the achievement score for that content area; and
- Schools where the actual score is less than the prediction minus the standard error of the estimate had a 10 point deduction applied to the achievement score for that content area.

The Achievement Change adjustment is calculated only if there are at least 10 students tested each year (2007, 2008 and 2009) in the content area and grade level.

A school district has the opportunity to appeal any data that affect its grade or AYP status if it has evidence that the data may be inaccurate. For example, the school district might identify corrected data regarding the number of students that were enrolled and should have been assessed. The Department of Education will do all that it can to correct errors that are brought to its attention. The purpose of the appeal window is to address substantive issues regarding the *Education YES!* grade or AYP status. The school district must cite specific data that are challenged in the appeal. Appeals that have no effect on the *Education YES!* grade or AYP status will not be considered.

The scoring and grading for the Indicators of School Performance are based on the school's self-rating of each component for each indicator. Each school team assigned the school a rating for each component, using the following scale:

- Systematically and Consistently Meeting Criteria;
- Progressing Toward Criteria;
- Starting to Meet Criteria; or
- Not Yet Meeting Criteria.

The ratings were scored on a scale where the number of possible points for each indicator is 36. The number of points possible for each component varies based on the number of components in the indicator. This method equally weights each indicator. For example, an indicator with 3 components receives 12 points per component whereas an indicator with 4 components receives 9 points per component. The possible score for all schools is 396 (11 indicators times 36 points). A single grade is assigned to the group of 11 indicators. The school's grade is based on the percentage of the possible points that the school could score for the total of all 11 indicators.

A "window" to update the School Self Assessments, including updating the self-rating and evidence for the Indicators of School Performance, ended in February, 2009. Beginning in 2004-05, the Department published both the school's self-rating and the evidence reported for each component. The school's self-rating for each component, and the evidence provided, is available in the online Report Card at <https://oeaa.state.mi.us/ayp/>.

The State Board of Education has approved a new School Improvement Framework that is intended to form the basis of revisions to the Indicators of School Performance for 2007-08. Draft rubrics have been developed and a pilot study was done in the spring of 2009.

Scores and grades are calculated for each content area for each school. The content areas remain the same, using only English Language Arts and Mathematics at the elementary level, and adding Science and Social Studies at the middle school and high school levels. The score and grade for each content area is based on the score for achievement status, as adjusted by averaging it with the score for achievement change.

The composite school grade is derived from the school scores and letter grades and the school's status in terms of Adequate Yearly Progress (AYP) under the federal No Child Left Behind Act. The weighting of the components of *Education YES!* in the composite grade has been as follows:

Table 14.1. Education YES! Composite Score Weighting

Component	Point Value
School Performance Indicators	33
Achievement Status	34
Achievement Change	33
Achievement Growth	
Total	100

The scores for each content area are averaged to calculate an achievement score and grade for each school. An achievement score for each content area has been computed by averaging the Status and Change (or adjusted Change) scores for a content area. A preliminary aggregate achievement score is derived by averaging the scores from each content area. The preliminary aggregate achievement score is weighted 67% and the School Self-Assessment (Indicator score) is weighted 33% in calculating the preliminary score and grade for a school.

In 2004-05, the State Board of Education approved a change to the *Education YES!* policy so that the school's indicator score cannot improve the school's composite score and grade by more than one letter grade more than the school's achievement grade. This means that a school that receives an "F" for achievement can receive a composite grade no higher than "D/Alert."

After the computation of a school’s composite grade for achievement described above, a final “filter” will be applied, consisting of the question of whether or not a school or district met or did not meet AYP. The answer to this question is an additional determining factor for a school’s final composite grade on the report card. A school that does not make AYP shall not be given a grade of “A.” A school that makes AYP shall not be listed as unaccredited. A school’s composite school grade will be used to prioritize assistance to underperforming schools and to prioritize interventions to improve student achievement.

Table 14.2. Unified Accountability for Michigan Schools

<i>Education YES! Composite Score</i>	90-100	B (iv)	A
	80-89	B (iv)	B (iv)
	70-79	C (iii)	C (iii)
	60-69	D/Alert (ii)	C (iii)
	50-59	Unaccredited (i)	D/Alert (ii)
		<i>Did Not Make AYP</i>	<i>Makes AYP</i>

(i) – (iv) Priorities for Assistance and Intervention

Schools that are labeled “A”, “B”, “C”, or “D / Alert” will be accredited. Schools that receive an “A” will be summary accredited. Schools that receive a “B”, “C”, or “D/Alert” will be in interim status. Unaccredited schools will also be labeled as such. Summary accreditation, interim status, and unaccredited are labels from Section 1280 of the Revised School Code.

Results of accountability analyses for 2008 and 2009 are summarized in Tables 14.3 through 14.6 below.

Table 14.3. Results of Accountability Analyses

Report on Michigan School AYP 2008 and 2009				
	Total Number of Schools	Elementary	Middle School	High Schools
Final Results for 2009				
Total Number of Schools	3,671	1,968	597	537
Made AYP	3,147 85.7%	1,839 93.4%	569 95.3%	382 71.1%
Did Not Make AYP	524 14.3%	129 6.6%	28 4.7%	155 28.9%
Final Results for 2008				
Total Number of Schools	3,761	2,058	609	513
Made AYP	3,003 79.8%	1,910 92.8%	543 89.2%	241 47.0%
Did Not Make AYP	758 20.2%	148 7.2%	66 10.8%	272 53.0%

Table 14.4. Report on School AYP 2007-08 and 2008-09

		2007-08	2008-09
Total Number of Schools Assigned AYP status		3,761	3,671
Total Number of Schools Not Making AYP		758	524
Percent of Schools Not Making AYP		20.0%	14.0%
Schools that make AYP using Interim Flexibility Option 1 - Students with Disabilities group		343	327
Schools Identified for Improvement		478	514
Schools Identified for Improvement by Phase	Identified for Improvement	185	165
	Identified for Improvement (SES)	77	111
	Corrective Action	62	69
	Restructuring	72	61
Schools with Graduation Rates under 80%		229	204
Schools not meeting Participation target by group*	All Students	150	147
	Black	97	88
	American Indian	0	0
	Asian American	0	1
	Hispanic	10	8
	White	68	48
	Multiracial	0	0
	Limited English Proficient	9	76
	Students with Disabilities	78	76
	Economically Disadvantaged	151	117
Schools not meeting Proficiency target by group*	All Students	324	189
	Black	111	57
	Asian American	1	1
	Hispanic	9	5
	White	40	15
	Limited English Proficient	9	2
	Students with Disabilities	236	123
	Economically Disadvantaged	186	93

*A school is counted in all student subgroups that did not make AYP. So, there are duplicated counts of schools.

Table 14.5. Report on Michigan District AYP 2008 and 2009

	Total Number of Districts	Number Met AYP	Percent Met AYP	Number Not Met AYP	Percent Not Met AYP
Final Results for 2009					
All School Districts	548	536	97.8%	12	2.2%
K-12 Districts	492	486	98.8%	6	1.2%
Charters	31	27	87.1%	4	12.9%
ISDs	26	24	92.3%	2	7.7%
Final Results for 2008					
All School Districts	546	525	96.2%	21	3.8%
K-12 Districts	492	483	98.2%	9	1.8%
Charters	33	25	75.8%	8	24.2%
ISDs	26	17	65.4%	9	34.6%

Table 14.6. State Accreditation Letter Grades 2008 and 2009

Grade	2008		2009	
	Number of Schools	Percent of Schools	Number of Schools	Percent of Schools
A	1,526	40.6%	1,680	45.8%
B	1,127	30.0%	1,002	27.3%
C	453	12.0%	382	10.4%
D-Alert	195	5.2%	142	3.9%
Unaccredited	8	0.2%	5	0.1%
No Grade	452	12.0%	460	12.5%
Total	3,761		3,671	

Chapter 15: State Summary Data

The summary data for the spring 2009 administration are presented in Table 15.1. For each content area, Table 15.1 presents the average score and the percentages of students falling into each of the four performance levels. Frequency distributions for the MME scale scores are presented in Figures 15.1 through 15.6, and in Tables 15.2 through 15.7. Tables 15.8 through 15.12 present the summary statistics for the item parameter estimates.

Table 15.1. Spring 2009 Michigan State Average Scores and Percentages in each Performance Level – All Examinees

Content Area	N	Average	Percentages within Performance Levels			
			Not Proficient	Partially Proficient	Proficient	Advanced
Reading	124,385	1106	16%	23%	58%	3%
Writing	125,579	1091	10%	45%	40%	4%
ELA	124,099	1099	12%	35%	50%	3%
Mathematics	123,284	1095	35%	15%	37%	13%
Science	123,873	1099	29%	15%	48%	8%
Social Studies	123,969	1127	8%	11%	39%	43%

Frequency Count

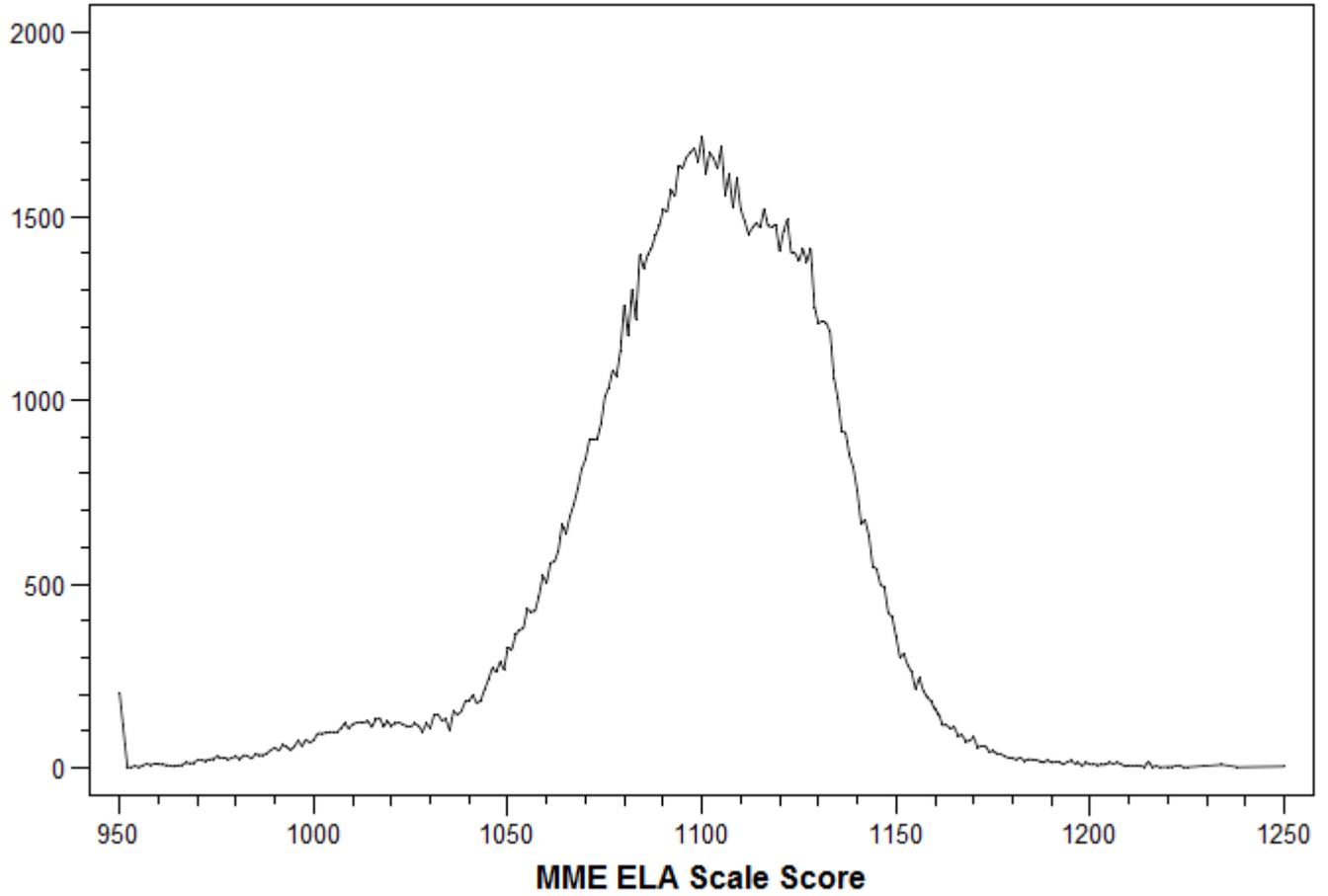


Figure 15.1. Frequency plot for MME Spring 2009 English Language Arts scale score total group – All forms included.

Frequency Count

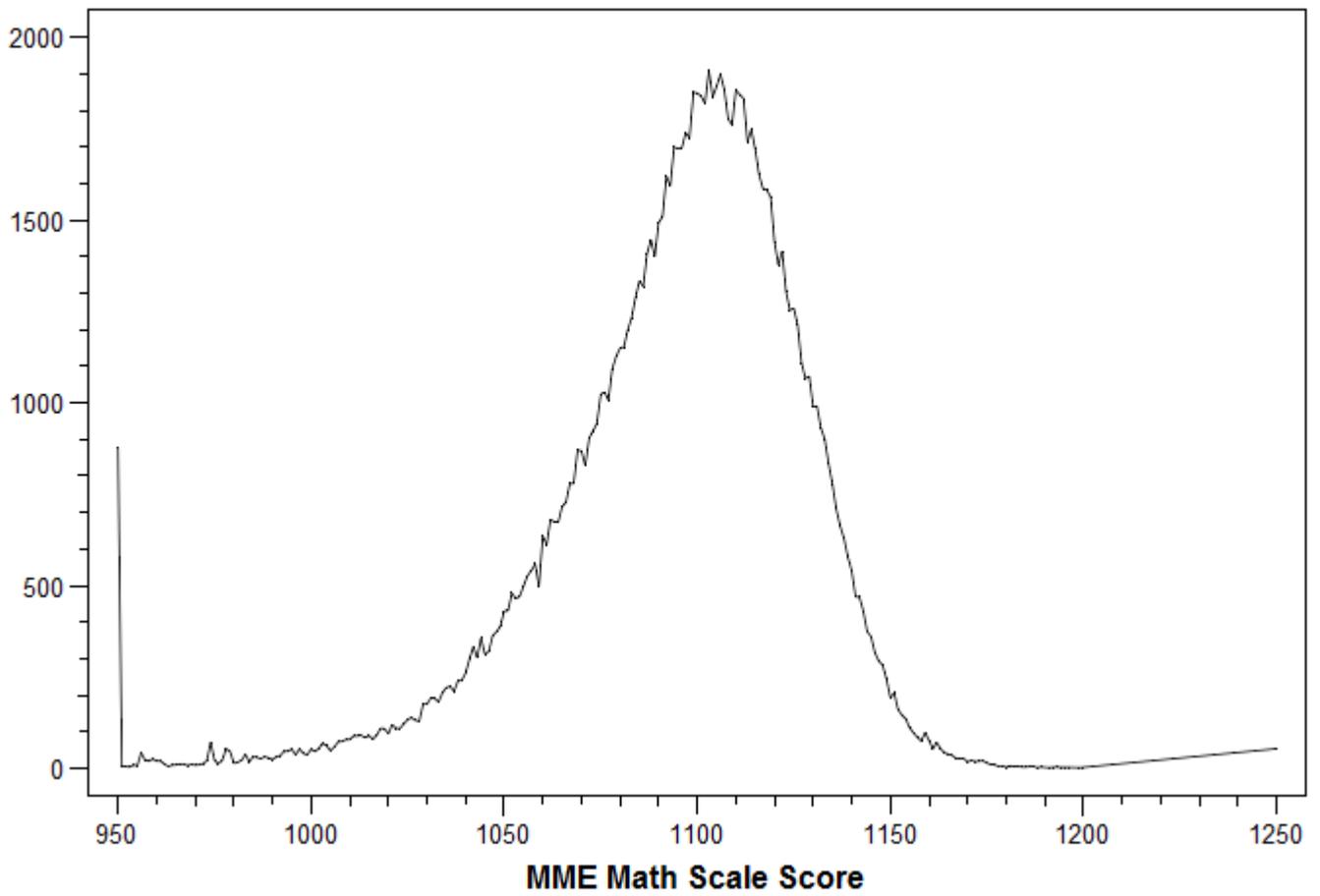


Figure 15.2. Frequency plot for MME Spring 2009 Mathematics scale score total group – All forms included.

Frequency Count

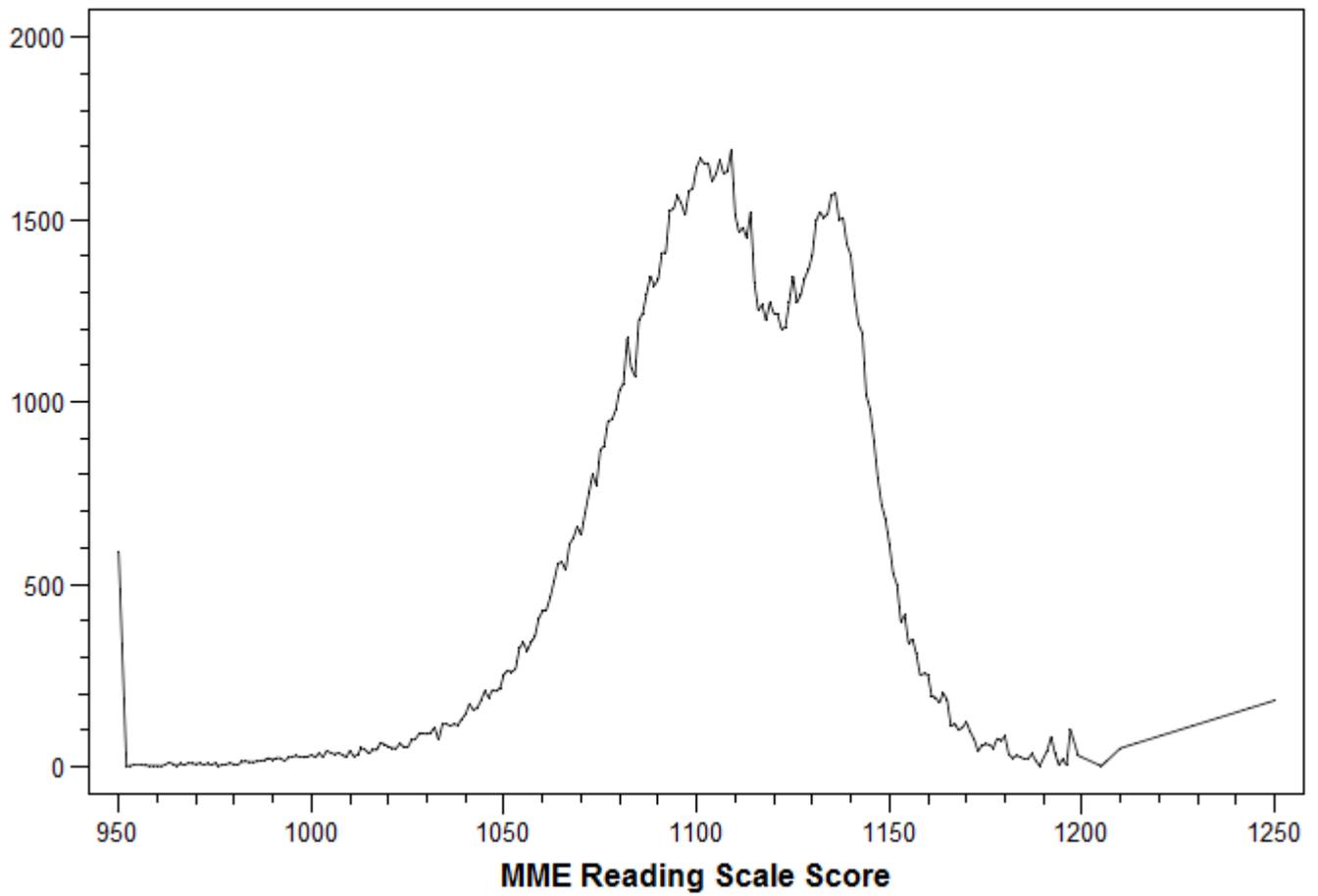


Figure 15.3. Frequency plot for MME Spring 2009 Reading scale score total group – All forms included.

Frequency Count

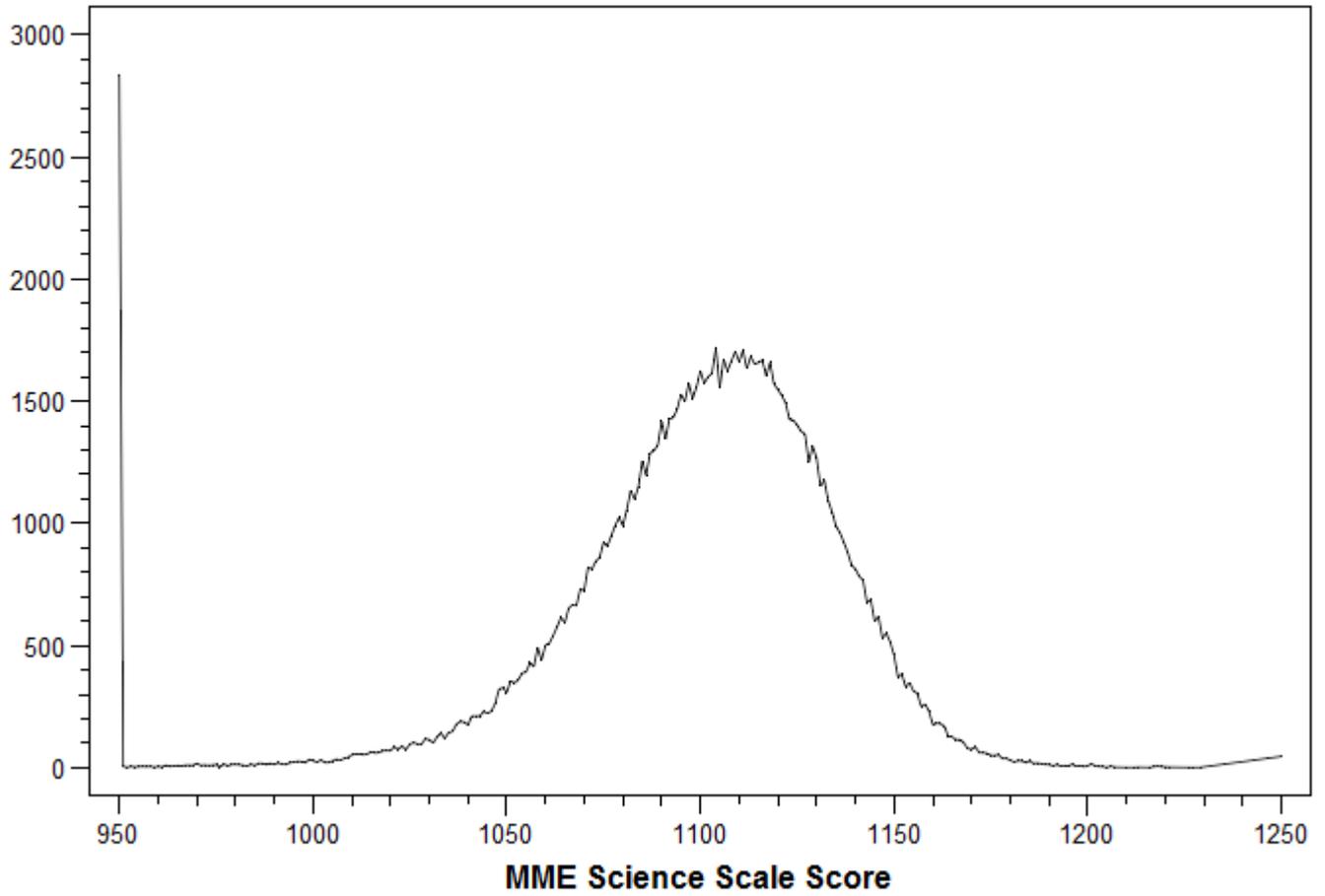


Figure 15.4. Frequency plot for MME Spring 2009 Science scale score total group – All forms included.

Frequency Count

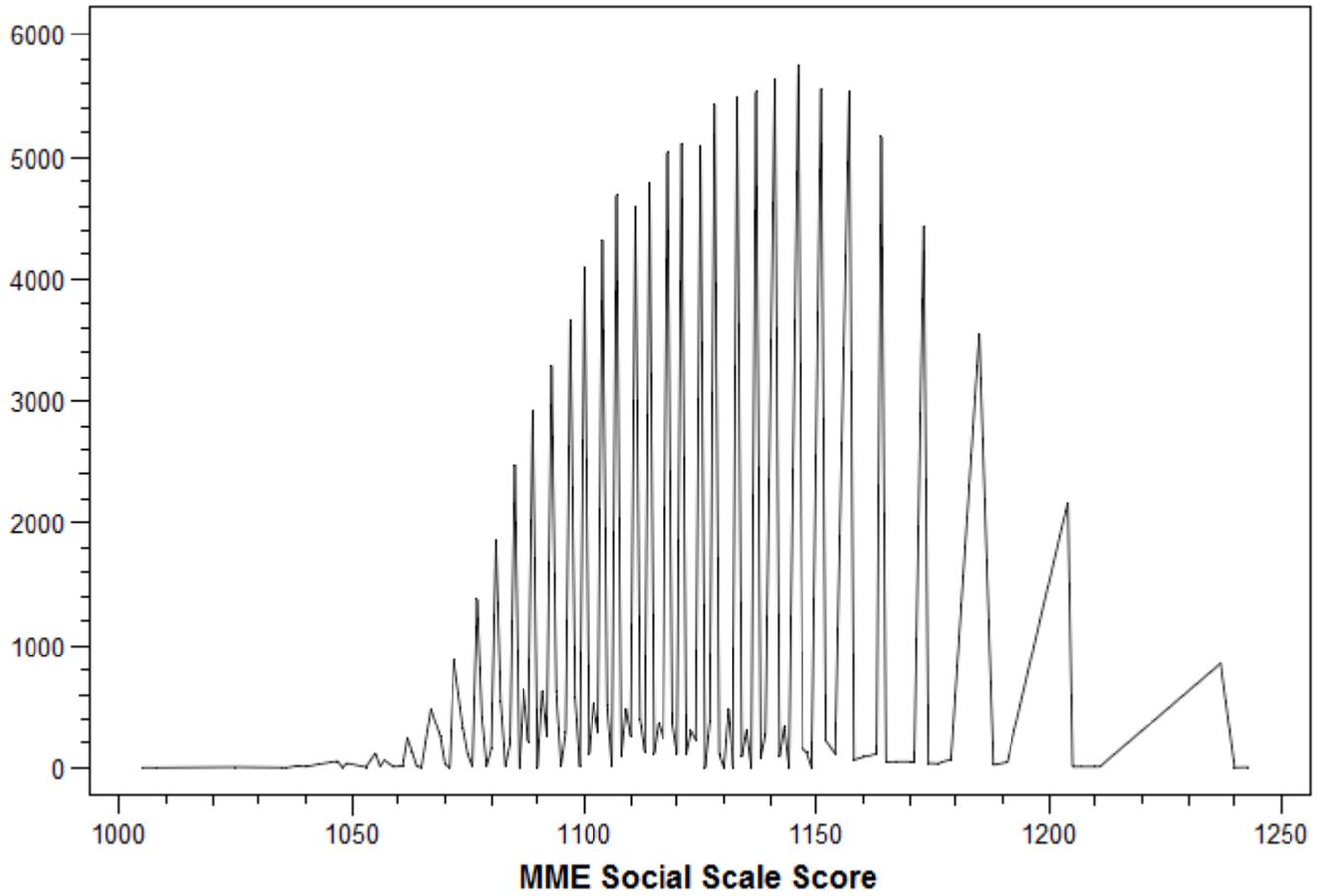


Figure 15.5. Frequency plot for MME Spring 2009 Social Studies scale score total group – All forms included.

Frequency Count

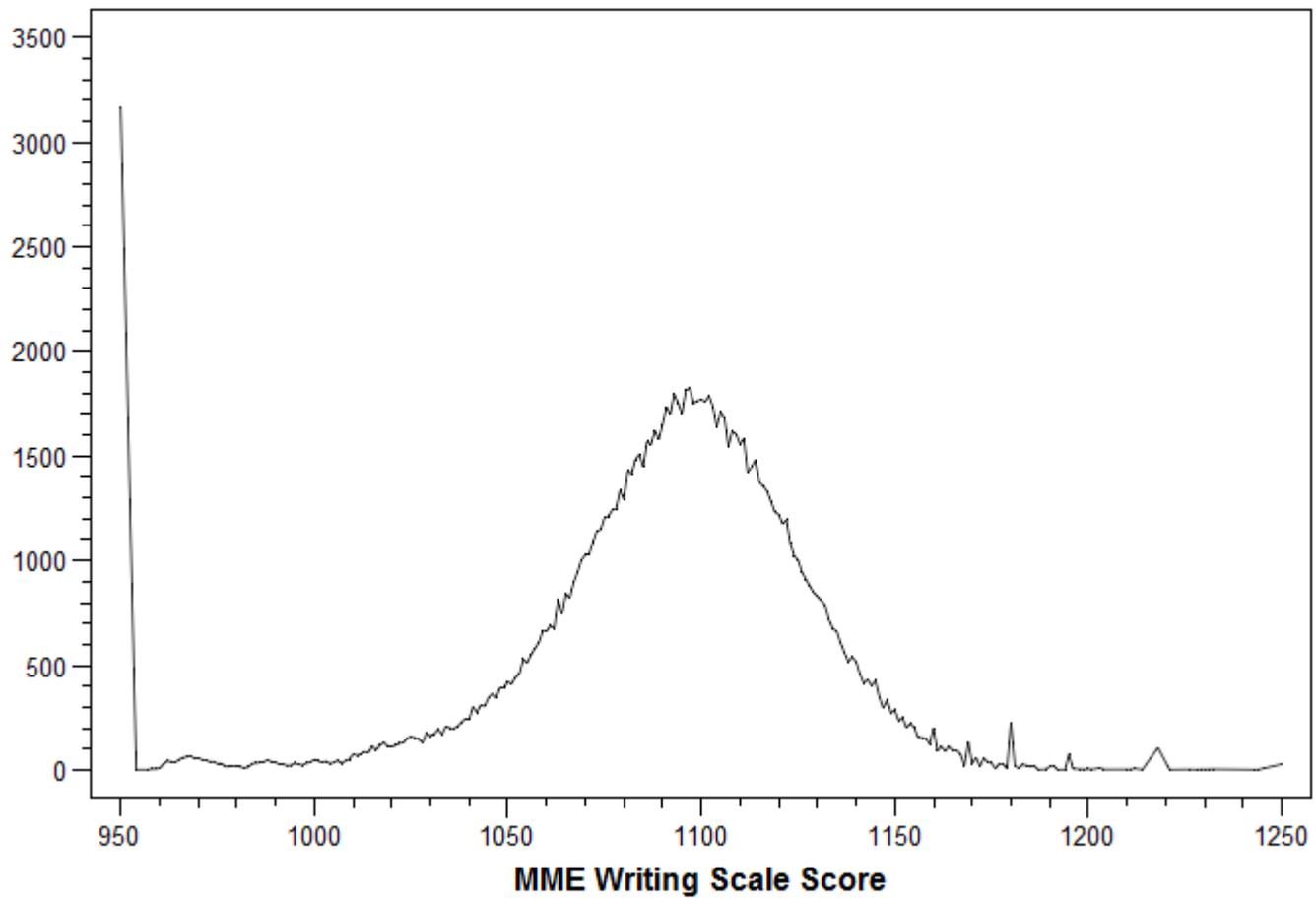


Figure 15.6. Frequency plot for MME Spring 2009 Writing scale score total group – All forms included.

Table 15.2. MME Spring 2009 English Language Arts Frequencies for Total Group -- All Forms Included

Scale Score	Frequency	Percent
950	203	0.16
952	1	0.00
953	1	0.00
954	4	0.00
955	1	0.00
956	6	0.00
957	10	0.01
958	7	0.01
959	10	0.01
960	9	0.01
961	8	0.01
962	7	0.01
963	5	0.00
964	3	0.00
965	6	0.00
966	6	0.00
967	13	0.01
968	12	0.01
969	12	0.01
970	20	0.02
971	20	0.02
972	18	0.01
973	22	0.02
974	23	0.02
975	30	0.02
976	27	0.02
977	27	0.02
978	22	0.02
979	27	0.02
980	30	0.02
981	23	0.02
982	34	0.03
983	31	0.02
984	24	0.02
985	37	0.03
986	34	0.03
987	34	0.03
988	39	0.03
989	48	0.04
990	55	0.04
991	45	0.04
992	62	0.05
993	56	0.05

994	48	0.04
995	57	0.05
996	74	0.06
997	58	0.05
998	77	0.06
999	67	0.05
1000	75	0.06
1001	93	0.07
1002	93	0.07
1003	94	0.08
1004	97	0.08
1005	97	0.08
1006	96	0.08
1007	107	0.09
1008	121	0.10
1009	106	0.09
1010	118	0.10
1011	122	0.10
1012	124	0.10
1013	123	0.10
1014	126	0.10
1015	111	0.09
1016	131	0.11
1017	135	0.11
1018	112	0.09
1019	126	0.10
1020	114	0.09
1021	120	0.10
1022	123	0.10
1023	117	0.09
1024	113	0.09
1025	111	0.09
1026	120	0.10
1027	114	0.09
1028	98	0.08
1029	120	0.10
1030	108	0.09
1031	143	0.12
1032	145	0.12
1033	128	0.10
1034	132	0.11
1035	101	0.08
1036	154	0.12
1037	144	0.12
1038	153	0.12

1039	179	0.14
1040	183	0.15
1041	197	0.16
1042	175	0.14
1043	183	0.15
1044	212	0.17
1045	240	0.19
1046	271	0.22
1047	260	0.21
1048	288	0.23
1049	267	0.22
1050	327	0.26
1051	321	0.26
1052	361	0.29
1053	375	0.30
1054	379	0.31
1055	432	0.35
1056	424	0.34
1057	426	0.34
1058	466	0.38
1059	522	0.42
1060	504	0.41
1061	558	0.45
1062	560	0.45
1063	590	0.48
1064	662	0.53
1065	638	0.51
1066	684	0.55
1067	717	0.58
1068	759	0.61
1069	813	0.66
1070	839	0.68
1071	892	0.72
1072	894	0.72
1073	894	0.72
1074	936	0.75
1075	1010	0.81
1076	1035	0.83
1077	1081	0.87
1078	1066	0.86
1079	1135	0.91
1080	1260	1.02
1081	1175	0.95
1082	1302	1.05
1083	1221	0.98
1084	1396	1.12
1085	1359	1.10
1086	1396	1.12

1087	1414	1.14
1088	1449	1.17
1089	1475	1.19
1090	1520	1.22
1091	1513	1.22
1092	1573	1.27
1093	1555	1.25
1094	1640	1.32
1095	1632	1.32
1096	1662	1.34
1097	1674	1.35
1098	1687	1.36
1099	1651	1.33
1100	1719	1.39
1101	1618	1.30
1102	1675	1.35
1103	1659	1.34
1104	1630	1.31
1105	1692	1.36
1106	1560	1.26
1107	1615	1.30
1108	1524	1.23
1109	1604	1.29
1110	1520	1.22
1111	1488	1.20
1112	1453	1.17
1113	1470	1.18
1114	1483	1.20
1115	1470	1.18
1116	1521	1.23
1117	1475	1.19
1118	1472	1.19
1119	1479	1.19
1120	1406	1.13
1121	1460	1.18
1122	1492	1.20
1123	1401	1.13
1124	1401	1.13
1125	1380	1.11
1126	1412	1.14
1127	1378	1.11
1128	1412	1.14
1129	1254	1.01
1130	1209	0.97
1131	1216	0.98
1132	1211	0.98
1133	1187	0.96
1134	1062	0.86

1135	1006	0.81
1136	916	0.74
1137	912	0.73
1138	850	0.68
1139	817	0.66
1140	754	0.61
1141	663	0.53
1142	675	0.54
1143	632	0.51
1144	548	0.44
1145	541	0.44
1146	497	0.40
1147	491	0.40
1148	420	0.34
1149	410	0.33
1150	357	0.29
1151	300	0.24
1152	310	0.25
1153	279	0.22
1154	261	0.21
1155	214	0.17
1156	244	0.20
1157	210	0.17
1158	191	0.15
1159	181	0.15
1160	158	0.13
1161	146	0.12
1162	115	0.09
1163	116	0.09
1164	106	0.09
1165	111	0.09
1166	86	0.07
1167	89	0.07
1168	71	0.06
1169	73	0.06
1170	83	0.07
1171	55	0.04
1172	58	0.05
1173	59	0.05
1174	43	0.03
1175	45	0.04
1176	38	0.03
1177	37	0.03
1178	30	0.02
1179	27	0.02
1180	25	0.02
1181	22	0.02

1182	28	0.02
1183	18	0.01
1184	23	0.02
1185	22	0.02
1186	20	0.02
1187	18	0.01
1188	13	0.01
1189	21	0.02
1190	13	0.01
1191	15	0.01
1192	16	0.01
1193	8	0.01
1194	15	0.01
1195	19	0.02
1196	11	0.01
1197	13	0.01
1198	5	0.00
1199	14	0.01
1200	8	0.01
1201	9	0.01
1202	7	0.01
1203	8	0.01
1204	8	0.01
1205	14	0.01
1206	10	0.01
1207	14	0.01
1208	9	0.01
1209	3	0.00
1210	6	0.00
1211	5	0.00
1212	6	0.00
1213	4	0.00
1214	2	0.00
1215	17	0.01
1216	1	0.00
1217	4	0.00
1218	1	0.00
1220	2	0.00
1221	2	0.00
1223	6	0.00
1224	1	0.00
1225	1	0.00
1234	8	0.01
1238	1	0.00
1250	3	0.00

Table 15.3 MME Spring 2009 Mathematics Frequencies for Total Group -- All Forms Included

Scale Score	Frequency	Percent
950	878	0.71
951	5	0.00
952	4	0.00
953	3	0.00
954	8	0.01
955	7	0.01
956	42	0.03
957	22	0.02
958	19	0.02
959	25	0.02
960	19	0.02
961	20	0.02
962	11	0.01
963	5	0.00
964	8	0.01
965	9	0.01
966	11	0.01
967	11	0.01
968	7	0.01
969	10	0.01
970	8	0.01
971	9	0.01
972	12	0.01
973	20	0.02
974	71	0.06
975	21	0.02
976	12	0.01
977	21	0.02
978	53	0.04
979	46	0.04
980	13	0.01
981	16	0.01
982	21	0.02
983	37	0.03
984	17	0.01
985	31	0.03
986	29	0.02
987	26	0.02
988	31	0.03
989	28	0.02
990	23	0.02
991	31	0.03
992	34	0.03

993	46	0.04
994	46	0.04
995	52	0.04
996	36	0.03
997	53	0.04
998	40	0.03
999	36	0.03
1000	52	0.04
1001	45	0.04
1002	54	0.04
1003	68	0.06
1004	61	0.05
1005	48	0.04
1006	57	0.05
1007	72	0.06
1008	73	0.06
1009	78	0.06
1010	78	0.06
1011	88	0.07
1012	88	0.07
1013	90	0.07
1014	83	0.07
1015	89	0.07
1016	79	0.06
1017	90	0.07
1018	109	0.09
1019	108	0.09
1020	94	0.08
1021	119	0.10
1022	109	0.09
1023	108	0.09
1024	121	0.10
1025	133	0.11
1026	138	0.11
1027	131	0.11
1028	128	0.10
1029	175	0.14
1030	175	0.14
1031	192	0.16
1032	191	0.15
1033	181	0.15
1034	206	0.17
1035	220	0.18
1036	224	0.18
1037	210	0.17

1038	238	0.19
1039	240	0.19
1040	260	0.21
1041	298	0.24
1042	332	0.27
1043	302	0.24
1044	356	0.29
1045	312	0.25
1046	319	0.26
1047	362	0.29
1048	373	0.30
1049	388	0.31
1050	430	0.35
1051	431	0.35
1052	479	0.39
1053	464	0.38
1054	470	0.38
1055	498	0.40
1056	526	0.43
1057	540	0.44
1058	559	0.45
1059	498	0.40
1060	635	0.52
1061	612	0.50
1062	681	0.55
1063	674	0.55
1064	673	0.55
1065	716	0.58
1066	728	0.59
1067	779	0.63
1068	779	0.63
1069	870	0.71
1070	869	0.70
1071	828	0.67
1072	903	0.73
1073	923	0.75
1074	943	0.76
1075	1023	0.83
1076	1030	0.84
1077	1007	0.82
1078	1094	0.89
1079	1127	0.91
1080	1150	0.93
1081	1153	0.94
1082	1200	0.97
1083	1233	1.00
1084	1288	1.04
1085	1333	1.08

1086	1319	1.07
1087	1409	1.14
1088	1446	1.17
1089	1402	1.14
1090	1494	1.21
1091	1509	1.22
1092	1620	1.31
1093	1595	1.29
1094	1700	1.38
1095	1696	1.38
1096	1698	1.38
1097	1739	1.41
1098	1726	1.40
1099	1850	1.50
1100	1848	1.50
1101	1840	1.49
1102	1819	1.48
1103	1910	1.55
1104	1838	1.49
1105	1866	1.51
1106	1899	1.54
1107	1858	1.51
1108	1778	1.44
1109	1761	1.43
1110	1856	1.51
1111	1844	1.50
1112	1833	1.49
1113	1715	1.39
1114	1748	1.42
1115	1695	1.37
1116	1625	1.32
1117	1587	1.29
1118	1583	1.28
1119	1564	1.27
1120	1441	1.17
1121	1378	1.12
1122	1415	1.15
1123	1304	1.06
1124	1254	1.02
1125	1259	1.02
1126	1213	0.98
1127	1110	0.90
1128	1067	0.87
1129	1073	0.87
1130	990	0.80
1131	991	0.80
1132	932	0.76
1133	901	0.73

1134	834	0.68
1135	778	0.63
1136	709	0.58
1137	663	0.54
1138	629	0.51
1139	578	0.47
1140	540	0.44
1141	469	0.38
1142	468	0.38
1143	430	0.35
1144	374	0.30
1145	360	0.29
1146	316	0.26
1147	293	0.24
1148	282	0.23
1149	244	0.20
1150	193	0.16
1151	206	0.17
1152	160	0.13
1153	145	0.12
1154	135	0.11
1155	110	0.09
1156	97	0.08
1157	84	0.07
1158	74	0.06
1159	97	0.08
1160	76	0.06
1161	53	0.04
1162	70	0.06
1163	53	0.04
1164	42	0.03
1165	38	0.03
1166	35	0.03
1167	26	0.02
1168	26	0.02
1169	26	0.02
1170	17	0.01
1171	21	0.02
1172	18	0.01
1173	19	0.02
1174	20	0.02
1175	14	0.01
1176	10	0.01
1177	9	0.01
1178	5	0.00
1179	6	0.00
1180	2	0.00
1181	6	0.00

1182	4	0.00
1183	5	0.00
1184	3	0.00
1185	3	0.00
1187	5	0.00
1188	1	0.00
1189	3	0.00
1191	1	0.00
1192	2	0.00
1193	3	0.00
1194	2	0.00
1195	2	0.00
1196	2	0.00
1199	1	0.00
1200	2	0.00
1250	53	0.04

Table 15.4. MME Spring 2009 Reading Frequencies for Total Group -- All Forms Included

Scale Score	Frequency	Percent
950	588	0.47
952	1	0.00
953	2	0.00
954	4	0.00
956	4	0.00
957	3	0.00
958	2	0.00
959	2	0.00
960	2	0.00
961	1	0.00
962	4	0.00
963	11	0.01
964	7	0.01
965	1	0.00
966	8	0.01
967	3	0.00
968	9	0.01
969	8	0.01
970	6	0.00
971	9	0.01
972	4	0.00
973	8	0.01
974	4	0.00
975	9	0.01
976	1	0.00
977	5	0.00
978	6	0.00
979	9	0.01
980	5	0.00
981	5	0.00
982	13	0.01
983	14	0.01
984	12	0.01
985	12	0.01
986	14	0.01
987	15	0.01
988	18	0.01
989	23	0.02
990	18	0.01
991	23	0.02
992	23	0.02
993	15	0.01
994	24	0.02

995	26	0.02
996	29	0.02
997	26	0.02
998	25	0.02
999	26	0.02
1000	32	0.03
1001	24	0.02
1002	37	0.03
1003	25	0.02
1004	43	0.03
1005	37	0.03
1006	33	0.03
1007	37	0.03
1008	32	0.03
1009	25	0.02
1010	43	0.03
1011	27	0.02
1012	32	0.03
1013	51	0.04
1014	45	0.04
1015	35	0.03
1016	46	0.04
1017	45	0.04
1018	66	0.05
1019	60	0.05
1020	54	0.04
1021	50	0.04
1022	49	0.04
1023	63	0.05
1024	53	0.04
1025	52	0.04
1026	72	0.06
1027	75	0.06
1028	90	0.07
1029	90	0.07
1030	89	0.07
1031	93	0.07
1032	107	0.09
1033	72	0.06
1034	115	0.09
1035	117	0.09
1036	111	0.09
1037	116	0.09
1038	113	0.09

1039	130	0.10
1040	143	0.11
1041	171	0.14
1042	155	0.12
1043	160	0.13
1044	180	0.14
1045	207	0.17
1046	189	0.15
1047	210	0.17
1048	206	0.17
1049	216	0.17
1050	253	0.20
1051	264	0.21
1052	259	0.21
1053	268	0.22
1054	325	0.26
1055	340	0.27
1056	317	0.25
1057	341	0.27
1058	356	0.29
1059	406	0.33
1060	426	0.34
1061	429	0.34
1062	464	0.37
1063	509	0.41
1064	557	0.45
1065	561	0.45
1066	538	0.43
1067	610	0.49
1068	627	0.50
1069	659	0.53
1070	634	0.51
1071	693	0.56
1072	748	0.60
1073	802	0.64
1074	771	0.62
1075	869	0.70
1076	880	0.71
1077	948	0.76
1078	951	0.76
1079	978	0.79
1080	1033	0.83
1081	1051	0.84
1082	1175	0.94
1083	1094	0.88
1084	1071	0.86
1085	1225	0.98
1086	1243	1.00

1087	1297	1.04
1088	1346	1.08
1089	1319	1.06
1090	1336	1.07
1091	1410	1.13
1092	1410	1.13
1093	1525	1.23
1094	1530	1.23
1095	1566	1.26
1096	1545	1.24
1097	1513	1.22
1098	1579	1.27
1099	1586	1.28
1100	1645	1.32
1101	1668	1.34
1102	1653	1.33
1103	1656	1.33
1104	1605	1.29
1105	1625	1.31
1106	1663	1.34
1107	1626	1.31
1108	1633	1.31
1109	1693	1.36
1110	1513	1.22
1111	1465	1.18
1112	1479	1.19
1113	1450	1.17
1114	1521	1.22
1115	1327	1.07
1116	1251	1.01
1117	1267	1.02
1118	1224	0.98
1119	1273	1.02
1120	1243	1.00
1121	1241	1.00
1122	1200	0.96
1123	1204	0.97
1124	1272	1.02
1125	1345	1.08
1126	1273	1.02
1127	1293	1.04
1128	1339	1.08
1129	1363	1.10
1130	1402	1.13
1131	1497	1.20
1132	1521	1.22
1133	1506	1.21
1134	1516	1.22

1135	1568	1.26
1136	1573	1.26
1137	1500	1.21
1138	1506	1.21
1139	1432	1.15
1140	1403	1.13
1141	1292	1.04
1142	1213	0.98
1143	1190	0.96
1144	1019	0.82
1145	981	0.79
1146	894	0.72
1147	792	0.64
1148	718	0.58
1149	678	0.55
1150	612	0.49
1151	529	0.43
1152	499	0.40
1153	397	0.32
1154	415	0.33
1155	338	0.27
1156	350	0.28
1157	310	0.25
1158	250	0.20
1159	255	0.21
1160	252	0.20
1161	193	0.16
1162	189	0.15
1163	175	0.14
1164	201	0.16
1165	183	0.15
1166	110	0.09
1167	117	0.09
1168	99	0.08

1169	107	0.09
1170	121	0.10
1171	95	0.08
1172	75	0.06
1173	42	0.03
1174	56	0.05
1175	61	0.05
1176	60	0.05
1177	50	0.04
1178	77	0.06
1179	71	0.06
1180	83	0.07
1181	33	0.03
1182	22	0.02
1183	30	0.02
1184	26	0.02
1185	20	0.02
1186	20	0.02
1187	35	0.03
1188	16	0.01
1189	1	0.00
1191	44	0.04
1192	81	0.07
1193	36	0.03
1194	4	0.00
1195	20	0.02
1196	4	0.00
1197	101	0.08
1199	30	0.02
1205	1	0.00
1210	50	0.04
1250	181	0.15

Table 15.5. MME Spring 2009 Science Frequencies for Total Group—All Forms Included

Scale Score	Frequency	Percent
950	2835	2.29
951	5	0.00
952	1	0.00
953	5	0.00
954	3	0.00
955	4	0.00
956	5	0.00
957	4	0.00
958	5	0.00
959	2	0.00
960	4	0.00
961	3	0.00
962	10	0.01
963	5	0.00
964	6	0.00
965	9	0.01
966	6	0.00
967	10	0.01
968	10	0.01
969	9	0.01
970	18	0.01
971	9	0.01
972	11	0.01
973	9	0.01
974	10	0.01
975	12	0.01
976	3	0.00
977	15	0.01
978	7	0.01
979	13	0.01
980	13	0.01
981	12	0.01
982	10	0.01
983	6	0.00
984	14	0.01
985	9	0.01
986	18	0.01
987	17	0.01
988	12	0.01
989	18	0.01
990	13	0.01
991	20	0.02
992	14	0.01

993	13	0.01
994	20	0.02
995	23	0.02
996	26	0.02
997	25	0.02
998	20	0.02
999	34	0.03
1000	29	0.02
1001	22	0.02
1002	32	0.03
1003	21	0.02
1004	21	0.02
1005	27	0.02
1006	35	0.03
1007	31	0.03
1008	37	0.03
1009	42	0.03
1010	52	0.04
1011	57	0.05
1012	55	0.04
1013	53	0.04
1014	55	0.04
1015	65	0.05
1016	60	0.05
1017	62	0.05
1018	69	0.06
1019	70	0.06
1020	69	0.06
1021	85	0.07
1022	75	0.06
1023	88	0.07
1024	74	0.06
1025	97	0.08
1026	102	0.08
1027	97	0.08
1028	96	0.08
1029	118	0.10
1030	115	0.09
1031	101	0.08
1032	125	0.10
1033	143	0.12
1034	119	0.10
1035	144	0.12
1036	148	0.12
1037	177	0.14

1038	189	0.15
1039	187	0.15
1040	175	0.14
1041	211	0.17
1042	211	0.17
1043	211	0.17
1044	230	0.19
1045	223	0.18
1046	230	0.19
1047	261	0.21
1048	321	0.26
1049	327	0.26
1050	308	0.25
1051	355	0.29
1052	346	0.28
1053	360	0.29
1054	387	0.31
1055	392	0.32
1056	429	0.35
1057	414	0.33
1058	491	0.40
1059	441	0.36
1060	501	0.40
1061	509	0.41
1062	541	0.44
1063	576	0.46
1064	615	0.50
1065	593	0.48
1066	653	0.53
1067	666	0.54
1068	663	0.54
1069	732	0.59
1070	724	0.58
1071	821	0.66
1072	812	0.66
1073	845	0.68
1074	860	0.69
1075	923	0.75
1076	907	0.73
1077	949	0.77
1078	991	0.80
1079	1024	0.83
1080	991	0.80
1081	1055	0.85
1082	1134	0.92
1083	1100	0.89
1084	1150	0.93
1085	1255	1.01

1086	1197	0.97
1087	1287	1.04
1088	1297	1.05
1089	1320	1.07
1090	1417	1.14
1091	1345	1.09
1092	1428	1.15
1093	1433	1.16
1094	1467	1.18
1095	1528	1.23
1096	1500	1.21
1097	1573	1.27
1098	1512	1.22
1099	1558	1.26
1100	1624	1.31
1101	1574	1.27
1102	1600	1.29
1103	1614	1.30
1104	1719	1.39
1105	1554	1.25
1106	1673	1.35
1107	1625	1.31
1108	1664	1.34
1109	1704	1.38
1110	1662	1.34
1111	1708	1.38
1112	1634	1.32
1113	1683	1.36
1114	1653	1.33
1115	1658	1.34
1116	1668	1.35
1117	1604	1.29
1118	1658	1.34
1119	1571	1.27
1120	1548	1.25
1121	1524	1.23
1122	1494	1.21
1123	1425	1.15
1124	1423	1.15
1125	1402	1.13
1126	1377	1.11
1127	1366	1.10
1128	1251	1.01
1129	1315	1.06
1130	1268	1.02
1131	1152	0.93
1132	1176	0.95
1133	1093	0.88

1134	1047	0.85
1135	989	0.80
1136	963	0.78
1137	924	0.75
1138	883	0.71
1139	830	0.67
1140	814	0.66
1141	783	0.63
1142	768	0.62
1143	675	0.54
1144	688	0.56
1145	598	0.48
1146	620	0.50
1147	531	0.43
1148	551	0.44
1149	515	0.42
1150	461	0.37
1151	372	0.30
1152	383	0.31
1153	332	0.27
1154	348	0.28
1155	314	0.25
1156	304	0.25
1157	251	0.20
1158	260	0.21
1159	232	0.19
1160	176	0.14
1161	182	0.15
1162	180	0.15
1163	169	0.14
1164	128	0.10
1165	126	0.10
1166	114	0.09
1167	115	0.09
1168	102	0.08
1169	78	0.06
1170	75	0.06
1171	85	0.07
1172	61	0.05
1173	61	0.05
1174	59	0.05
1175	50	0.04
1176	50	0.04
1177	53	0.04
1178	37	0.03
1179	38	0.03
1180	30	0.02
1181	24	0.02

1182	28	0.02
1183	28	0.02
1184	20	0.02
1185	28	0.02
1186	17	0.01
1187	19	0.02
1188	16	0.01
1189	13	0.01
1190	13	0.01
1191	7	0.01
1192	13	0.01
1193	6	0.00
1194	9	0.01
1195	6	0.00
1196	15	0.01
1197	5	0.00
1198	9	0.01
1199	4	0.00
1200	7	0.01
1201	16	0.01
1202	6	0.00
1203	5	0.00
1204	4	0.00
1205	1	0.00
1206	6	0.00
1207	2	0.00
1210	1	0.00
1212	1	0.00
1213	2	0.00
1216	1	0.00
1217	4	0.00
1218	6	0.00
1219	4	0.00
1220	3	0.00
1221	2	0.00
1226	1	0.00
1227	1	0.00
1229	1	0.00
1250	46	0.04

Table 15.6. MME Spring 2009 Social Studies Frequencies for Total Group—All Forms Included

Scale Score	Frequency	Percent
1005	2	0.00
1008	1	0.00
1025	6	0.00
1035	2	0.00
1036	1	0.00
1038	14	0.01
1040	10	0.01
1047	54	0.04
1048	2	0.00
1049	37	0.03
1053	6	0.00
1055	118	0.10
1056	12	0.01
1057	66	0.05
1059	8	0.01
1061	16	0.01
1062	239	0.19
1063	131	0.11
1064	11	0.01
1065	6	0.00
1067	480	0.39
1069	255	0.21
1070	29	0.02
1071	3	0.00
1072	880	0.71
1074	316	0.25
1075	111	0.09
1076	12	0.01
1077	1375	1.11
1078	423	0.34
1079	14	0.01
1080	151	0.12
1081	1857	1.50
1082	547	0.44
1083	14	0.01
1084	197	0.16
1085	2480	2.00
1086	5	0.00
1087	639	0.52
1088	214	0.17
1089	2919	2.35
1090	2	0.00
1091	627	0.51
1092	261	0.21
1093	3291	2.65
1094	621	0.50
1095	14	0.01
1096	288	0.23
1097	3664	2.96
1098	569	0.46
1099	9	0.01
1100	4095	3.30
1101	116	0.09
1102	527	0.43
1103	288	0.23
1104	4325	3.49
1105	510	0.41
1106	12	0.01
1107	4684	3.78
1108	95	0.08
1109	488	0.39
1110	255	0.21
1111	4585	3.70
1112	398	0.32
1113	130	0.10
1114	4785	3.86
1115	117	0.09
1116	367	0.30
1117	235	0.19
1118	5038	4.06
1119	362	0.29
1120	114	0.09
1121	5112	4.12
1122	115	0.09
1123	296	0.24
1124	227	0.18
1125	5088	4.10
1126	1	0.00
1127	377	0.30
1128	5424	4.38
1129	109	0.09
1130	1	0.00
1131	481	0.39
1132	5	0.00
1133	5485	4.42
1134	89	0.07
1135	305	0.25
1136	2	0.00

1137	5535	4.46
1138	85	0.07
1139	272	0.22
1141	5636	4.55
1142	92	0.07
1143	330	0.27
1144	5	0.00
1146	5756	4.64
1147	156	0.13
1148	120	0.10
1149	1	0.00
1151	5560	4.48
1152	220	0.18
1154	115	0.09
1157	5548	4.48
1158	63	0.05
1160	89	0.07
1163	111	0.09
1164	5165	4.17
1165	46	0.04
1167	50	0.04
1170	50	0.04
1171	42	0.03
1173	4437	3.58
1174	38	0.03
1176	34	0.03
1179	68	0.05
1185	3549	2.86
1188	23	0.02
1191	46	0.04
1204	2174	1.75
1205	9	0.01
1207	9	0.01
1210	9	0.01
1211	10	0.01
1237	857	0.69
1240	2	0.00
1243	5	0.00

Table 15.7. MME Spring 2009 Writing Frequencies for Total Group—All Forms Included

Scale Score	Frequency	Percent
950	3168	2.52
954	1	0.00
957	1	0.00
958	4	0.00
959	4	0.00
960	7	0.01
961	27	0.02
962	42	0.03
963	39	0.03
964	35	0.03
965	47	0.04
966	58	0.05
967	61	0.05
968	67	0.05
969	53	0.04
970	56	0.04
971	48	0.04
972	46	0.04
973	38	0.03
974	39	0.03
975	28	0.02
976	31	0.02
977	15	0.01
978	16	0.01
979	19	0.02
980	18	0.01
981	14	0.01
982	9	0.01
983	13	0.01
984	30	0.02
985	32	0.03
986	35	0.03
987	38	0.03
988	47	0.04
989	37	0.03
990	35	0.03
991	27	0.02
992	26	0.02
993	20	0.02
994	22	0.02
995	32	0.03
996	26	0.02
997	22	0.02

998	32	0.03
999	36	0.03
1000	49	0.04
1001	42	0.03
1002	35	0.03
1003	41	0.03
1004	30	0.02
1005	34	0.03
1006	45	0.04
1007	31	0.02
1008	42	0.03
1009	50	0.04
1010	74	0.06
1011	68	0.05
1012	75	0.06
1013	87	0.07
1014	83	0.07
1015	112	0.09
1016	95	0.08
1017	118	0.09
1018	131	0.10
1019	111	0.09
1020	110	0.09
1021	117	0.09
1022	126	0.10
1023	130	0.10
1024	151	0.12
1025	157	0.13
1026	153	0.12
1027	147	0.12
1028	130	0.10
1029	175	0.14
1030	161	0.13
1031	172	0.14
1032	192	0.15
1033	170	0.14
1034	208	0.17
1035	200	0.16
1036	196	0.16
1037	208	0.17
1038	227	0.18
1039	244	0.19
1040	240	0.19
1041	301	0.24
1042	275	0.22

1043	309	0.25
1044	304	0.24
1045	344	0.27
1046	364	0.29
1047	346	0.28
1048	393	0.31
1049	393	0.31
1050	421	0.34
1051	411	0.33
1052	443	0.35
1053	461	0.37
1054	529	0.42
1055	512	0.41
1056	550	0.44
1057	580	0.46
1058	607	0.48
1059	662	0.53
1060	663	0.53
1061	688	0.55
1062	677	0.54
1063	811	0.65
1064	745	0.59
1065	842	0.67
1066	821	0.65
1067	895	0.71
1068	943	0.75
1069	1001	0.80
1070	1027	0.82
1071	1030	0.82
1072	1089	0.87
1073	1141	0.91
1074	1148	0.91
1075	1206	0.96
1076	1212	0.97
1077	1243	0.99
1078	1249	0.99
1079	1335	1.06
1080	1294	1.03
1081	1429	1.14
1082	1415	1.13
1083	1484	1.18
1084	1507	1.20
1085	1449	1.15
1086	1570	1.25
1087	1554	1.24
1088	1621	1.29
1089	1579	1.26
1090	1651	1.31

1091	1736	1.38
1092	1702	1.36
1093	1796	1.43
1094	1753	1.40
1095	1705	1.36
1096	1813	1.44
1097	1826	1.45
1098	1752	1.40
1099	1764	1.40
1100	1769	1.41
1101	1761	1.40
1102	1786	1.42
1103	1737	1.38
1104	1639	1.31
1105	1713	1.36
1106	1683	1.34
1107	1547	1.23
1108	1618	1.29
1109	1604	1.28
1110	1558	1.24
1111	1583	1.26
1112	1422	1.13
1113	1447	1.15
1114	1480	1.18
1115	1376	1.10
1116	1355	1.08
1117	1333	1.06
1118	1284	1.02
1119	1233	0.98
1120	1221	0.97
1121	1176	0.94
1122	1194	0.95
1123	1083	0.86
1124	1023	0.81
1125	1003	0.80
1126	946	0.75
1127	912	0.73
1128	880	0.70
1129	848	0.68
1130	829	0.66
1131	810	0.65
1132	787	0.63
1133	716	0.57
1134	675	0.54
1135	663	0.53
1136	607	0.48
1137	560	0.45
1138	517	0.41

1139	539	0.43
1140	518	0.41
1141	458	0.36
1142	414	0.33
1143	432	0.34
1144	402	0.32
1145	426	0.34
1146	349	0.28
1147	295	0.23
1148	333	0.27
1149	272	0.22
1150	286	0.23
1151	236	0.19
1152	249	0.20
1153	202	0.16
1154	221	0.18
1155	203	0.16
1156	155	0.12
1157	153	0.12
1158	150	0.12
1159	123	0.10
1160	200	0.16
1161	91	0.07
1162	111	0.09
1163	88	0.07
1164	111	0.09
1165	90	0.07
1166	93	0.07
1167	75	0.06
1168	21	0.02
1169	134	0.11
1170	30	0.02
1171	58	0.05
1172	17	0.01
1173	55	0.04
1174	37	0.03
1175	32	0.03
1176	6	0.00
1177	29	0.02

1178	28	0.02
1179	5	0.00
1180	228	0.18
1181	18	0.01
1182	8	0.01
1183	25	0.02
1184	21	0.02
1185	15	0.01
1186	20	0.02
1187	1	0.00
1189	2	0.00
1190	15	0.01
1191	21	0.02
1192	1	0.00
1194	1	0.00
1195	74	0.06
1196	5	0.00
1198	3	0.00
1199	2	0.00
1200	4	0.00
1201	2	0.00
1203	8	0.01
1204	2	0.00
1210	2	0.00
1211	2	0.00
1212	6	0.00
1214	1	0.00
1218	106	0.08
1221	1	0.00
1226	2	0.00
1228	1	0.00
1229	2	0.00
1230	2	0.00
1231	2	0.00
1232	3	0.00
1243	1	0.00
1244	1	0.00
1250	27	0.02

Table 15.8. Mean and SD of Item Parameter Estimates for Mathematics

		2009 Spring Mathematics		
		a	b	c
		Initial Form 1		
mean		1.435	-0.058	0.193
SD		0.585	1.005	0.078
		Initial Form 2		
mean		1.430	-0.002	0.190
SD		0.574	1.021	0.076
		Initial Form 3		
mean		1.441	-0.055	0.193
SD		0.576	0.980	0.080
		Initial Form 4		
mean		1.413	-0.052	0.195
SD		0.567	0.991	0.079
		Initial Form 5		
mean		1.451	-0.012	0.192
SD		0.580	1.000	0.079
		Initial Form 6		
mean		1.489	-0.051	0.196
SD		0.583	0.981	0.081
		Initial Form 7		
mean		1.463	-0.024	0.191
SD		0.573	0.991	0.078
		Initial Form 8		
mean		1.462	-0.032	0.192
SD		0.561	0.990	0.078
		Initial Form 9		
mean		1.466	-0.036	0.197
SD		0.564	0.991	0.084
		Initial Form 10		
mean		1.434	-0.056	0.188
SD		0.558	0.983	0.078
		Makeup Form		
mean		1.504	0.116	0.201
SD		0.610	0.895	0.062
		Accommodated Form		
mean		1.622	0.031	0.197
SD		0.567	1.065	0.080

Table 15.9. Mean and SD of Item Parameter Estimates for Reading

		2009 Spring Reading		
		a	b	c
		Initial Form		
mean		1.124	0.118	0.200
SD		0.800	1.270	0.079
		Makeup Form		
mean		0.962	-0.026	0.207
SD		0.396	1.307	0.082
		Accommodated Form		
mean		0.998	0.076	0.210
SD		0.411	1.434	0.066

Table 15.10. Mean and SD of Item Parameter Estimates for Science

		2009 Spring Science		
		a	b	c
		Initial Form 1		
mean		0.826	0.353	0.207
SD		0.285	0.955	0.085
		Initial Form 2		
mean		0.826	0.430	0.219
SD		0.266	1.007	0.080
		Initial Form 3		
mean		0.792	0.335	0.208
SD		0.240	1.038	0.080
		Initial Form 4		
mean		0.829	0.369	0.206
SD		0.295	1.074	0.079
		Initial Form 5		
mean		0.821	0.516	0.210
SD		0.293	0.971	0.078
		Initial Form 6		
mean		0.837	0.471	0.218
SD		0.226	0.972	0.085
		Initial Form 7		
mean		0.821	0.333	0.206
SD		0.302	1.008	0.083
		Initial Form 8		
mean		0.815	0.439	0.215
SD		0.230	0.933	0.079
		Initial Form 9		
mean		0.828	0.360	0.214
SD		0.215	0.934	0.081
		Initial Form 10		
mean		0.833	0.360	0.212
SD		0.278	1.079	0.081
		Makeup Form		
mean		0.880	0.573	0.224
SD		0.277	1.106	0.060
		Accommodated Form		
mean		0.898	0.602	0.227
SD		0.357	1.044	0.081

Table 15.11. Mean and SD of Item Parameter Estimates for Writing

		2009 Spring Writing MC Items		
		a	b	c
mean SD	Initial Form			
		0.961 0.279	0.461 0.782	0.219 0.114
mean SD	Makeup Form			
		0.855 0.252	0.254 0.920	0.209 0.083
mean SD	Accommodated Form			
		0.907 0.273	0.260 0.759	0.201 0.050

		2009 Spring Writing CR Items						
		a	b	Step1	Step2	Step3	Step4	Step5
mean SD	Initial Form							
		0.441 0.001	0.798 0.018	3.539 0.031	2.878 0.009	0.956 0.018	-2.460 0.032	-4.913 0.091
mean SD	Makeup Form							
		0.504 0.000	0.824 0.032	3.050 0.004	2.253 0.037	0.687 0.041	-2.094 0.030	-3.896 0.052
mean SD	Accommodated Form							
		0.392 0.007	0.945 0.047	3.121 0.058	2.245 0.030	0.313 0.019	-2.504 0.068	-3.175 0.039

Table 15.12. Mean and SD of Item Parameter Estimates for Social Studies

		2009 Spring Social Studies	
		b	
mean SD	Initial Form		
		-0.1603 0.7605	
mean SD	Makeup Form		
		0.1062 0.6431	
mean SD	Accommodated Form		
		-0.0878 0.7759	

Chapter 16: MME Scale Score History

The first MME assessment was administered statewide in spring 2007. For each content area, Tables 16.1 to 16.3 present the average scores and the percentages of students in each of the four performance levels. Yearly samples from a state may vary, and changes to a testing program (such as the inclusion of a new measurement instrument like *Locating Information* in 2009) contribute to annual scale score means. Therefore, changes in means across years need to be considered in context. OEAA encourages those interested in using MME scale scores to reference the informational materials contained in this technical report, as well as information posted on the MME website (www.michigan.gov/mme). Additionally, the Guide to Reports should be used when using or referencing assessment data (see http://www.michigan.gov/documents/mde/9_MME_2009_Guide_to_Reports_283213_7_283862_7.pdf).

Table 16.1. Spring 2007 Michigan State Average Scores and Percentages in each Performance Level

Content Area	N	Average	Percentages within Performance Levels			
			Apprentice	Basic	Met Standards	Exceeded Standards
Reading	113,956	1104	17%	24%	58%	2%
Writing	111,479	1090	10%	50%	38%	2%
ELA	111,000	1098	12%	37%	49%	2%
Mathematics	113,839	1093	38%	16%	37%	10%
Science	113,630	1098	28%	16%	50%	6%
Social Studies	113,718	1124	7%	9%	42%	41%

Table 16.2. Spring 2008 Michigan State Average Scores and Percentages in each Performance Level

Content Area	N	Average	Percentages within Performance Levels			
			Not Proficient	Partially Proficient	Proficient	Advanced
Reading	130,226	1106	17%	21%	60%	3%
Writing	129,400	1090	11%	48%	39%	3%
ELA	128,818	1099	13%	34%	50%	2%
Mathematics	129,803	1093	38%	16%	36%	10%
Science	129,691	1099	27%	16%	51%	6%
Social Studies	130,957	1123	7%	13%	39%	41%

Table 16.3. Spring 2009 Michigan State Average Scores and Percentages in each Performance Level

Content Area	N	Average	Percentages within Performance Levels			
			Not Proficient	Partially Proficient	Proficient	Advanced
Reading	124,385	1106	16%	23%	58%	3%
Writing	125,579	1091	10%	45%	40%	4%
ELA	124,099	1099	12%	35%	50%	3%
Mathematics	123,284	1095	35%	15%	37%	13%
Science	123,873	1099	29%	15%	48%	8%
Social Studies	123,969	1127	8%	11%	39%	43%

References

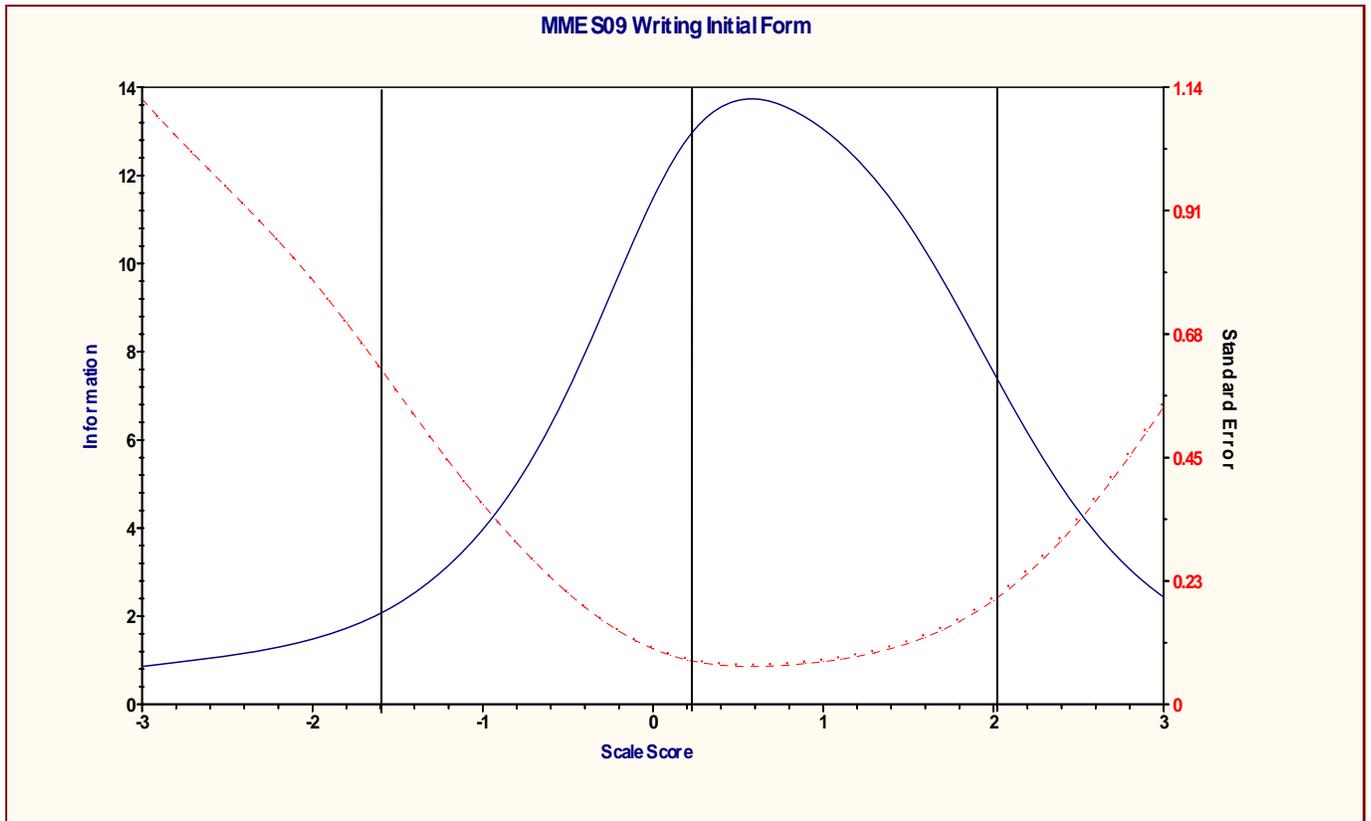
- ACT. (2007a). *ACT National Curriculum Survey 2005-2006*. Iowa City, IA: Author.
- ACT. (2009). *ACT Writing Test Technical Report*. Iowa City, IA: Author.
- ACT. (2007b). *The ACT technical manual*. Iowa City, IA: Author.
- ACT. (2008a). *WorkKeys Applied Mathematics technical manual*. Iowa City, IA: Author.
- ACT. (2008b). *WorkKeys Locating Information technical manual*. Iowa City, IA: Author.
- ACT. (2008c). *WorkKeys Reading for Information technical manual*. Iowa City, IA: Author.
- American Educational Research Association, American Psychological Association, National Council on Measurement in Education, & Joint Committee on Standards for Educational and Psychological Testing (U.S.). (1999) *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Anastasi, A. (1982). *Psychological testing (5th ed.)*. New York: Macmillan.
- Assessment and Examination Service. (2006). *Standard Setting Plan*.
- Assessment and Examination Service. (2006). *Standard Setting Report*.
- Crocker, L. & Algina, J. (1986). *Introduction to classical and modern test theory*. Holt, Rinehart and Winston, Inc.
- Dorans, N.J., & Schmitt, A. P. (1991). Constructed response and differential item functioning: A pragmatic approach (Research Rep. No. 91-47). Princeton, NJ: Educational Testing Service.
- Dossey, J.A. (2005). *Comparison of the ACT and WorkKeys assessments with the Mathematics and Science Content Expectations in the Michigan Curriculum Framework*.
- Hambleton, R.K., & Swaminathan, H. (1985). *Item response theory: Principles and Applications*. Boston: Kluwer.
- Holland, P.W., & Wainer, H. (1993). *Differential item functioning*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.
- Kolen, M.J. (1984). Effectiveness of analytic smoothing in equipercentile equating. *Journal of Educational Statistics*, 9, 25 – 44.
- Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices*, New York: Springer.
- Lee, W.-C., Hanson, B. A., & Brennan, R. L. (2002). Estimating consistency and accuracy indices for multiple classifications. *Applied Psychological Measurement*, 26(4), 412-432.

- Martineau, J. A. (2007). An extension and practical evaluation of expected classification accuracy. *Applied Psychological Measurement* 31(3), 181-194.
- Michigan Department of Education. (2006). *Michigan Department of Education Memorandum*.
- Michigan Department of Education. (2006). *Minutes of the State Board of Education*.
- Michigan Department of Education Web Site. (2007). <http://www.michigan.gov/mde/>.
- Millman, J., & Greene, U. (1989). *The specification and development of tests of achievement and ability*. In R.L. Linn (Ed.) *Educational Measurement* (3rd edition, pp. 335-366). New York: American Council on Education and Macmillan.
- Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, 16, 159 – 176.
- Muraki, E. (1996). A generalized partial credit model. In Van der Linden, W.J. & Hambleton, R.K. (Eds). *Handbook of modern item response theory* (pp. 153 – 168). New York: Springer.
- Muraki, E. , & Bock R. D. (1997). *PARSCALE: IRT item analysis and test scoring for rating-scale data*. Chicago, IL: Scientific Software International.
- Orlando, M. & Thissen, D. (2000). Likelihood-based item fit indices for dichotomous item response theory models. *Applied Psychological Measurement*, 24(1), 50-64.
- Shanahan, T. (2005). *Review of ACT Coverage of Michigan Language Arts Standards*.
- Stocking, M. L. & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, 7, 201-210.
- Webb, N.L. (2005). *Alignment Analysis of Language Arts Standards and Assessment: Michigan Grades 9–12*.
- Webb, N.L. (2005). *Alignment Analysis of Mathematics Standards and Assessments: Michigan High School*.
- Webb, N.L. (2006). *Alignment Analysis of Mathematics Standards and Michigan Merit Examination*.
- Webb, N.L. (2006). *Alignment Analysis of Reading and Language Arts Standards and Michigan Merit Examination*.
- Webb, N.L. (2006). *Alignment Analysis of Science Standards and Michigan Merit Examination*.
- Yen, W. M. & Fitzpatrick, A. R. (2006). Item Response Theory. In R. L. Brennan (Ed.) *Educational Measurement* (4th edition, pp. 111-153). Westport, CT: Praeger.

Appendices

Appendix A: Plots of PARSCALE Information Functions

Spring 2009 Writing Initial Form



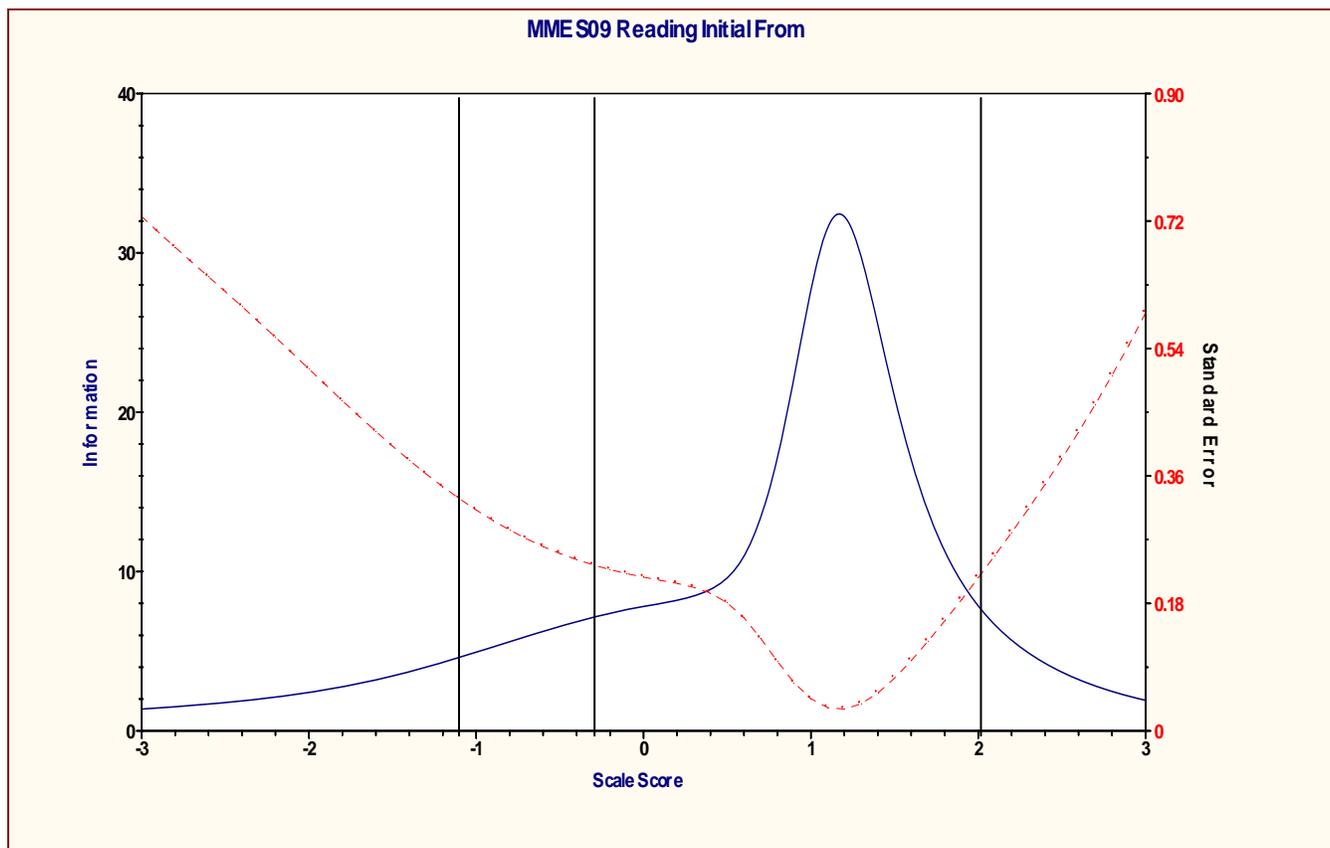
Test information curve: solid line

Standard error curve: dotted line

The total test information for a specific scale score is read from the left vertical axis.

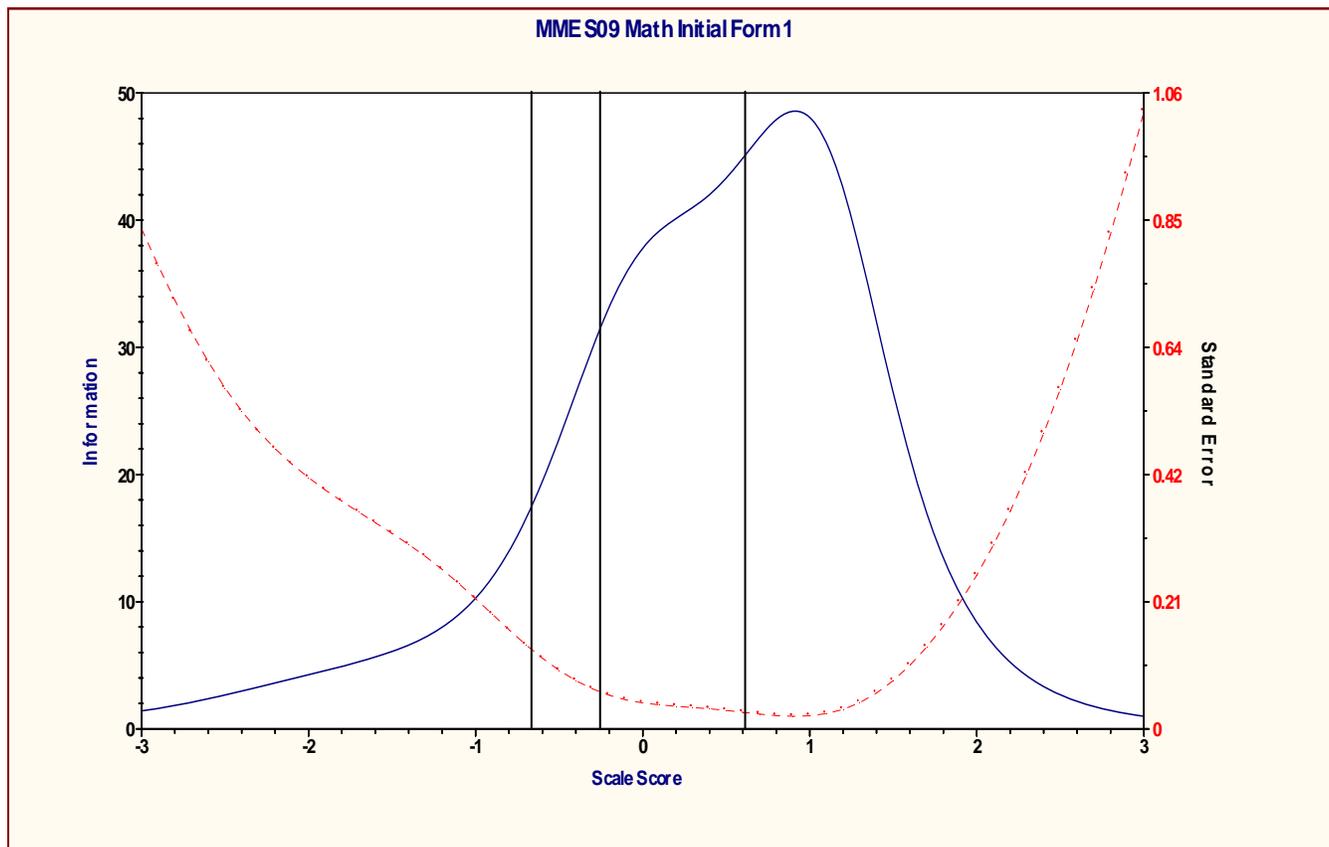
The standard error for a specific scale score is read from the right vertical axis.

Spring 2009 Reading Initial Form



Test information curve: solid line **Standard error curve: dotted line**
The total test information for a specific scale score is read from the left vertical axis.
The standard error for a specific scale score is read from the right vertical axis.

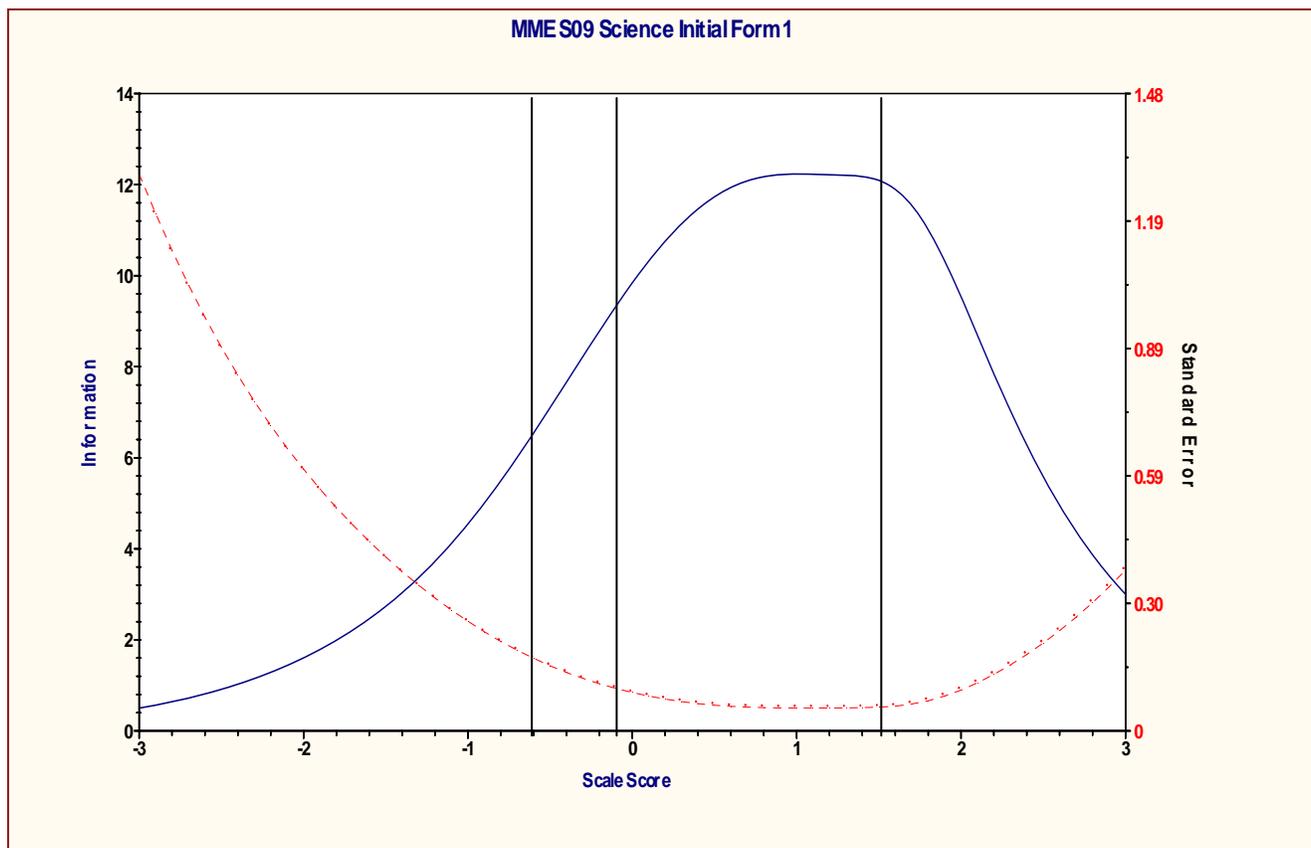
Spring 2009 Mathematics Initial Form



Test information curve: solid line **Standard error curve: dotted line**

The total test information for a specific scale score is read from the left vertical axis.
The standard error for a specific scale score is read from the right vertical axis.

Spring 2009 Science Initial Form



Test information curve: solid line **Standard error curve: dotted line**
The total test information for a specific scale score is read from the left vertical axis.
The standard error for a specific scale score is read from the right vertical axis.

Appendix B: Data Created for Field-Test Items

Field Format	Field Name	Field Description	Notes	Computation Description
A2	TESTCYCLEID	Year cycle (2 characters)		(From Test Map)
A2	SUBJECT	Subject (MA, SC, SS)	Mathematics, Science, Social Studies	(From Test Map)
A2	LEVEL	Grade-Level of GLCE		(From Test Map)
A4	FORM	Form the item appears on.	For matrix items the first form.	From Test Map
F2	NFORMS	Number of Forms Item Appears On (1 - 10)	Indicates how many forms a matrix item appears on, ranges 1-10.	Computed from Test Map Information
A60	FORMS	Form Numbers (string of 3x20 characters)	Indicates which forms a matrix item appears on, there will be as many form numbers as there are forms that item appears on.	Computed from Test Map Information
F12	ITEMCODE	Item Code (Both 7 and 12 digits used)	Unique Company ID number for an item (PEM)	(From Test Map)
A2	TYPE	Item Type (MC)	MC - multiple-choice	(From Test Map)
A50	SCENARIO	Scenario Title/Passage Type		(From Test Map)
F1	PART	Section of Test		(From Test Map)
A3	FUNCTION	1 = Common, 2 = Matrix; in the future, will be 99 if item is non-scorable		(From Test Map)
F12	TESTPOS	Item position on the test form.	For matrix items the first form.	
A60	POSITS	Test Positions (string of 3x20 characters)	Indicates positions in the test for each form that a matrix item appears on, there will be as many position numbers as there are forms that an item appears on.	Computed from Test Map Information
F1	ANCHOR	Anchor Item (0 = NO, 1 = YES); 1 Means "USE FOR PRE-EQUATING"		(From Test Map)
F1	POINTS	Maximum Score Points For Item		(From Test Map)
A1	STANDARD	Reported Standard		(From Test Map)
A4	DOMAIN	Reported Domain		(From Test Map)

Field Format	Field Name	Field Description	Notes	Computation Description
A2	BENCHMARK	Reported Benchmark		(From Test Map)
A10	GLCE	Reported Grade Level Expectation		(From Test Map)
A1	KEY	Item Answer Key (A, B, C, D; in the future will be 9 if non-scorable)	For MC items	(From Test Map)
A1	CALC	Calculator (Y) or Non-calculator (N)		(From Test Map)
A3	MATURITY	Maturity (FT or OP)	Field-Test, Operational	(From Test Map)
A9	MEAPID	MEAP Item ID	Michigan item identifier-concatenate(standard, domain, benchmark)	(From Test Map)
F2	GRADE	Grade-Level the item will be tested in	Grade in which an item administered	(From Test Map)
F3	MATRIX	Order for item processing	This is an ACT variable that contains the item processing order within each subject.	
F6	NCOUNT	N-count	Number of calibration cases used to produce statistics	Total number of calibration students who took the item regardless of the number of forms on which that item appears. Inclusion/exclusion rules for calibration students will be defined by OEAA
F6	N_MAL	N-count Males	N-counts for break-down groups	Total number of calibration male students who took the item regardless of the number of forms on which that item appears
F6	N_FEM	N-count Females		Total number of calibration female students who took the item regardless of the number of forms on which that item appears
F6	N_WHI	N-count White		Total number of calibration white students who took the item regardless of the number of forms on which that item appears

Field Format	Field Name	Field Description	Notes	Computation Description
F6	N_BLA	N-count Black		Total number of calibration black students who took the item regardless of the number of forms on which that item appears
F2	A	Percent (option A)	Percent of ALL calibration cases	Number of students who chose option A divided by the total number of calibration students
F2	B	Percent (option B)		Number of students who chose option B divided by the total number of calibration students
F2	C	Percent (option C)		Number of students who chose option C divided by the total number of calibration students
F2	D	Percent (option D)		Number of students who chose option D divided by the total number of calibration students
F2	M	Percent (mult. marks)		Number of students who chose multiple marks divided by the total number of calibration students
F2	O	Percent (Omits)		Number of students who had omits divided by the total number of calibration students
F2	MAA	Male Percent (A)		Percent for MALE calibration cases
F2	MAB	Male Percent (B)	Number of male students who chose option B divided by the total number of male calibration students	
F2	MAC	Male Percent (C)	Number of male students who chose option C divided by the total number of male calibration students	
F2	MAD	Male Percent (D)	Number of male students who chose option D divided by the total number of male calibration students	
F2	MAM	Male Percent (MM)	Number of male students who chose multiple marks divided by the total number of male calibration students	

Field Format	Field Name	Field Description	Notes	Computation Description	
F2	MAO	Male Percent (Omits)		Number of male students who had omits divided by the total number of male calibration students	
F2	FEA	Female Percent (A)	Percent for FEMALE calibration cases	Number of female students who chose option A divided by the total number of female calibration students	
F2	FEB	Female Percent (B)		Number of female students who chose option B divided by the total number of female calibration students	
F2	FEC	Female Percent (C)		Number of female students who chose option C divided by the total number of female calibration students	
F2	FED	Female Percent (D)		Number of female students who chose option D divided by the total number of female calibration students	
F2	FEM	Female Percent (MM)		Number of female students who chose multiple marks divided by the total number of female calibration students	
F2	FEO	Female Percent (Omits)		Number of female students who had omits divided by the total number of female calibration students	
F2	WHA	White Percent (A)		Percent for WHITE calibration cases	Number of white students who chose option A divided by the total number of white calibration students
F2	WHB	White Percent (B)			Number of white students who chose option B divided by the total number of white calibration students
F2	WHC	White Percent (C)	Number of white students who chose option C divided by the total number of white calibration students		
F2	WHD	White Percent (D)	Number of white students who chose option D divided by the total number		

Field Format	Field Name	Field Description	Notes	Computation Description	
				of white calibration students	
F2	WHM	White Percent (MM)		Number of white students who chose multiple marks divided by the total number of white calibration students	
F2	WHO	White Percent (Omits)		Number of white students who had omits divided by the total number of white calibration students	
F2	BLA	Black Percent (A)	Percent for BLACK calibration cases	Number of black students who chose option A divided by the total number of black calibration students	
F2	BLB	Black Percent (B)		Number of black students who chose option B divided by the total number of black calibration students	
F2	BLC	Black Percent (C)		Number of black students who chose option C divided by the total number of black calibration students	
F2	BLD	Black Percent (D)		Number of black students who chose option D divided by the total number of black calibration students	
F2	BLM	Black Percent (MM)		Number of black students who chose multiple marks divided by the total number of black calibration students	
F2	BLO	Black Percent (Omits)		Number of black students who had omits divided by the total number of black calibration students	
F8.4	PVAL	P-value		P-value of item scores (all cases)	The sum of students' gained score divided by the total number of all students
F8.4	MPVAL	P-value for Male		Impact analysis: item means for break-down groups	The sum of male students' gained score divided by the total number of male students
F8.4	FPVAL	P-value for Female			The sum of female students' gained score divided by the total number of female students
F8.4	WPVAL	P-value for White	The sum of white students' gained score divided by the total number of		

Field Format	Field Name	Field Description	Notes	Computation Description
				white students
F8.4	BPVAL	P-value for Black		The sum of black students' gained score divided by the total number of black students
F8.4	ADJPVAL	Adjusted P-value	Adjusted P-value = (Arithmetic mean - MIN item score) / (MAX item score - MIN item score)	Difference between the arithmetic mean and the minimum item score divided by the item score range
A5	DIFFICFL	Difficulty flag	Based on Test Construction Specifications	For MC item p LT .3 or p GT .9.
F8.4	SDEV	Item Standard Deviation	Standard deviation of item scores	Standard deviation of item score distribution
F8.4	ITOT	Item-Total Correlation	Pearson product-moment correlation (Point-Biserial correlation for dichotomous items)	Point-biserial correlation for MC items (see Crocker & Algina, 1986, page 317)
F8.4	ITOTBIS	Biserial Correlation	For MC: biserial	Biserial correlation for MC items (see Crocker & Algina, 1986, page 317)
F8.4	ITOTC	Point-Biserial Correlation (corrected)	For MC items (corrected for maximal possible value)	Corrected point-biserial correlation (see Crocker & Algina, 1986, page 317)
A2	ITOTFL	Item-Total correlation flag	Based on Test Construction Specifications	For MC item if pb LT .25.
F8.4	APB	P-b correlation for option A	Options point-biserial correlations (for CR items only Omits Rpb is supplied)	Point-biserial correlation for option A for a MC item when those students who chose option A is scored as 1
F8.4	BPB	P-b correlation for option B		Point-biserial correlation for option B for a MC item when those students who chose option B is scored as 1
F8.4	CPB	P-b correlation for option C		Point-biserial correlation for option C for a MC item when those students who chose option C is scored as 1
F8.4	DPB	P-b correlation for option D		Point-biserial correlation for option D for a MC item when those students who chose option D is scored as 1
F8.4	OPB	P-b correlation for Omits		Point-biserial correlation for omits for a MC item when those students

Field Format	Field Name	Field Description	Notes	Computation Description	
				who omitted the item is scored as 1	
A7	MISKFL	Flag for potential miskeying	Based on Test Construction Specifications	For MC, if keyed option not the highest percentage, or any option LT 2% or any non-keyed item pb GT 0, or omit pb GT .03.	
F8.4	MCHI_MF	Mantel CHSQ Male-Female	DIF analyses: Mantel chi-square (for both dichotomous and polytomous items), Mantel-Haenszel Delta and corresponding lower and upper 95% confidence interval limits for dichotomous items (not supplied for polytomous items)	Mantel Chi-square for male versus female comparison (See Holland & Wainer, 1993 page 40)	
F8.4	MHDL_MF	Lower Limit of 95% Confidence Interval for MHD_MF			
F8.4	MHD_MF	Mantel-Haenszel Delta Male-Female		Mantel Haenszel delta for male versus female comparison (See Holland & Wainer, 1993 page 41)	
F8.4	MHDU_MF	Upper Limit of 95% Confidence Interval for MHD_MF			
F8.4	MCHI_WB	Mantel CHSQ White-Black		Mantel Chi-square for white versus black comparison (See Holland & Wainer, 1993 page 40)	
F8.4	MHDL_WB	Lower Limit of 95% Confidence Interval for MHD_WB			
F8.4	MHD_WB	Mantel-Haenszel Delta White-Black		Mantel Haenszel delta for white versus black comparison (See Holland & Wainer, 1993 page 41)	
F8.4	MHDU_WB	Upper Limit of 95% Confidence Interval for MHD_WB			
A2	DIF_MF	DIF category for M-F (A, B, C)		DIF level categorization: A - no or negligible, B - moderate, C -	Items are classified as A category of DIF if either MH D-DIF is not

Field Format	Field Name	Field Description	Notes	Computation Description
A2	DIF_WB	DIF category for W-B (A, B, C)	substantial.	statistically different from zero (using the 5% significance level) or if the magnitude of the MH D-DIF values is less than one delta unit in absolute value. Items are classified as C category of DIF if MH D-DIF both exceeds 1.5 in absolute value and is statistically significantly larger than 1.0 in absolute value (using the 5% significance level). All other items are classified as category B.
A6	FG_MF	Favored group for M-F (Male, Female)	Favored group if DIF level equal to B or C	
A6	FG_WB	Favored group for W-B (White, Black)		
F8.5	APAR	A parameter (scaled)	For dichotomous items.	Item discrimination parameter from IRT calibration and equating
F8.5	ASE	SE for A parameter (scaled)	For dichotomous items.	Standard error for item discrimination parameter from IRT calibration and equating
F8.5	BPAR	B parameter (scaled)	For dichotomous items.	Item difficulty parameter from IRT calibration and equating
F8.5	BSE	SE for B parameter (scaled)	For dichotomous items.	Standard error for item difficulty parameter from IRT calibration and equating
F8.5	CPAR	C parameter (scaled)		Item pseudo-guessing parameter from IRT calibration and equating
F8.5	CSE	SE for C parameter (scaled)		Standard error for item pseudo-guessing parameter from IRT calibration and equating
F8.4	MSQIN	Mean-square infit	Rasch fit index and flag: blank (0.5 < 1.5), MM (misfit moderate: 1.5 < 2.0), MH (misfit high: 2.0 <), TP (too predictable: < 0.5). Not supplied for 3PL and 2PPC models.	Infit index output from Winsteps calibration
F8.4	MSQOUT	Mean-square outfit		Outfit index output from Winsteps calibration
A2	MSQFITFL	Mean-square fit flag (blank, MM, MH, TP)		

Field Format	Field Name	Field Description	Notes	Computation Description
F1	FITLEV	Misfit level (0, 1, 2)		Mean-squares > 2 indicate distorting or degrading the measurement system, flagged as misfit level 2. 1.5 – 2 means unproductive for construction of measurement, but not degrading, flagged as misfit level 1. < 0.5 means less productive for measurement, but not degrading. It may produce misleadingly good reliabilities and separations, flagged as misfit level 1. Otherwise, no flag with a misfit level of 0
F10.3	CHISQ	Chi-square statistics for 3PL and GPC fit index computed by PARSCALE	For dichotomous items.	Use ITEMFIT = 10 to specify the number (10) of frequency score groups to be used for computation of item-fit index in PARSCALE calibration runs. Note 10 deciles are used for other item statistics.
F5.0	DF	Degrees of freedom associated with the Chi-square fit index computed by PARSCALE.	For dichotomous items.	
F5.3	P_CHISQ	P-value associated with the Chi-square fit index computed by PARSCALE.	For dichotomous items.	
F8.3	sx2	IRT fit statistic for PARSCALE calibrated items.	Replaces ZQ1 fit statistic.	
F3	df_sx2	degrees of freedom for sx2 statistic.		
F8.3	p_sx2	p-value for sx2 statistic		
A2	sx2fitflag	Fit Flag based on sx2 statistic	Replaces ZQ1 fit flag.	Equals NF (no fit) if p-value < .05, otherwise blank.
F8.5	BASIC	(Theta cut for Basic)		
F8.5	MET	(Theta cut for Met)		
F8.5	EXCEED	(Theta cut for Exceed)		
F8.5	ICC1	(ICC at cut for Basic)		
F8.5	ICC2	(ICC at cut for Met)		
F8.5	ICC3	(ICC at cut for Exceed)		

Field Format	Field Name	Field Description	Notes	Computation Description
F8.5	INFO1	Item information at cut point 1	Item information at performance level cut-points.	Item information computed at cut score 1 based on Hambleton & Swaminathan (1985, page 106-107)
F8.5	INFO2	Item information at cut point 2		Item information computed at cut score 2 based on Hambleton & Swaminathan (1985, page 106-107)
F8.5	INFO3	Item information at cut point 3		Item information computed at cut score 3 based on Hambleton & Swaminathan (1985, page 106-107)
F8.3	TH01	Theta point 1	Theta points for plotting conditional item means.	Theta point corresponding to decile 1 (lowest 10%)
F8.3	TH02	Theta point 2		Theta point corresponding to decile 2
F8.3	TH03	Theta point 3		Theta point corresponding to decile 3
F8.3	TH04	Theta point 4		Theta point corresponding to decile 4
F8.3	TH05	Theta point 5		Theta point corresponding to decile 5
F8.3	TH06	Theta point 6		Theta point corresponding to decile 6
F8.3	TH07	Theta point 7		Theta point corresponding to decile 7
F8.3	TH08	Theta point 8		Theta point corresponding to decile 8
F8.3	TH09	Theta point 9		Theta point corresponding to decile 9
F8.3	TH10	Theta point 10		Theta point corresponding to decile 10 (highest 10%)
F8.3	AD01	Conditional Item Mean for Decile 1	Conditional item means plot: All	Item mean for decile 1 for all students
F8.3	AD02	Conditional Item Mean for Decile 2		Item mean for decile 2 for all students
F8.3	AD03	Conditional Item Mean for Decile 3		Item mean for decile 3 for all students
F8.3	AD04	Conditional Item Mean for Decile 4		Item mean for decile 4 for all students
F8.3	AD05	Conditional Item Mean for Decile 5		Item mean for decile 5 for all students
F8.3	AD06	Conditional Item Mean for Decile 6		Item mean for decile 6 for all students
F8.3	AD07	Conditional Item Mean for Decile 7		Item mean for decile 7 for all students
F8.3	AD08	Conditional Item Mean for		Item mean for decile 8 for all

Field Format	Field Name	Field Description	Notes	Computation Description
		Decile 8		students
F8.3	AD09	Conditional Item Mean for Decile 9		Item mean for decile 9 for all students
F8.3	AD10	Conditional Item Mean for Decile 10		Item mean for decile 10 for all students
F8.3	MD01	Conditional Item Mean for Decile 1	Conditional item means plot: Males	Item mean for decile 1 for male students
F8.3	MD02	Conditional Item Mean for Decile 2		Item mean for decile 2 for male students
F8.3	MD03	Conditional Item Mean for Decile 3		Item mean for decile 3 for male students
F8.3	MD04	Conditional Item Mean for Decile 4		Item mean for decile 4 for male students
F8.3	MD05	Conditional Item Mean for Decile 5		Item mean for decile 5 for male students
F8.3	MD06	Conditional Item Mean for Decile 6		Item mean for decile 6 for male students
F8.3	MD07	Conditional Item Mean for Decile 7		Item mean for decile 7 for male students
F8.3	MD08	Conditional Item Mean for Decile 8		Item mean for decile 8 for male students
F8.3	MD09	Conditional Item Mean for Decile 9		Item mean for decile 9 for male students
F8.3	MD10	Conditional Item Mean for Decile 10		Item mean for decile 10 for male students
F8.3	FD01	Conditional Item Mean for Decile 1	Conditional item means plot: Females	Item mean for decile 1 for female students
F8.3	FD02	Conditional Item Mean for Decile 2		Item mean for decile 2 for female students
F8.3	FD03	Conditional Item Mean for Decile 3		Item mean for decile 3 for female students
F8.3	FD04	Conditional Item Mean for Decile 4		Item mean for decile 4 for female students
F8.3	FD05	Conditional Item Mean for Decile 5		Item mean for decile 5 for female students
F8.3	FD06	Conditional Item Mean for Decile 6		Item mean for decile 6 for female students

Field Format	Field Name	Field Description	Notes	Computation Description
F8.3	FD07	Conditional Item Mean for Decile 7		Item mean for decile 7 for female students
F8.3	FD08	Conditional Item Mean for Decile 8		Item mean for decile 8 for female students
F8.3	FD09	Conditional Item Mean for Decile 9		Item mean for decile 9 for female students
F8.3	FD10	Conditional Item Mean for Decile 10		Item mean for decile 10 for female students
F8.3	WD01	Conditional Item Mean for Decile 1	Conditional item means plot: Whites	Item mean for decile 1 for white students
F8.3	WD02	Conditional Item Mean for Decile 2		Item mean for decile 2 for white students
F8.3	WD03	Conditional Item Mean for Decile 3		Item mean for decile 3 for white students
F8.3	WD04	Conditional Item Mean for Decile 4		Item mean for decile 4 for white students
F8.3	WD05	Conditional Item Mean for Decile 5		Item mean for decile 5 for white students
F8.3	WD06	Conditional Item Mean for Decile 6		Item mean for decile 6 for white students
F8.3	WD07	Conditional Item Mean for Decile 7		Item mean for decile 7 for white students
F8.3	WD08	Conditional Item Mean for Decile 8		Item mean for decile 8 for white students
F8.3	WD09	Conditional Item Mean for Decile 9		Item mean for decile 9 for white students
F8.3	WD10	Conditional Item Mean for Decile 10		Item mean for decile 10 for white students
F8.3	BD01	Conditional Item Mean for Decile 1	Conditional item means plot: Blacks	Item mean for decile 1 for black students
F8.3	BD02	Conditional Item Mean for Decile 2		Item mean for decile 2 for black students
F8.3	BD03	Conditional Item Mean for Decile 3		Item mean for decile 3 for black students
F8.3	BD04	Conditional Item Mean for Decile 4		Item mean for decile 4 for black students
F8.3	BD05	Conditional Item Mean for		Item mean for decile 5 for black

Field Format	Field Name	Field Description	Notes		Computation Description
		Decile 5			students
F8.3	BD06	Conditional Item Mean for Decile 6			Item mean for decile 6 for black students
F8.3	BD07	Conditional Item Mean for Decile 7			Item mean for decile 7 for black students
F8.3	BD08	Conditional Item Mean for Decile 8			Item mean for decile 8 for black students
F8.3	BD09	Conditional Item Mean for Decile 9			Item mean for decile 9 for black students
F8.3	BD10	Conditional Item Mean for Decile 10			Item mean for decile 10 for black students
F8.3	A95_A0	95th percentile	Box & whisker plot: All	Option A	95th percentile of theta for all students for Option A
F8.3	A75_A0	75th percentile			75th percentile of theta for all students for Option A
F8.3	A50_A0	50th percentile			50th percentile of theta for all students for Option A
F8.3	A25_A0	25th percentile			25th percentile of theta for all students for Option A
F8.3	A05_A0	5th percentile			5th percentile of theta for all students for Option A
F8.3	M95_A0	95th percentile	Box & whisker plot: Males		95th percentile of theta for male students for Option A
F8.3	M75_A0	75th percentile			75th percentile of theta for male students for Option A
F8.3	M50_A0	50th percentile			50th percentile of theta for male students for Option A
F8.3	M25_A0	25th percentile			25th percentile of theta for male students for Option A
F8.3	M05_A0	5th percentile			5th percentile of theta for male students for Option A
F8.3	F95_A0	95th percentile	Box & whisker plot: Females		95th percentile of theta for female students for Option A
F8.3	F75_A0	75th percentile			75th percentile of theta for female students for Option A
F8.3	F50_A0	50th percentile			50th percentile of theta for female students for Option A

Field Format	Field Name	Field Description	Notes		Computation Description	
F8.3	F25_A0	25th percentile			25th percentile of theta for female students for Option A	
F8.3	F05_A0	5th percentile			5th percentile of theta for female students for Option A	
F8.3	W95_A0	95th percentile	Box & whisker plot: Whites		95th percentile of theta for white students for Option A	
F8.3	W75_A0	75th percentile			75th percentile of theta for white students for Option A	
F8.3	W50_A0	50th percentile			50th percentile of theta for white students for Option A	
F8.3	W25_A0	25th percentile			25th percentile of theta for white students for Option A	
F8.3	W05_A0	5th percentile			5th percentile of theta for white students for Option A	
F8.3	B95_A0	95th percentile		Box & whisker plot: Blacks		95th percentile of theta for black students for Option A
F8.3	B75_A0	75th percentile				75th percentile of theta for black students for Option A
F8.3	B50_A0	50th percentile			50th percentile of theta for black students for Option A	
F8.3	B25_A0	25th percentile			25th percentile of theta for black students for Option A	
F8.3	B05_A0	5th percentile			5th percentile of theta for black students for Option A	
F8.3	A95_B1	95th percentile	Box & whisker plot: All		Option B	95th percentile of theta for all students for Option B
F8.3	A75_B1	75th percentile				75th percentile of theta for all students for Option B
F8.3	A50_B1	50th percentile				50th percentile of theta for all students for Option B
F8.3	A25_B1	25th percentile				25th percentile of theta for all students for Option B
F8.3	A05_B1	5th percentile				5th percentile of theta for all students for Option B
F8.3	M95_B1	95th percentile	Box & whisker plot: Males		95th percentile of theta for male students for Option B	
F8.3	M75_B1	75th percentile			75th percentile of theta for male	

Field Format	Field Name	Field Description	Notes	Computation Description
				students for Option B
F8.3	M50_B1	50th percentile		50th percentile of theta for male students for Option B
F8.3	M25_B1	25th percentile		25th percentile of theta for male students for Option B
F8.3	M05_B1	5th percentile		5th percentile of theta for male students for Option B
F8.3	F95_B1	95th percentile	Box & whisker plot: Females	95th percentile of theta for female students for Option B
F8.3	F75_B1	75th percentile		75th percentile of theta for female students for Option B
F8.3	F50_B1	50th percentile		50th percentile of theta for female students for Option B
F8.3	F25_B1	25th percentile		25th percentile of theta for female students for Option B
F8.3	F05_B1	5th percentile		5th percentile of theta for female students for Option B
F8.3	W95_B1	95th percentile	Box & whisker plot: Whites	95th percentile of theta for white students for Option B
F8.3	W75_B1	75th percentile		75th percentile of theta for white students for Option B
F8.3	W50_B1	50th percentile		50th percentile of theta for white students for Option B
F8.3	W25_B1	25th percentile		25th percentile of theta for white students for Option B
F8.3	W05_B1	5th percentile		5th percentile of theta for white students for Option B
F8.3	B95_B1	95th percentile	Box & whisker plot: Blacks	95th percentile of theta for black students for Option B
F8.3	B75_B1	75th percentile		75th percentile of theta for black students for Option B
F8.3	B50_B1	50th percentile		50th percentile of theta for black students for Option B
F8.3	B25_B1	25th percentile		25th percentile of theta for black students for Option B
F8.3	B05_B1	5th percentile		5th percentile of theta for black students for Option B

Field Format	Field Name	Field Description	Notes		Computation Description
F8.3	A95_C2	95th percentile	Box & whisker plot: All	Option C	95th percentile of theta for all students for Option C
F8.3	A75_C2	75th percentile			75th percentile of theta for all students for Option C
F8.3	A50_C2	50th percentile			50th percentile of theta for all students for Option C
F8.3	A25_C2	25th percentile			25th percentile of theta for all students for Option C
F8.3	A05_C2	5th percentile			5th percentile of theta for all students for Option C
F8.3	M95_C2	95th percentile	Box & whisker plot: Males		95th percentile of theta for male students for Option C
F8.3	M75_C2	75th percentile			75th percentile of theta for male students for Option C
F8.3	M50_C2	50th percentile			50th percentile of theta for male students for Option C
F8.3	M25_C2	25th percentile			25th percentile of theta for male students for Option C
F8.3	M05_C2	5th percentile			5th percentile of theta for male students for Option C
F8.3	F95_C2	95th percentile	Box & whisker plot: Females		95th percentile of theta for female students for Option C
F8.3	F75_C2	75th percentile			75th percentile of theta for female students for Option C
F8.3	F50_C2	50th percentile			50th percentile of theta for female students for Option C
F8.3	F25_C2	25th percentile			25th percentile of theta for female students for Option C
F8.3	F05_C2	5th percentile			5th percentile of theta for female students for Option C
F8.3	W95_C2	95th percentile	Box & whisker plot: Whites		95th percentile of theta for white students for Option C
F8.3	W75_C2	75th percentile			75th percentile of theta for white students for Option C
F8.3	W50_C2	50th percentile			50th percentile of theta for white students for Option C
F8.3	W25_C2	25th percentile			25th percentile of theta for white

Field Format	Field Name	Field Description	Notes		Computation Description	
					students for Option C	
F8.3	W05_C2	5th percentile			5th percentile of theta for white students for Option C	
F8.3	B95_C2	95th percentile	Box & whisker plot: Blacks		95th percentile of theta for black students for Option C	
F8.3	B75_C2	75th percentile			75th percentile of theta for black students for Option C	
F8.3	B50_C2	50th percentile			50th percentile of theta for black students for Option C	
F8.3	B25_C2	25th percentile			25th percentile of theta for black students for Option C	
F8.3	B05_C2	5th percentile			5th percentile of theta for black students for Option C	
F8.3	A95_D3	95th percentile		Box & whisker plot: All	Option D	95th percentile of theta for all students for Option D
F8.3	A75_D3	75th percentile				75th percentile of theta for all students for Option D
F8.3	A50_D3	50th percentile				50th percentile of theta for all students for Option D
F8.3	A25_D3	25th percentile				25th percentile of theta for all students for Option D
F8.3	A05_D3	5th percentile				5th percentile of theta for all students for Option D
F8.3	M95_D3	95th percentile	Box & whisker plot: Males		95th percentile of theta for male students for Option D	
F8.3	M75_D3	75th percentile			75th percentile of theta for male students for Option D	
F8.3	M50_D3	50th percentile			50th percentile of theta for male students for Option D	
F8.3	M25_D3	25th percentile			25th percentile of theta for male students for Option D	
F8.3	M05_D3	5th percentile			5th percentile of theta for male students for Option D	
F8.3	F95_D3	95th percentile	Box & whisker plot: Females		95th percentile of theta for female students for Option D	
F8.3	F75_D3	75th percentile			75th percentile of theta for female students for Option D	

Field Format	Field Name	Field Description	Notes		Computation Description
F8.3	F50_D3	50th percentile			50th percentile of theta for female students for Option D
F8.3	F25_D3	25th percentile			25th percentile of theta for female students for Option D
F8.3	F05_D3	5th percentile			5th percentile of theta for female students for Option D
F8.3	W95_D3	95th percentile	Box & whisker plot: Whites		95th percentile of theta for white students for Option D
F8.3	W75_D3	75th percentile			75th percentile of theta for white students for Option D
F8.3	W50_D3	50th percentile			50th percentile of theta for white students for Option D
F8.3	W25_D3	25th percentile			25th percentile of theta for white students for Option D
F8.3	W05_D3	5th percentile			5th percentile of theta for white students for Option D
F8.3	B95_D3	95th percentile	Box & whisker plot: Blacks		95th percentile of theta for black students for Option D
F8.3	B75_D3	75th percentile			75th percentile of theta for black students for Option D
F8.3	B50_D3	50th percentile			50th percentile of theta for black students for Option D
F8.3	B25_D3	25th percentile			25th percentile of theta for black students for Option D
F8.3	B05_D3	5th percentile			5th percentile of theta for black students for Option D
F8.3	A95_OM	95th percentile	Box & whisker plot: All	Omits	95th percentile of theta for all students for omits
F8.3	A75_OM	75th percentile			75th percentile of theta for all students for omits
F8.3	A50_OM	50th percentile			50th percentile of theta for all students for omits
F8.3	A25_OM	25th percentile			25th percentile of theta for all students for omits
F8.3	A05_OM	5th percentile			5th percentile of theta for all students for omits
F8.3	M95_OM	95th percentile	Box & whisker plot: Males		95th percentile of theta for male students for omits
F8.3	M75_OM	75th percentile			75th percentile of theta for male students for omits

Field Format	Field Name	Field Description	Notes	Computation Description
F8.3	M50_OM	50th percentile		50th percentile of theta for male students for omits
F8.3	M25_OM	25th percentile		25th percentile of theta for male students for omits
F8.3	M05_OM	5th percentile		5th percentile of theta for male students for omits
F8.3	F95_OM	95th percentile	Box & whisker plot: Females	95th percentile of theta for female students for omits
F8.3	F75_OM	75th percentile		75th percentile of theta for female students for omits
F8.3	F50_OM	50th percentile		50th percentile of theta for female students for omits
F8.3	F25_OM	25th percentile		25th percentile of theta for female students for omits
F8.3	F05_OM	5th percentile		5th percentile of theta for female students for omits
F8.3	W95_OM	95th percentile	Box & whisker plot: Whites	95th percentile of theta for white students for omits
F8.3	W75_OM	75th percentile		75th percentile of theta for white students for omits
F8.3	W50_OM	50th percentile		50th percentile of theta for white students for omits
F8.3	W25_OM	25th percentile		25th percentile of theta for white students for omits
F8.3	W05_OM	5th percentile		5th percentile of theta for white students for omits
F8.3	B95_OM	95th percentile	Box & whisker plot: Blacks	95th percentile of theta for black students for omits
F8.3	B75_OM	75th percentile		75th percentile of theta for black students for omits
F8.3	B50_OM	50th percentile		50th percentile of theta for black students for omits
F8.3	B25_OM	25th percentile		25th percentile of theta for black students for omits
F8.3	B05_OM	5th percentile		5th percentile of theta for black students for omits

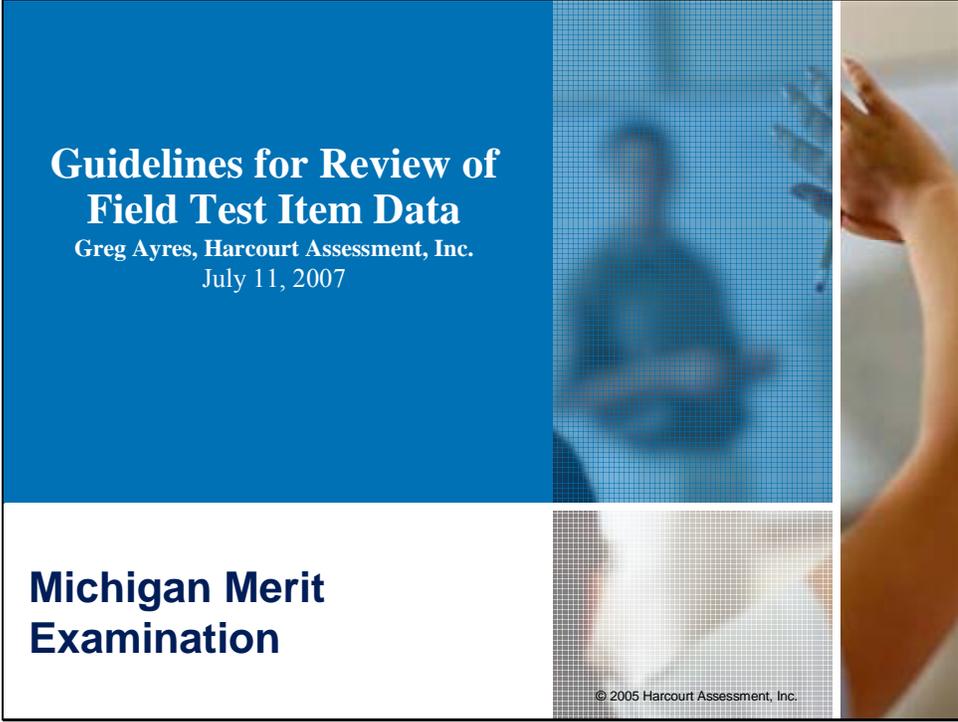
Appendix C: Statistics and Terms Used on Item Labels for Item Review Committees

CID	Company identification number for the item.
Maturity	Function of the reviewed item.
Form	Form numbers that contain the reviewed item.
Position	Position numbers in the test for the reviewed item (given for each form that the item appears on).
Type	Item type: MC – multiple-choice item, CR – constructed-response item, WR – writing.
Key	The correct answer for an MC item.
Max	The maximum score point for a CR or a writing item.
P-value	The percent of students who answered the item correctly. Its theoretical range is 0-1. It indicates item difficulty. Items with high p-values, such as .90, are relatively easy items. Those with p-values below .50 are relatively difficult items. P-values depend on the group of examinees who take the test.
Adj. P value	Item mean divided by the difference between minimum and maximum score points. It is equivalent to the p-value for the MC items when the score point is awarded either 1 or 0.
N-count	The number of tested students who were administered the item.
Rasch Difficulty	The usual range of Rasch difficulties is from -3 to +3 with mean of 0 and standard deviation of 1. 0 means medium difficulty. Positive values mean difficult items. Negative values mean easy items.
PB	Point-biserial correlation shows the relationship between a student's performance
Correlation	on the item and performance on the test as a whole. A high point-biserial correlation (e.g., above .50) indicates that students who answered the item correctly on the item achieved higher total scores on the test than those who answered the item incorrectly on the item. Values less than .25 may indicate a weaker than desired relationship. Note that extremely difficult or extremely easy items may have point-biserial correlation artificially reduced.
Item-Total Corr.	Item-total correlation shows the relationship between a student's performance on the item and performance on the test as a whole. A high item-total correlation (e.g., above .50) indicates that students who earned more points on the item achieved higher total scores on the test than those who earned fewer points on the item. Values less than .25 may indicate a weaker than desired relationship. Note that extremely difficult or extremely easy items may have item-total correlation artificially reduced.
FIT Flag	This flag indicates that two fit indices are out of the desired range. It means the Item may have not misfit or overfit the measurement model specified for the test analysis.

Difficulty Flag	This flag indicates that P-value, or adjusted p-value, or Rasch difficulty is out of the desired range.
PB Correlation Flag	This flag indicates that a MC item point-biserial correlation is smaller than the desired range of larger than 0.25.
Item-Total Corr. Flag	This flag indicates that a CR or a Writing item point-biserial correlation is smaller than the desired range of larger than 0.25.
Option Quality Flag	This flag indicates that a MC item may have a key problem. It could be that the key is not correct or it was miskeyed in scoring.
Score Point Dist. Flag	This flag indicates that a CR or a Writing item may have a scoring rubric problem. It could be the sample answer for each score point was not correctly identified.
Option Analysis	Percent of students who selected options A, B, C, and D, or did not choose any option (Omit) for all students and for subgroups by gender and ethnicity.
Score Point Distribution	Percent of students who earned each valid score point and who did not answer the CR or writing item for all students and for subgroups by gender and ethnicity.
Option PB Correlation	Point-biserial correlation for each of a MC item options. The key option point-biserial correlation should be positive and high. The non-keyed option point-biserial should be negative and low.
Omit PB Correlation	Point-biserial correlation for omit of a CR or Writing item. The omit point-biserial correlation should be negative.
Invalid Codes	The codes for invalid responses for a CR or a writing item.
DIF	Differential Item Functioning index. It indicates whether the reviewed item favors a particular subgroup of the student population; thus that group of students may have a higher chance of answering the item correctly or earn higher score point than the contrasted group. The focused group is often the minority group such as female in the gender group comparison, and black in the ethnic group comparison. The reference group is often the majority group which is male in the gender group comparison, and white in the ethnic group comparison.

Appendix D: Guidelines for Bias Review of Field Test Item Data

Slide 1

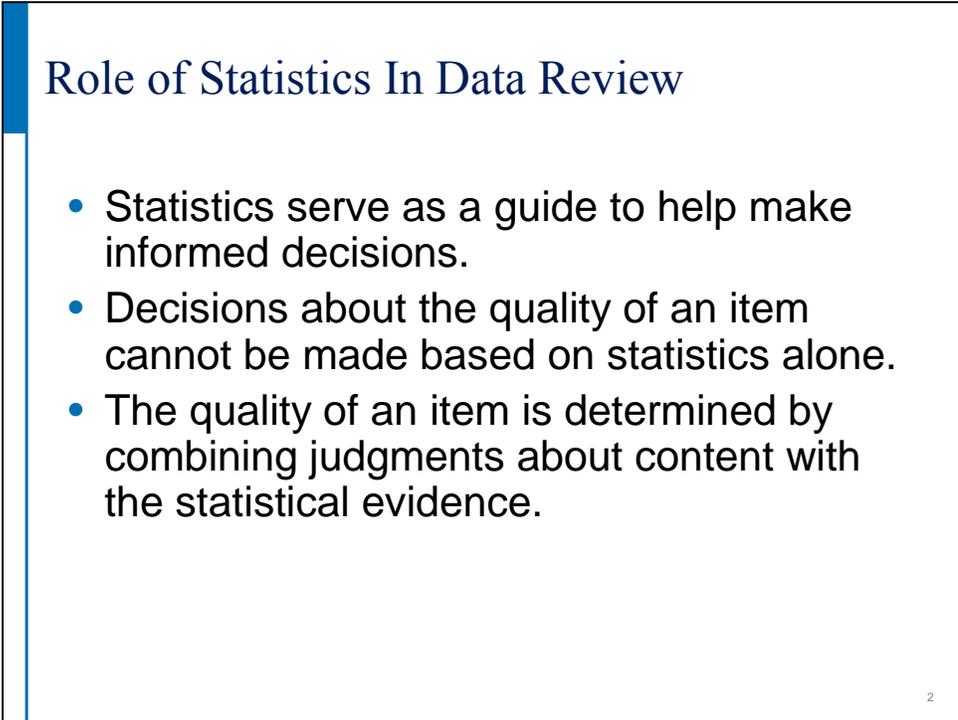


**Guidelines for Review of
Field Test Item Data**
Greg Ayres, Harcourt Assessment, Inc.
July 11, 2007

**Michigan Merit
Examination**

© 2005 Harcourt Assessment, Inc.

Slide 2



Role of Statistics In Data Review

- Statistics serve as a guide to help make informed decisions.
- Decisions about the quality of an item cannot be made based on statistics alone.
- The quality of an item is determined by combining judgments about content with the statistical evidence.

2

Statistical Evidence

- Psychometricians collect evidence about item and test characteristics.
- Statistical evidence needs to be weighed to determine whether the item is a good candidate for an operational form.

Item Statistics

MME **Grade: 11** **Subject: Math** **Admin: Spring 2007**

ID: 3423345

GLCE: F.2.h.06

Accept as is

Form: 8

Reject

Position: 13

Accept with revision

Scenario: NA

Table 1. Item Information

Type: MC	P-value: .62	B parameter:	Difficulty Flag:
Key: B	N-count: 3695	PB Correlation: 0.50	PB Correlation Flag:
	Maturity: FT	Fit Flag:	Option Quality Flag:

Slide 5

Table 2. Breakout Group Descriptives and Option Analysis

		N-count	Percent of Students Selected Option				
			A	B *	C	D	Omit
Group	All	3695	7	62	15	16	0
	Male	1797	8	59	16	17	0
	Female	1898	7	64	13	15	0
	White	2913	7	65	13	14	0
	Black	519	7	44	22	26	0
Option PB Correlations			-0.15	0.50	-0.27	-0.29	-0.03

Table 3. Differential Item Functioning

Reference/ Focal Group	Male/ Female	White/ Black
Flag	B	
Favored Group	female	

5

Slide 6

Table 2. Breakout Group Descriptives and Score Point Distributions

		N-count	Item Mean	Percent of Students at Each Score Point						
				0	1	2	3	4	5	6
Group	All	1977	1.94	5	38	27	20	8	2	
	Male	998	1.75	7	43	26	17	6	1	
	Female	979	2.13	3	33	29	23	10	3	
	White	1572	2.03	4	35	28	22	9	2	
	Black	277	1.43	10	52	24	12	2		
Omit PB Correlation										

Table 3. Condition Code Distributions

Frequency of Students at Each Condition Code		
A	B	C
.400		1.21

Table 4. Differential Item Functioning

Reference/ Focal Group	Male/ Female	White/ Black
Flag	C	
Favored Group	female	

6

Classical Item Difficulty: P-value

- MC items: **P-value** is the percentage of students who answered the item correctly.
- CR items: Adjusted **P-value** is the item mean divided by its range (max score – min score).
- Theoretical range from 0 to 1, with values over 0.9 indicating items that may be too easy, and values below 0.3 indicating items that may be too difficult
- Group dependent (not comparable across administration years)

7

Item Discrimination: Item-Total Correlation

- Item-total correlation indicates agreement between **item** scores and **total** test scores.
 - Point-biserial correlation is a specific type of item-total correlation used for dichotomous items (e.g., MC items).
- Theoretical range from -1 to 1
- High item-total correlation indicates that students who answered an item correctly, or who received a higher score-point on an item, also have higher total test scores (and vice versa).
- Item-total correlation greater than 0.25 are acceptable; those below 0.25 should be scrutinized.

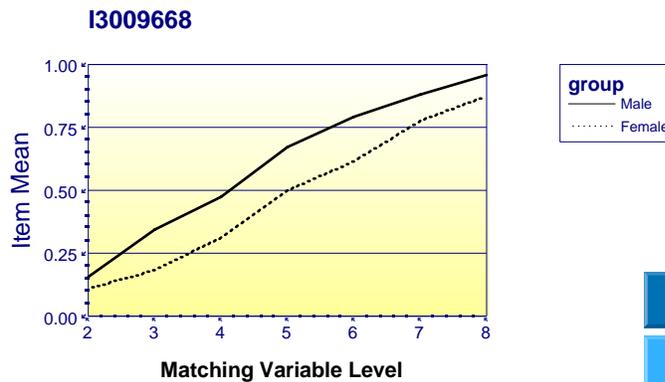
8

Option Analysis / Score Point Distribution

- Shows the percentage of students choosing each option on MC items, or earning a score point on CR items
- This percentage is given for all students and students grouped by ethnicity and gender.
- Option point-biserial correlation indicates the agreement between choosing each option (or earning a score point) and the total score on the test.

Differential Item Functioning (DIF) Analysis

- DIF refers to the **unexpected** differences in performance on a studied item between a reference and a focal group **after they have been matched** with respect to the total score on the test.



DIF and Item Bias

- An item is biased if it measures attributes irrelevant to the intended construct or is somehow a less acceptable measure of the construct for one subgroup.
- DIF does not necessarily mean that an item is biased. DIF only indicates that the examinees of equal proficiency from different subgroups have an unequal probability of responding correctly to an item.
- The results of DIF analyses provide a convenient starting point for the study of item bias.

11

DIF Levels

- Items are classified into one of the three DIF categories.
 - Category A: Negligible DIF, no group favored
 - Category B: Moderate DIF, one group is slightly favored by the studied item
 - Category C: Large DIF, one group is strongly favored by the studied item
- Items in category B and C are flagged and should be carefully examined for potential bias against a particular group.

12

DIF Table

- DIF flag: An indication of moderate DIF (flag B) or large DIF (flag C)
- Fav group: The flag for indicating which group is favored by the studied item



Summary

- Make informed decisions based on the data.
- Information on content and statistics determines the quality of an item.
- Weigh the statistical evidence and content, and then determine whether the items are good candidates for a live form.

Next

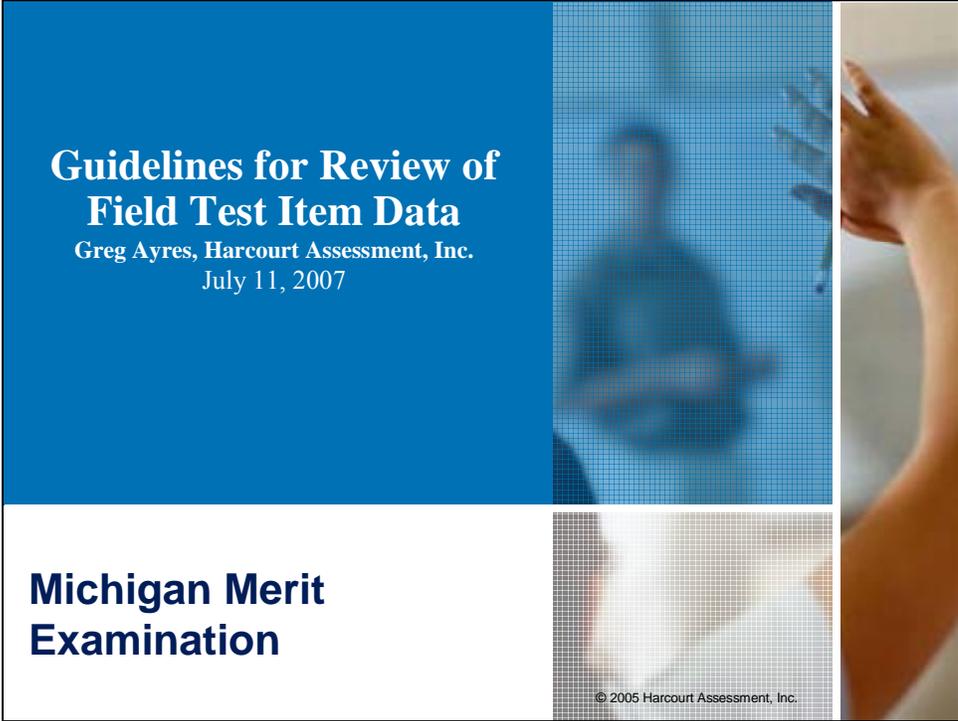
- Ask any questions that you may have
- Work in your respective subject area groups
- Enjoy the process

Thank you!



Appendix E: Guidelines for Content Review of Field Test Item Data

Slide 1

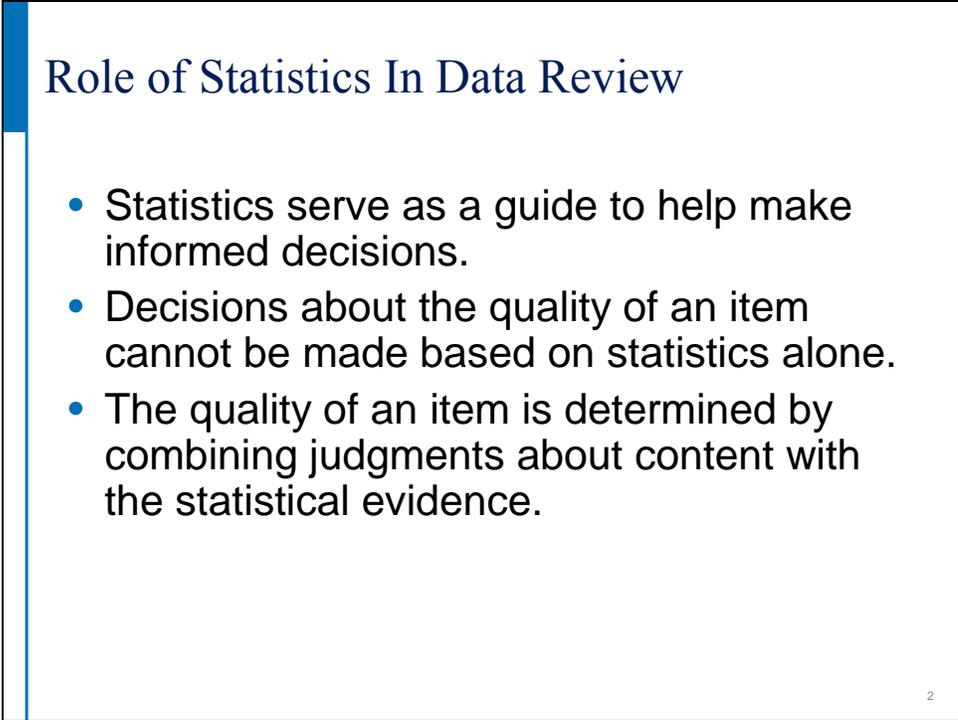


**Guidelines for Review of
Field Test Item Data**
Greg Ayres, Harcourt Assessment, Inc.
July 11, 2007

**Michigan Merit
Examination**

© 2005 Harcourt Assessment, Inc.

Slide 2



Role of Statistics In Data Review

- Statistics serve as a guide to help make informed decisions.
- Decisions about the quality of an item cannot be made based on statistics alone.
- The quality of an item is determined by combining judgments about content with the statistical evidence.

2

Statistical Evidence

- Psychometricians collect evidence about item and test characteristics.
- Statistical evidence needs to be weighed to determine whether the item is a good candidate for an operational form.

3

Item Statistics

MME **Grade: 11** **Subject: Math** **Admin: Spring 2007**

ID: 3423170

GLCE: G.1.h.05

Accept as is

Form: 2

Reject

Position: 14

Accept with revision

Scenario: NA

Table 1. Item Information

Type: MC	P-value: .32	B parameter:	Difficulty Flag:
Key: C	N-count: 3718	PB Correlation: 0.24	PB Correlation Flag: CL
	Maturity: FT	Fit Flag:	Option Quality Flag: P

4

Slide 5

Table 2. Breakout Group Descriptives and Option Analysis

		N-count	Percent of Students Selected Option				
			A	B	C *	D	Omit
Group	All	3718	14	31	32	22	0
	Male	1810	14	29	37	21	0
	Female	1908	14	34	29	23	0
	White	2898	13	31	33	22	0
	Black	539	17	34	28	20	0
Option PB Correlations			-0.24	-0.16	0.24	0.12	-0.04

Table 3. Differential Item Functioning

Reference/ Focal Group	Male/ Female	White/ Black
Flag		
Favored Group		

5

Slide 6

Table 2. Breakout Group Descriptives and Score Point Distributions

		N-count	Item Mean	Percent of Students at Each Score Point						
				0	1	2	3	4	5	6
Group	All	1977	1.94	5	38	27	20	8	2	
	Male	998	1.75	7	43	26	17	6	1	
	Female	979	2.13	3	33	29	23	10	3	
	White	1572	2.03	4	35	28	22	9	2	
	Black	277	1.43	10	52	24	12	2		
Omit PB Correlation										

Table 3. Condition Code Distributions

Frequency of Students at Each Condition Code		
A	B	C
.400		1.21

Table 4. Differential Item Functioning

Reference/ Focal Group	Male/ Female	White/ Black
Flag	C	
Favored Group	female	

6

Classical Item Difficulty: P-value

- MC items: **P-value** is the percentage of students who answered the item correctly.
- CR items: Adjusted **P-value** is the item mean divided by its range (max score – min score).
- Theoretical range from 0 to 1, with values over 0.9 indicating items that may be too easy, and values below 0.3 indicating items that may be too difficult
- Group dependent (not comparable across administration years)

7

Item Discrimination: Item-Total Correlation

- Item-total correlation indicates agreement between **item** scores and **total** test scores.
 - Point-biserial correlation is a specific type of item-total correlation used for dichotomous items (e.g., MC items).
- Theoretical range from -1 to 1
- High item-total correlation indicates that students who answered an item correctly, or who received a higher score-point on an item, also have higher total test scores (and vice versa).
- Item-total correlation greater than 0.25 are acceptable; those below 0.25 should be scrutinized.

8

Option Analysis / Score Point Distribution

- Shows the percentage of students choosing each option on MC items, or earning a score point on CR items
- This percentage is given for all students and students grouped by ethnicity and gender.
- Option point-biserial correlation indicates the agreement between choosing each option (or earning a score point) and the total score on the test.

9

Summary

- Make informed decisions based on the data.
- Information on content and statistics determines the quality of an item.
- Weigh the statistical evidence and content, and then determine whether the items are good candidates for a live form.

10

Next

- Ask any questions that you may have
- Work in your respective subject area groups
- Enjoy the process

Thank you!

