Illinois Alternate Assessment
2009 Technical Manual

Illinois State Board of Education
Division of Assessment

Table of Contents

1. PURPOSE AND DESIGN OF THE IAA TESTING PROGRAM 4
NCLB Requirements 4
Excerpts from the August 2005 Non-Regulatory Guidance 5
Test Development and Test Blueprint 6
Item Development 10
Item Development Cycle 10
Item Specifications 11
Qualifications of Item Writers and Method of Recruitment 13
Training Practices/Activities (in consideration of both content and bias) 14
Test Administration Training 14
Test Implementation Manual 14
Test Booklets 14
Answer Sheets 14
Online Test Platform 14
Teacher Training 15
Data Review Outcomes 15
Differential Item Functioning Analyses 16
2. RELIABILITY AND GENERALIZABILITY 19
Internal Consistency of Overall Scores 19
Standard Error of Measurement 20
IRT Test Information Function 24
IRT Conditional SEM 27
Reliability of Scores 27
Reliability of Performance Classification 27
3. VALIDITY 30
Performance-Based Measurement 30
Content Validity 32
Construct Validity 33
Dimensionality 33
Internal Construct 36
Criterion-related Validity 38
Agreement between Teacher Scores and Expert Scores 39
Table 3.5: Agreement between Teacher Scores and Expert Scores 40
Correlations between Teacher Scores and Expert Scores 40
4. CALIBRATION AND SCALING 42
Calibration 42
Scaling 42
5. STANDARDS VALIDATION 45
Panelists 45
The IAA Standards Validation Process 46
Linked Cut Scores 48
Grade 3 Writing 49
Recommended Raw Cut Scores 50
Rationale for Changing the Linked Cut Scores 50
Approved Cut Scores 51
Estimated Consequences of the Final Cut Scores 52
Panelist Variability 54
REFERENCES 58
APPENDIX A: IAA Scoring Rubric 60
APPENDIX B: Conditional Standard Errors of Measurement Associated With IAA Scale Scores 61
APPENDIX C: Classification Consistency 68

APPENDIX D: First Ten Eigenvalues from the Principal Component Analysis....... 72
APPENDIX E: Principal Component Analysis Scree Plot 74
APPENDIX F: Rater Agreement by Item... 78
APPENDIX G: IAA Standard Validation Evaluation form... 87
APPENDIX H: Alignment Study 89

1. PURPOSE AND DESIGN OF THE IAA TESTING PROGRAM

In 1997, the Illinois Standard Achievement Test (ISAT) was authorized by state law to measure how well students learned the knowledge and skills identified in the Illinois Learning Standards. The Illinois Alternate Assessment (IAA) was added to the assessment program in 2000 to meet the requirements of the Individuals with Disabilities Education Act of 1997 (IDEA) and the No Child Left Behind Act (NCLB) of 2001. These laws mandated that an alternate assessment be in place for those students with significant cognitive disabilities who are unable to take the standard form of the state assessment even with accommodations. Eligibility for participation in the IAA is determined by the student's Individualized Education Program (IEP) team. The original IAA was a portfolio-based assessment. In 2006, Pearson was contracted by the Illinois State Board of Education (ISBE) to develop, administer and maintain a new IAA. Writing, the first subject area developed for this new assessment was piloted in the fall of 2006 and administered operationally in the spring of 2007. Reading, Mathematics, and Science subject areas for the IAA were developed and piloted in fall 2007, and operationally administered in spring 2008.

This Technical Manual provides technical information of the 2009 IAA tests. It addresses test development, implementation, scoring, and technical attributes of the IAA. Other sources of information regarding the IAA, provided in paper or online format, include the *IAA Implementation Manual* and training materials are not included in this manual.

NCLB Requirements

In December 2003, the US Department of Education released regulations allowing states to develop alternate achievement standards for students with the most significant cognitive disabilities. These standards had to have the same characteristics as grade-level achievement standards; specifically, they must be aligned with the State's academic content standards, they must describe at least three proficiency levels, reference the competencies associated with each achievement level, and include cut scores that differentiate among the levels. The regulations also stipulated that a recognized and validated procedure must be used to determine each achievement level.

States were not required to adopt alternate achievement standards. However, if they chose to do so, the standards and the assessment used to measure students with the most significant cognitive disabilities against those standards would be subject to federal peer review. The *Alternate Achievement Standards for Students with the Most Significant Cognitive Disabilities: Non-regulatory Guidance* (2005) provides guidance on developing alternate achievement standards specified states could develop alternate assessments based on alternate achievement standards, but provided little guidance as to the format of these assessments, other than stipulating they must meet the same requirements as all other assessments under Title I, i.e., the same technical requirements as the regular assessment.

The non-regulatory guidance provides states significant latitude in designing the format of alternate assessments for alternate achievement standards. They specifically state that there is no typical format and suggest that an alternate assessment may reduce the breadth and or depth of those standards (US Department of Education, 2005, p.16). Essentially, the US Department of Education has indicated that it is most concerned with the technical adequacy of the alternate assessments and their alignment with state content standards. Provided states follow best psychometric practices in developing their alternate assessments and document their processes, the format of any alternate assessment is secondary to the requirement to measure the content standards.

The most relevant NCLB requirements for the IAA were those that had been explicitly addressed to ISBE through the peer review letter. Points that were made regarding the IAA are provided below and have been addressed and documented in the work Pearson and ISBE have completed and/or planned under the current IAA contract:

4.0 - TECHNICAL QUALITY

5. Documentation of the technical adequacy of the Illinois Alternate Assessment (IAA):

    - The use of procedures for sensitivity and bias reviews and evidence of how results are used; and

    - Clear documentation of the standard-setting process.

5.0 – ALIGNMENT

5. Details of the alignment study planned for the IAA. This evidence should include the assurance that tasks used are appropriately aligned/linked to the academic performance indicators.

Excerpts from the August 2005 Non-Regulatory Guidance

According to the December 9, 2003 regulation, and as determined by each child's IEP team, students with disabilities may, as appropriate, now be assessed through the following means, as appropriate:

- The regular grade-level State assessment
- The regular grade-level State assessment with accommodations, such as changes in presentation, response, setting, and timing (see http://education.umn.edu/NCEO/OnlinePubs/Policy16.htm).
- Alternate assessments aligned with grade-level achievement standards

– Alternate assessments based on alternate achievement standards.

The 2004 IDEA amendments reinforce the principle that children with disabilities may be appropriately assessed through one of these four alternatives. To qualify as an assessment under Title I, an alternate assessment must be aligned with the State's content standards, must yield results separately in both reading/language arts and mathematics, and must be designed and implemented in a manner that supports use of the results as an indicator of AYP. Alternate assessments can measure progress based on alternate achievement standards and can also measure proficiency based on grade-level achievement standards. Alternate assessments may be needed for students who have a broad variety of disabilities; consequently, a State may employ more than one alternate assessment.

When used as part of the State assessment program, alternate assessments must have an explicit structure, guidelines that determine for which students may participate, clearly defined scoring criteria and procedures, and a report format that communicates student performance in terms of the academic achievement standards defined by the State. The requirements for high technical quality, as set forth in 34 C.F.R. §§200.2(b) and 200.3(a)(1), include validity, reliability, accessibility, objectivity, and consistency with nationally recognized professional and technical standards, all of which apply to both alternate assessments and regular State assessments.

Test Development and Test Blueprint

In the spring of 2006, a team of Illinois educators created the new Illinois Alternate Assessment Frameworks (refer to www.isbe.net/assessment/iaa.htm). The purpose of the frameworks is to prioritize skills and knowledge from the Illinois Learning Standards in order to develop a new Illinois Alternate Assessment for students who have the most significant cognitive disabilities. Pearson was responsible for facilitating the development of the IAA Frameworks and providing statewide staff development on how to access grade-level curriculum.

The first task was to define the critical function; what the educators expect ALL students to know or to do in order to meet an assessment objective. Pearson trained a group of educators to assist in development of the IAA Frameworks by starting with the intent of the standard, providing examples of how a variety of students can access the standard and related curricula and materials, and then defining the critical function based on this work. The educators were reminded that students taking the IAA will receive instruction on grade level content standards (maybe at a lower complexity level) within the context of grade level curriculum, ensuring that the intent of the grade level content standard remains intact through the alignment process.

ISBE contracted Pearson and their subcontractor partners, the Inclusive Large Scale Standards and Assessment (ILSSA) group, and Beck Evaluation and Testing Associates, Inc. (BETA) in 2006 to develop the new IAA in grades 3–8 and 11 for

Reading and Mathematics; in grades 4, 7, and 11 for Science; and in grades 3, 5, 6, 8, and 11 for Writing. The Pearson team, working with ISBE and the Assessment Committee for Students with Disabilities (ACSD), developed an item-based assessment that includes performance tasks to best measure achievement through links to the Illinois Learning Standards.

An item-based assessment provides more objective measurement than does a portfolio-based alternate assessment, and requires less teacher and student time to administer. Several factors were taken into consideration during planning and development of the IAA program including:

- The IAA will reflect the breadth and depth of content of the tested content areas and grade level.
- The IAA will promote access to the general curriculum.
- The IAA will reflect and promote high expectation and achievement levels.
- The IAA will allow access to students with the most significant cognitive impairments, including those with sensory impairments.
- The IAA will be free from racial, gender, ethnicity, socioeconomic, geographical region, and cultural bias.
- The IAA will not increase the teachers' burden to assess and is non-obtrusive to the instructional process.
- The IAA will meet federally mandated requirements.

Besides being based on instructional activities in the general curriculum, the test development utilized the theory and elements of Universal Design for Learning. Specifically, multiple means of expression and representation were addressed. In addition, an alternate assessment design specialist from BETA recommended instructional and assessment strategies that could be used effectively with the test.

The IAA is administered on a one-on-one basis by qualified and trained teachers. Training was provided to teachers prior to the administration. Although IAA items are in multiple-choice format, the scoring is done through a 1–4 point scoring rubric. The rubric was developed in collaboration with the ISBE, the ACSD, and educators.

The item format was modified after the pilot test and before construction of the 2008 test. An analytical study was conducted to investigate the impact of the modification of the test format. The results of this study showed virtually no difference in the performance of these two item types. In other words, this modification would not significantly alter the fall 2007 pilot test results such that they would be unusable for data and bias review (refer to the *IAA 2008 Technical Manual*). A more cautious approach, however, was taken to minimize any potential impacts of format change. The IAA originally intended to be a pre-equated test with the item statistics derived from the fall 2006 and fall 2007 pilot tests was changed to a post-equating model. In light of this, it was decided that item statistics from the fall 2007 pilot test would not be submitted to the item bank. Instead, only item statistics for items administered operationally or in field-test positions from the spring 2008 and future administrations would be included in the item bank.

In 2009, the IAA is further improved in two respects; it has a standardized test administration procedure and an increased test length. Standardization of IAA administration is achieved through three ways: (1) it incorporates supplement testing materials into test booklet, (2) it uses a prescriptive scoring rubric to increase consistency in scoring, and (3) it inserts the rubric in the booklet for convenience in the administration process. The 2009 blueprint of census item for each subject is listed in Table 1.1a through Table 1.1d. Test lengths of the 2008 and 2009 census items can be found in Table 5.2.

Table 1.1a: Reading Blueprint

Grade	Goal	Number of Items	Percent of Items
03	1	11	79
03	2	3	21
04	1	10	71
04	2	4	29
05	1	9	64
05	2	5	36
06	1	9	64
06	2	5	36
07	1	10	71
07	2	4	29
08	1	9	64
08	2	5	36
11	1	11	100

Table 1.1b: Mathematics Blueprint

Grade	Goal	Number of Items	Percent of Items
03	6	6	40
03	7	2	13
03	8	3	20
03	9	2	13
03	10	2	13
04	6	6	40
04	7	3	20
04	8	2	13
04	9	2	13
04	10	2	13
05	6	5	33
05	7	3	20
05	8	2	13
05	9	3	20
05	10	2	13
06	6	4	27
06	7	2	13
06	8	3	20
06	9	3	20
06	10	3	20
07	6	3	20
07	7	3	20
07	8	3	20
07	9	3	20
07	10	3	20
08	6	4	27
08	7	2	13
08	8	4	27
08	9	2	13
08	10	3	20
11	6	5	33
11	7	2	13
11	8	2	13
11	9	4	27
11	10	2	13

Table 1.1c: Science Blueprint

Grade	Goal	Number of Items	Percent of Items
04	11	2	13
04	12	10	67
04	13	3	20
07	11	3	19
07	12	11	69
07	13	2	13
11	11	2	13
11	12	11	73
11	13	2	13

Table 1.1d: Writing Blueprint

Grade	Goal	Number of Items	Percent of Items
03	3	7	100
05	3	7	100
06	3	7	100
08	3	7	100
11	3	7	100

Item Development

Item Development Cycle

New items are acquired each year to establish an adequate item pool for test construction. The planning of new item development is based on content coverage and the number of test items needed for the test. Each new item is evaluated by content experts and teacher panels through qualitative and quantitative approaches before use in a test. The cycle of IAA item development is described as follows:

1. **Information Gathering** – review ISBE's documentation, attend planning meetings, synthesize item and test specification, and determine plans for releasing items.
2. **Project-specific Document Creation** – develop project development plans and content- and state-specific task writer training materials.
3. **Item Writer Recruitment and Training** – recruit and train potential writers on industry best practices and IAA-specific styles and item requirements. ISBE reviews training, preparation, and presentation materials and participates in face-to-face, web-based, and/or conference call training.
4. **Item Development** – procure items; review and edit items created by item writers to address source and content accuracy, alignment to curriculum

and/or test specifications, principles of Universal Design, grade and cognitive level appropriateness, level of symbolic communication, scorability with the rubric, and language usage; copy edit for sentence structure, grammar, spelling and punctuation; create art; evaluate tasks for potential bias/sensitivity concerns.

5. **Independent Review** – review by content specialists for overall task quality and alignment to ISBE's *Guidelines for Test Development* and the test specifications.
6. **Initial Customer Review** – review by and feedback from ISBE staff on a sampling of approximately 20 items per subject early in the development cycle to check for a common understanding of ISBE expectations for quality and for content and cognitive mapping.
7. **Committee Reviews** – review of passages and items by Illinois stakeholders for content and bias/sensitivity with Pearson staff. Items that are suspected having bias are not used in the test.
8. **Pilot Test Item Selection** – pilot test as a way to collect item information for quantitative evaluation. Pilot test items are selected from the items that passed the Committee Review. This selection is a cooperative effort between the Pearson and ISBE staff. These pilot test items are embedded in the census test to reduce field test effect.
9. **Pilot Test Administration** – test embedded pilot items along with census items. The IAA is tested annually between February and April.
10. **Data Review** – perform different item analyses on the pilot test items after test administration. The analyses results are presented to teacher panels for item quality review. Teacher panels are reminded in the Data Review meeting to use the statistics as a reference; the main purpose of the meeting is to review item quality through content and standard alignment aspects.
11. **Census Item Selection** – use census items for scoring. Items accepted in the Data Review meeting are eligible for census items. Based on test blueprint and the test design, Pearson and ISBE content experts work closely selecting census items. Psychometric review of item and test statistics is implemented to secure quantitative quality of the test.
12. **Census Test Administration** – test census items along with pilot items. The IAA is tested annually between February and April.

Item Specifications

The following is a general description of the Illinois student population being assessed by the IAA. This description was used as context for item development purposes only. These students have, or function as if they have, significant cognitive disabilities. Students in this population most likely:

- Have both physical and mental disabilities, and
- Use an alternate form of communication

These students exist along a disability continuum—some students may have one of the more severe forms of autism, some may have Down syndrome and others may have multiple cognitive and physical impairments that severely limit their ability to function in the classroom.

Based on this understanding of the population to be tested, the IAA items and stimuli were written in accordance with the following Universal Design principles to promote the maximization of readability and comprehensibility (see Synthesis Report 44)[1]:

1. Simple, clear, commonly-used words should be used, and any unnecessary words should be eliminated.
2. When technical terms must be used, they should be clearly defined.
3. Compound complex sentences should be broken down into several short sentences, stating the most important ideas first.
4. Only one idea, fact, or process should be introduced at a time; then develop the ideas logically.
5. All noun-pronoun relationships should be made clear.
6. When time and setting are important to the sentence, place them at the beginning of the sentence.
7. When presenting instructions, sequence steps in the exact order of the occurrence.
8. If processes are being described, they should be simply illustrated, labeled, and placed close to the text they support.

By applying writing and editing guidelines that promote clarity in language, style, and format, the IAA assessments maximize accessibility so students may better show what they know and are able to do. Following best practices in item writing for alternate assessments and the Universal Design philosophy, writers and editors were directed to adhere to strategies such as those outlined in the Table 1.2.

[1] Thompson, S. J., Johnstone, C. J., & Thurlow, M. L. (2002). *Universal design applied to large scale assessments* (Synthesis Report 44). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved August 19, 2003, from the World Wide Web: http://education.umn.edu/NCEO/OnlinePubs/Synthesis44.html.

Table 1.2. Plain Language Editing Strategies (from Synthesis Report 44)

Strategy	Description
Reduce excessive length.	Reduce wordiness and remove irrelevant material. Where possible, replace compound and complex sentences with simple ones.
Eliminate unusual or low frequency words and replace with common words.	For example, replace "utilize" with "use."
Avoid ambiguous words.	For example, "crane" could be a bird or a piece of heavy machinery.
Avoid irregularly spelled words.	For example, "trough" and "feign."
Avoid proper names.	Replace proper names with simple, common names such as first names.
Avoid inconsistent naming and graphic conventions.	Avoid multiple names for the same concept. Be consistent in the use of typeface.
Avoid unclear signals about how to direct attention.	Well-designed headings and graphic arrangement can convey information about the relative importance of information and the order in which it should be considered. For example, phrases such as "in the table below,…" can be helpful.
Mark all questions.	When asking more than one question, be sure that each is specifically marked with a bullet, letter, number, or other obvious graphic signal.

Qualifications of Item Writers and Method of Recruitment

The item writers were selected to represent the Illinois general and special educators, whose names were provided by ISBE for item writer recruitment. Table 1.3 provides the number of item writers who worked on the IAA tasks by subject. Since Writing had adequate items in the bank, there was no need to develop new items; thus, Writing item writers were not invited for this meeting.

Table 1.3: Number of Item Writers by Subject

Subject	Number of Item Writers
Mathematics	10
Reading	18
Science	6
Writing	0

Training Practices/Activities (in consideration of both content and bias)

All item writers are trained prior to performing their task. Training is presented by Pearson content specialist staff during the Item Writer Workshop. During the item writer training, materials are reviewed and discussed in detail, and sample items are submitted by the item writers. A general description of the IAA population and the IAA administration approach are also discussed.

Test Administration Training

Given that the IAA is administered by teachers to each of their students individually, standardization of the test administration is essential to the validity of the test. Thus, test administration training is put in place to bring teachers/administers to the same level of understanding. Training materials are developed and presented by Pearson in collaboration with ISBE at regional settings across Illinois.

Test Implementation Manual

The *IAA Test Implementation Manual* was developed by Pearson for ISBE using input from best practices and the field. Within the test implementation manual, the teacher can find all information necessary to prepare for, administer, and provide scores back to Pearson for the IAA. Additionally, links to teacher training material for the IAA are also included in the manual to be used as a refresher course. The manual is available online at www.isbe.net/assessment/iaa.htm.

Test Booklets

The 2009 IAA test booklets incorporate supplemental art and scoring rubrics for each individual item. This modification significantly increases the consistency among teachers of test administration. Each test booklet contains a set of census items and subset of embedded pilot test items. Items are scored using a four-point rubric that is provided in Appendix A.

Answer Sheets

The IAA answer sheets have been developed by Pearson and ISBE to be user friendly, efficient means of data capture. The answer sheet is located on the back cover of the Implementation Manual and posted online. Teachers record the student's scores on the answer sheet during test administration and then transfer the scores to the online platform at a later time.

Online Test Platform

Pearson *School Success* group provides an online platform for teachers to use in IAA score submission. Training for the online platform is provided by Pearson to teachers and test coordinators statewide. The online platform speeds data collection and minimizes student identification errors.

Teacher Training

Training Objectives

- Increase participants' familiarity with IAA calendar of events and timeline expectations.
- Improve participants' understanding of the Illinois Learning Standards and IAA Frameworks.
- Promote scoring reliability and validity through practice exercises using the newly devised IAA rubric.
- Present video clips of students engaged in the IAA to explore educators' rationale for score assignment and test preparation efforts.
- Detail best practices for test administration including assessment procedures, emphasis on students' primary mode of communication, materials modification, and creating optimal testing environments.
- Offer guidelines for materials modification, including the receipt, verification and return of secure test materials.
- Demonstrate capabilities of the online scoring tool.

Training Logistics

- Throughout January and February of 2009, Pearson, in partnership with ISBE, conducted multiple onsite trainings in locations statewide in preparation for the spring 2009 operational assessment.
- Each session was attended by approximately 100 Illinois IAA Coordinators and educators.

Training Facilitators

- Each onsite session was co-facilitated by Pearson and ISBE representatives.

Training Materials

- All materials in support of the IAA Regional Trainings and spring 2009 test administration were developed by Pearson in consultation with and approval from ISBE.
- Materials were accessible to educators via the ISBE IAA website at www.isbe.net/assessment/iaa.htm and/or distributed to Illinois educators in conjunction with IAA's spring 2009 packaging and distribution requirements
- Regional Training materials included an PowerPoint presentation, IAA rubric, student video clips, sample answer document to acquaint participants with required data fields that were used in the spring 2009 operational
- Test administration resources included the IAA Frameworks, the 30-page *Test Implementation Manual*, *Online User Guides for Teachers*, *Coordinators and Scoring Monitors*, and test books

Data Review Outcomes

One of the important aspects of test development is to provide fair and accurate ability estimates for all subgroups within the population. In order to achieve such goal, all IAA items are screened for potential bias by teacher panels, administrators, and vendor content experts. Items are checked during three stages: item writing, item review, and data review. First, all of the teachers who are involved in item writing are trained and instructed to balance ethnic and gender references and to avoid gender and ethnic stereotypes. Then, another group of teachers is invited to the item review meetings to screen for potential language and content bias. Items approved by the item review committee are pilot tested and analyzed for differential item functioning. Last, in data review meetings, Illinois administrators, vendor content experts, and a group of teachers review each item based on statistical inputs.

Differential Item Functioning Analyses

Differential item functioning (DIF) analysis is a statistical approach for screening potential item bias. DIF assesses whether an item presents different statistical characteristics for different groups of students after matching their abilities. It is important to note that DIF might cause by actual differences in relevant knowledge of individual item or statistical Type 1 error. As a result, DIF statistics are only used to identify potential item bias presence, not to determine the existence of item bias. Subsequent review by content experts and teacher committees are required to determine the source and meaning of performance differences.

DIF analysis is conducted between two groups; such as male versus female, white versus black, and white versus Hispanics. Male and white are usually referred to as the reference group, and the others are focal group. DIF procedures consist of three steps, first matches student abilities through total raw scores or latent ability, theta. Then compute the subgroup average performance of each matching ability levels. Last, test the significance of the total of subgroup average performance across matching levels. If the totals between the two subgroups are statistically different, the item is flagged for closer inspections. In the IAA DIF analyses, DIF statistics were estimated for subgroups of Black, Hispanic, and Female. Items with statistically significant differences in performance are flagged and present to teacher committees.

Two statistical indices are used to identify DIF in the IAA pilot items. First, the Mantel-Haenszel statistics that provided by the WINSTEPS program are used. The second DIF index is the Cohen's effect size estimate, d. Cohen (1988) defined d as the difference between subgroup means, M_a - M_b, divided by the pooled standard deviation, S_{pooled}. The pooled standard deviation is found as the root mean square of the standard deviations of the two subgroups (refer to equations 1.1 and 1.2).

$$d = M_a - M_b \ / \ S_{pooled}, \text{ where} \tag{1.1}$$

$$S_{pooled} = \sqrt{(S_a^2 + S_b^2)/2} \ . \tag{1.2}$$

Cohen suggested the effect sizes could be grouped into three categories: small (d = 0.2), medium (d = 0.5), and large (d = 0.8). Based on the Mantel-Haenszel statistics and Cohen's d, the IAA items are flagged into four DIF categories (0–3) defined by Pearson:

0 = No Indication of DIF

1 = Slight Indication of DIF

2 = Possible Indication of DIF

3 = DIF Indicated

The flagging rules are as follows:

1. If the Mantel-Haenszel statistic is *not* significant at the α = .05 level, *and* if the Cohen's *d* is smaller than suggested medium value, the item is considered no potential bias and receives a flag value of "0".
2. If the Mantel-Haenszel statistic is significant at the α = .05 level, *or* if the Cohen's *d* is greater or equal to the suggested medium value yet smaller than the large value, the item is considered having slight DIF and receives a flag value of "1".
3. If the Mantel-Haenszel statistic is significant at the α = .05 level, *and* if the Cohen's *d* is greater or equal to the suggested medium value yet smaller than the large value, the item might possibility has DIF and receives a flag value of "2".
4. If the Mantel-Haenszel statistic is *not* significant at the α = .05 level, *and* the Cohen's *d* is greater or equal to the large value, the item might possibility has DIF and receives a flag value of "2" as well.
5. If the Mantel-Haenszel statistic is significant at the α = .05 level, *and* the Cohen's *d* is greater or equal to the large value, the item receives a flag value of "3".

Table 1.2 summarizes the items selected as cores that present DIF. Note that items from the 'No Indication of DIF' category were the first chosen for test construction. However, when items from the category did not adequately fulfill the blueprint, items from the 'Slight Indication of DIF' category were selected. If the blueprint still was incomplete after choosing the 'Slight Indication of DIF' category items, then items from the next category were considered, and so forth.

Table 1.2: DIF between Male/Female, White/Black, and White/Hispanic

Subject	Grade	Male/Female			White/Black			White/Hispanics		
		1	2	3	1	2	3	1	2	3
Reading	3	4	0	0	0	0	0	1	0	0
	4	0	0	0	2	0	0	3	0	0
	5	1	0	0	0	0	0	0	0	0
	6	0	0	0	0	0	0	0	0	0
	7	0	0	0	1	0	0	0	0	0
	8	1	0	0	3	0	0	0	1	0
	11	0	0	0	2	0	0	0	0	0
Mathematics	3	0	0	0	1	0	0	0	0	0
	4	2	0	0	0	0	0	0	0	0
	5	1	0	0	1	0	0	0	0	0
	6	0	0	0	1	0	0	1	0	0
	7	0	0	0	0	0	0	1	0	0
	8	0	0	0	0	0	0	2	0	0
	11	0	0	0	1	0	0	0	0	0
Science	4	1	0	0	0	0	0	0	0	0
	7	0	0	0	1	0	0	0	0	0
	11	1	0	0	1	0	0	1	0	0
Writing	3	0	0	0	0	0	0	0	0	0
	5	1	0	0	0	0	0	0	0	0
	6	1	0	0	0	0	0	1	0	0
	8	1	0	0	0	0	0	1	0	0
	11	1	0	0	1	0	0	0	0	0

Note: 1 = Slight Indication of DIF, 2 = Possible Indication of DIF, 3 = DIF Indicated

2. RELIABILITY AND GENERALIZABILITY

The reliability of a test reflects the degree to which test scores are free from errors of measurement that arise from various sources. Test reliability indicates the extent to which differences in test scores reflect real differences in the construct being measured across some variation in one or more factors, such as time or specific test items used. Different coefficients can be distinguished accordingly. For example, test-retest reliability measures the extent to which scores remain constant over time. A low test-retest reliability coefficient means that a person's scores are likely to shift unpredictably from one time to another. Generalizability, which may be thought of as a liberalization of classical theory (Feldt & Brennan, 1989, p. 128), treats these error components and their impact on score precision singly and in interaction.

Internal Consistency of Overall Scores

Because achievement test items typically represent only a relatively small sample from a much larger domain of suitable questions, the test score consistency (generalizability) across items is of particular interest. That is, how precisely will tests line up students if different sets of items from the same domain are used? Unless the lineups are very similar, it is difficult or impossible to make educationally sound decisions on the basis of test scores. This characteristic of test scores is most commonly referred to as *internal consistency*, which is quantified in terms of an index called Cronbach's coefficient alpha. The Cronbach's alpha (1951) is defined as:

$$\alpha = \left(\frac{n}{n-1}\right)\left(1 - \frac{\sum_i \sigma_i^2}{\sigma_X^2}\right), \qquad (2.1)$$

where n is the number of items in the test, σ_i^2 is the variance of the i^{th} item, and σ_X^2 is the variance of the test score X. The coefficient, which can range from 0.00 to 1.00, corresponds to a generalizability coefficient for a person by item design or, more broadly, as a generalizability coefficient for the person by item by occasions design with one fixed occasion and k randomly selected items (Feldt & Brennan, 1989, p 135). Most well-constructed achievement tests have values above .90.

Table 2.1 presents alpha coefficients for the tests administered in the assessment. Included with coefficient alpha in the table is the number of students responding to the test, the mean score obtained, the standard deviation of the scores, and the standard error of measurement (SEM). As the table shows, the IAA tests are highly reliable, since the alpha coefficients are comparable to or higher than those typically reported in the literature. Note that the IAA is a relatively short test (under 20 items). The high reliability might benefit from standardized administration and clear scoring guidelines.

Standard Error of Measurement

Based on the classical test theory (CTT), the standard error of measurement (SEM) is the degree to which chance fluctuation in test scores that may be expected. The SEM represents inconsistencies occurring in repeated observations of observed scores around a student's true test score, which is assumed to remain constant across repeated measurements of the same trait in the absence of instruction. The SEM is inversely related to the reliability of a test; the greater the reliability is, the smaller the SEM, and the more confidence the test user can have in the precision of the observed test score. The CTT SEM is calculated with the formula:

$$\text{CTT SEM} = SD_X \sqrt{1 - r_{XX}}\ , \qquad (2.2)$$

where SD_x is the standard deviation of observed test scores and r_{xx} is the test reliability.

The SEM can be helpful for quantifying the extent of errors occurring on a test. A standard error of measurement band placed around the student's true score would result in a range of values most likely to contain the student's observed score. The observed score may be expected to fall within one SEM of the true score 68 percent of the time, assuming that measurement errors are normally distributed.

Table 2.1: Reliability Estimates: Whole Population

Subject	Grade	N	Mean	SD	Alpha	SEM
Reading	3	1959	45.65	11.38	0.94	2.86
	4	1942	44.73	11.51	0.94	2.81
	5	1862	44.69	11.28	0.94	2.83
	6	1904	46.36	11.11	0.94	2.63
	7	2011	47.06	10.47	0.94	2.60
	8	1964	46.84	10.68	0.94	2.60
	11	1977	38.33	8.91	0.95	1.94
Mathematics	3	1958	47.99	11.71	0.93	3.06
	4	1941	49.12	12.33	0.95	2.79
	5	1860	49.78	12.00	0.95	2.72
	6	1896	50.10	11.68	0.95	2.67
	7	2009	49.07	11.32	0.94	2.85
	8	1963	50.56	11.43	0.95	2.63
	11	1968	49.30	12.15	0.95	2.65
cience	4	1939	47.93	12.29	0.95	2.86
	7	2007	53.90	11.83	0.95	2.72
	11	1972	51.01	11.88	0.96	2.47
Writing	3	1955	22.23	6.18	0.90	1.97
	5	1854	22.37	5.90	0.90	1.91
	6	1899	22.78	5.78	0.90	1.86
	8	1961	23.69	5.50	0.90	1.77
	11	1974	23.91	5.82	0.92	1.62

Table 2.1a: Reliability Estimates by Ethnicity

Grade	Subgroup	Reading	Mathematics	Science	Writing
3	Missing	0.93	0.92		0.89
	Asian	0.92	0.92		0.91
	Black	0.95	0.94		0.92
	Hispanic	0.95	0.95		0.92
	Multiple	0.94	0.90		0.89
	White	0.93	0.92		0.88
4	Missing	0.94	0.95	0.95	
	Asian	0.95	0.96	0.95	
	Black	0.94	0.96	0.95	
	Hispanic	0.95	0.96	0.95	
	Multiple	0.92	0.92	0.92	
	White	0.93	0.94	0.93	
5	Missing	0.94	0.94		0.90
	Asian	0.94	0.96		0.89
	Black	0.95	0.96		0.91
	Hispanic	0.94	0.95		0.89
	Multiple	0.96	0.95		0.94
	White	0.93	0.94		0.89
6	Missing	0.95	0.95		0.91
	Asian	0.96	0.97		0.95
	Black	0.94	0.94		0.90
	Hispanic	0.95	0.96		0.91
	Multiple	0.95	0.95		0.88
	White	0.94	0.94		0.88
7	Missing	0.94	0.94	0.95	
	Asian	0.92	0.91	0.92	
	Black	0.95	0.95	0.96	
	Hispanic	0.94	0.93	0.94	
	Multiple	0.91	0.82	0.82	
	White	0.93	0.93	0.94	
8	Missing	0.95	0.95		0.90
	Asian	0.94	0.95		0.91
	Black	0.94	0.95		0.90
	Hispanic	0.94	0.95		0.89
	Multiple	0.95	0.96		0.89
	White	0.94	0.94		0.89
11	Missing	0.96	0.96	0.96	0.93
	Asian	0.96	0.98	0.97	0.96
	Black	0.96	0.96	0.96	0.94
	Hispanic	0.95	0.96	0.95	0.92
	White	0.94	0.94	0.95	0.90

Note 1: *N* counts of Native Americans are smaller than 21 for all grades.
Note 2: Grade 11 *N* count for Multiple Ethnicity is smaller than 21.

Table 2.1b: Reliability Estimates by LEP

Grade	Subgroup	Reading	Mathematics	Science	Writing
3	Missing	0.93	0.92		0.89
	LEP	0.93	0.94		0.90
	Non-LEP	0.94	0.93		0.90
4	Missing	0.94	0.95	0.95	
	LEP	0.95	0.95	0.95	
	Non-LEP	0.94	0.95	0.94	
5	Missing	0.94	0.94		0.90
	LEP	0.93	0.95		0.86
	Non-LEP	0.94	0.95		0.90
6	Missing	0.95	0.95		0.91
	LEP	0.97	0.98		0.95
	Non-LEP	0.94	0.94		0.89
7	Missing	0.94	0.94	0.95	
	LEP	0.95	0.95	0.95	
	Non-LEP	0.94	0.93	0.95	
8	Missing	0.95	0.95		0.90
	LEP	0.94	0.96		0.91
	Non-LEP	0.94	0.95		0.89
11	Missing	0.96	0.96	0.96	0.93
	LEP	0.91	0.90	0.94	0.90
	Non-LEP	0.95	0.95	0.95	0.92

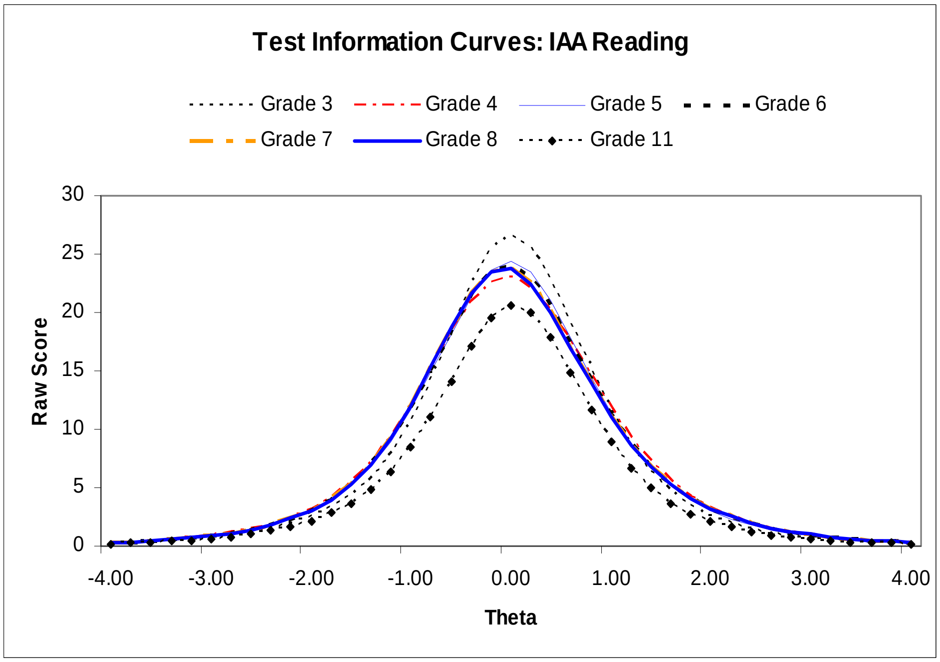
Table 2.1c: Reliability Estimates by Income

Grade	Subgroup	Reading	Mathematics	Science	Writing
3	Missing	0.93	0.92		0.89
	Low Income	0.94	0.94		0.90
	Not Low Income	0.94	0.93		0.90
4	Missing	0.94	0.95	0.95	
	Low Income	0.94	0.95	0.95	
	Not Low Income	0.94	0.95	0.94	
5	Missing	0.94	0.94		0.90
	Low Income	0.93	0.94		0.88
	Not Low Income	0.94	0.95		0.90
6	Missing	0.95	0.95		0.91
	Low Income	0.95	0.95		0.90
	Not Low Income	0.94	0.94		0.89
7	Missing	0.94	0.94	0.95	
	Low Income	0.94	0.94	0.95	
	Not Low Income	0.93	0.93	0.94	
8	Missing	0.95	0.95		0.90
	Low Income	0.94	0.95		0.90
	Not Low Income	0.94	0.94		0.89
11	Missing	0.96	0.96	0.96	0.93
	Low Income	0.95	0.94	0.95	0.91
	Not Low Income	0.95	0.95	0.95	0.92

IRT Test Information Function

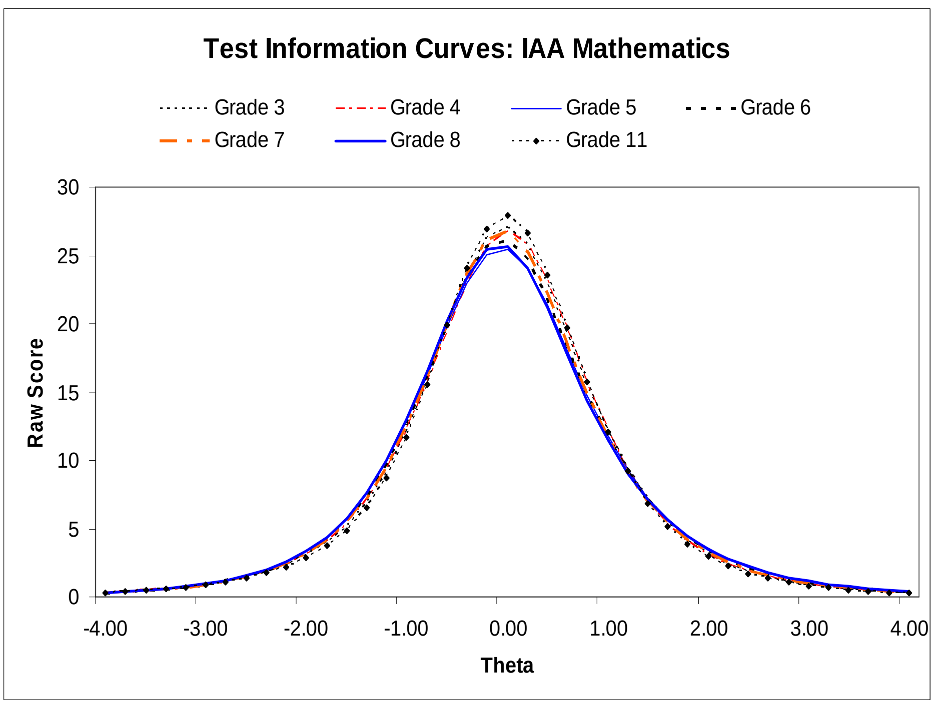
The reliability coefficients reported above were derived within the context of classical test theory and provide a single measure of precision for the entire test. With the Item Response Theory (IRT), it is possible to measure the relative precision of the test at different points on the scale. The amount of information at any point is directly related to the precision of the test. That is, precision is the highest where information is highest. Conversely, where information is the lowest, precision is the lowest, and ability is most likely poorly estimated. Figures 2.1–2.4 present the test information functions for the IAA Reading, Mathematics, Science, and Writing tests.

Figure 2.1: IAA Reading Grades 3-8 and Grade 11 Test Information Functions



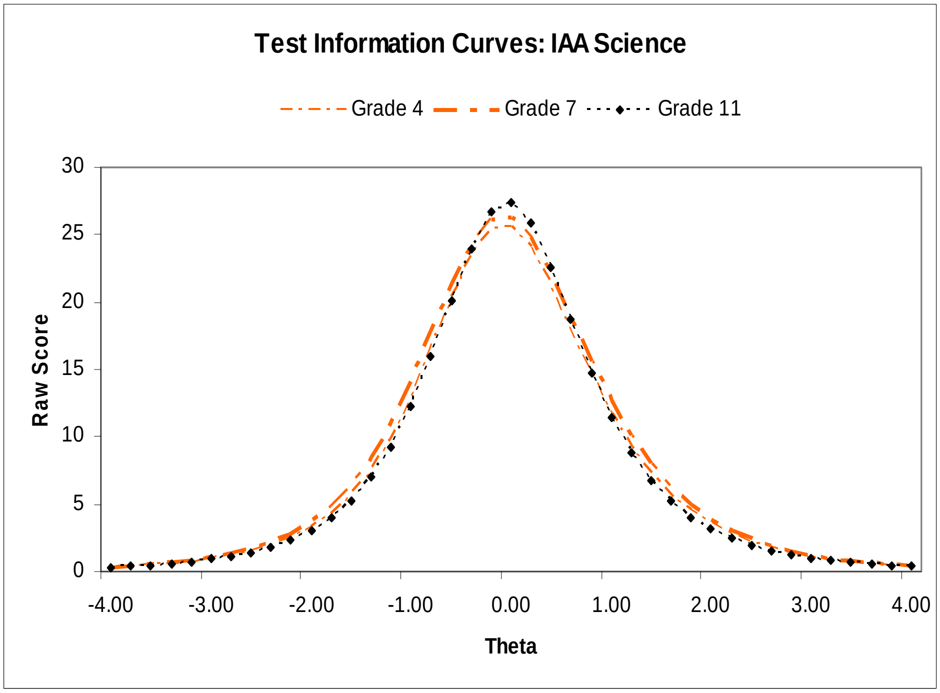
Note: Grades 3-8 have 14 items and grade 11 has 11 items.

Figure 2.2: IAA Mathematics Grades 3-8 and Grade 11 Test Information Functions



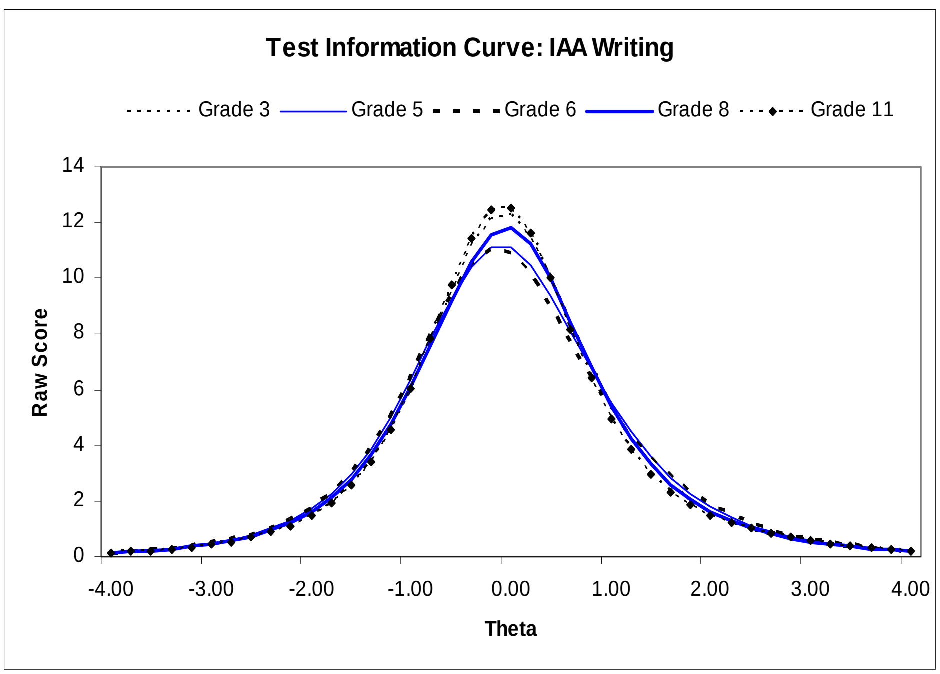
Note: Mathematics has 15 items for all grades.

Figure 2.3: IAA Science Grades 4, 7, and 11 Test Information Functions



Note: Science grades 4 and 11 have 15 items and grade 7 has 16 items.

Figure 2.4: IAA Writing Grades 3, 5, 6, 8, and 11 Test Information Functions



Note: Writing has 7 items for all grades.

IRT Conditional SEM

The standard error of measurement (SEM) reflects the degree of error in student scores. Classical test theory has a fixed SEM value for all students, but the SEM of item response theory varies across the ability range; thus, it is also referred to as the conditional SEM. The conditional SEM is defined as follows:

$$SEM(\theta) = \frac{1}{\sqrt{I(\theta)}}, \qquad (2.3)$$

where $I(\theta)$ is the test information function. The conditional SEM has an inverse normal distribution in which SEM values decrease as it moves toward the center. The SEM is first estimated on a theta scale by subject and grade. When reporting with IAA scale scores, the SEM is transformed onto the IAA scale by applying a scaling slope (see Appendix B).

Reliability of Scores

The IAA items were in a multiple-choice format but scored with a four-point rubric. The teachers administered the test to individual students and then decided the student's score on each item based on the administration guidelines. The reliability and validity of the rubric scores are presented in Chapter 3.

Reliability of Performance Classification

Student performance on the IAA is reported into four categories: Entry, Foundational, Satisfactory, and Mastery. The procedure of defining the cut scores that separate these categories is documented in Chapter 5, Standards Validation. The Standards Validation procedure utilized raw scores to define performance cuts. In order to transform the raw score cuts to scale score cuts, thetas corresponding to those raw score cuts were identified and then transformed to scale scores (refer to Chapter 5 Tables 5.5a – 5.5d).

The reliability of such classifications, which are criterion-referenced, is related to the reliability of the test on which they are based, but they are not identical. Glaser (1963) was among the first to draw attention to this distinction, and Feldt and Brennan (1989) extensively reviewed the topic. As Feldt and Brennan (1989, p. 140) point out, approaches to the development of reliability coefficients for criterion-referenced interpretations of test scores have been based either on squared-error loss or threshold loss. The former, also referred to as classification accuracy, investigates the accuracy of student performance level classification. The accuracy is associated with the error in student score estimates, SEM. For example, a student's true ability is above the Satisfactory level, but due to random measurement error, the student might have an observed score that is below the Satisfactory, vice versa. The impact of SEM is greater when the student's ability is around the cut scores.

Rudner (2005) incorporated the standard error into classification accuracy computation. His formula is expressed as below;

$$p(level_k) = \sum_{\theta=c}^{d} (\phi(\frac{b-\theta}{se(\theta)}) - \phi(\frac{a-\theta}{se(\theta)}))\, f(\theta), \qquad (2.4)$$

and *level* is $(a < \hat{\theta} < b \mid \theta)$. (2.5)

In equations 2.3 and 2.4, θ is the true score, $\hat{\theta}$ is a normally distributed observed score with a mean of θ and a standard deviation of $se(\theta)$. The $\phi(z)$ is the cumulative normal distribution function, and $f(\theta)$ is the standard normal density function. The c and d are cut score intervals, and the a and b are the lower and higher bound of the observed score. In this report, the empirical data distribution is used to compute the $f(\theta)$ and free the model from distribution constraint. This aspect is important for alternate assessments because it has been found that alternate assessment score distributions tend to be highly skewed towards a higher ability range.

An example of Rudner's analysis result is presented in Table 2.8. The R1 through R4 refer to the performance levels of Entry through Mastery for Reading, respectively. The column True is the classification based on the IRT true score estimates. The row Ex is the observed performance level classification. Note that this equation adds the probabilities along the raw score range to obtain the observed percentage. This approach could produce slightly different values (less than 0.5 in differences) than the reported observed percentages. Classification accuracy tables, similar to Table 2.8, for all subjects and grades can be found in Appendix C. The accuracy of each performance level is represented by the values on the diagonal. For example 18.0 is the accuracy of Entry level and 18.3 is the accuracy of Foundational level. The sum of the diagonal values, 18.0, 18.3, 24.1, and 15.3, is the overall test classification accuracy. The overall test classification is presented on Table 2.9 by subject and grade.

Table 2.8: Reading Grade 3 Classification Accuracy

Level	R1	R2	R3	R4	True
R1	18.0	2.9	0.1	0.4	21.3
R2	1.7	18.3	6.7	1.1	27.8
R3	0.0	2.8	24.1	7.1	33.9
R4	0.0	0.0	1.8	15.3	17.0
Ex	19.7	23.9	32.6	23.8	100.0

Table 2.9: Classification Accuracy

Grade	Reading	Mathematics	Science	Writing
3	76	74		71
4	77	77	78	
5	74	78		75
6	77	78		71
7	77	77	77	
8	75	76		68
11	68	79	79	70

3. VALIDITY

Test validity refers to the degree to which a test measures what it is intended to measure. Evidence that supports the validity of a test is gathered from different aspects and through different methods. The three most recognized aspects are content validity, construct validity, and criterion-related validity. Content validity refers to how well a test covers the content of interest. It examines the correspondence between test blueprints that describe the intended content and test items. Construct validity is comprised of analyses of a test's internal constructs in order to confirm that the test indeed functions as it is intended to function. Factor analysis and correlation analysis among test components, such as subtests and items, are two common approaches to examining the construct validity of a test. Criterion-related validity refers to the extent to which relationships between assessment scores and external criterion measures are consistent with the expected relations in the construct being assessed. That is, constructs of an assessment should reasonably account for the external pattern of correlations. A convergent pattern would indicate a correspondence between measures of the same construct (Cronbach & Meehl, 1955; Crocker & Algina, 1986; Clark & Watson, 1995).

Validity is essential to defensible score interpretation and use for any test (Cronbach & Meehl, 1955; Messick, 1995). Without adequate validity evidences, there can be no assurance that a test is measuring the content and construct that are intended. In this chapter, the IAA assessment framework is presented first to guide the evaluation of the IAA validity. Then, the validity of the IAA was examined through three aspects: content validity, construct validity, and criterion-related validity.

Performance-Based Measurement

The development of a validity test relies on appropriate understanding, definition, and measurement of the construct of interest, or as posited by Dawis (1987), an existing, accurate *theory of the scale* for the assessment. In the case of the IAA, the theory of the scale is proposed *a priori* and is the basis for evaluating the validity of the IAA.

Rosenthal & Rosnow (1991) stated that the measurement of actual performance is the gold standard of applied human behavior assessment. The keys to measurement of actual performance are: a) identifying the performance of interest to measure, b) understanding the performance of interest within a larger model of behavior and influencing factors, c) specifying an appropriate measurement model, and c) designing data collection that will best meet model requirements. Many models of human performance exist, from molecular cognitive models to molar models of human performance within organizations (e.g., Naylor & Ilgen, 1984). The selection of an appropriate model depends largely on the level of performance to be measured. For example, student performance related to the demonstration of IAA content standard, grade-level knowledge is not at the molecular cognitive process level, or at the person interacting within the classroom level, but at the level of individual observable performance in response to IAA items. Because of the large variance in individual needs across students coming into the assessment situation for the IAA population, a valid performance model for the IAA is the one that provides both the right type and right amount of standardization in the face of a plethora of meaningful individual difference dimensions. A valid assessment of a common construct across students who are each unique in how they retrieve, process, and convey relevant information is to assess each on the construct using the modality that is appropriate for that student. Construct-relevant factors are held constant, or standardized, and construct-irrelevant factors are allowed to vary according to the student needs.

Based on our work with various relevant performance models, the basic structure of the IAA performance model was posited (Figure 3.1) as a guide for examining the validity of IAA. In this model, standardization is built into the IAA performance items, teacher training, administration materials, scoring rubric, and protocol. Flexibility is provided through each teacher's best judgment of a student's unique needs regarding an assessment modality (i.e., mode of communication). Students interact with and respond to IAA performance items in a manner consistent with their needs and through a knowledgeable teacher's administration. Teacher scoring is standardized through training to a protocol and the use of a rubric validated through expert judgment and field testing. The basic framework of the IAA student performance model is designed such that the students' actual performance is elicited in response to the IAA items administered in a way that the given student's content knowledge is assessed and scored in a standardized manner.

Also included in Figure 3.1 is a validation component of the performance model that involves specially trained subject matter experts (SMEs) with sufficient knowledge of the IAA content, administration, and student population. A detailed description of this validation study can be found in the criterion-related validity section of this chapter.

As implied by the IAA performance model in Figure 3.1 and posited by Messick (1989), validity of the assessment is built up through relevant, integrated factors. The validity of the IAA rests on the content frameworks, assessment materials, teacher training, scoring materials, appropriate flexibility of the assessment item to account for student needs, and the accuracy of teacher scoring. Throughout this technical manual, the validity of these various IAA tests has been presented through logical development processes and qualitative judgments. In the next three sections, three forms of validity evidences are presented: content validity, construct validity, and criterion-related validity.

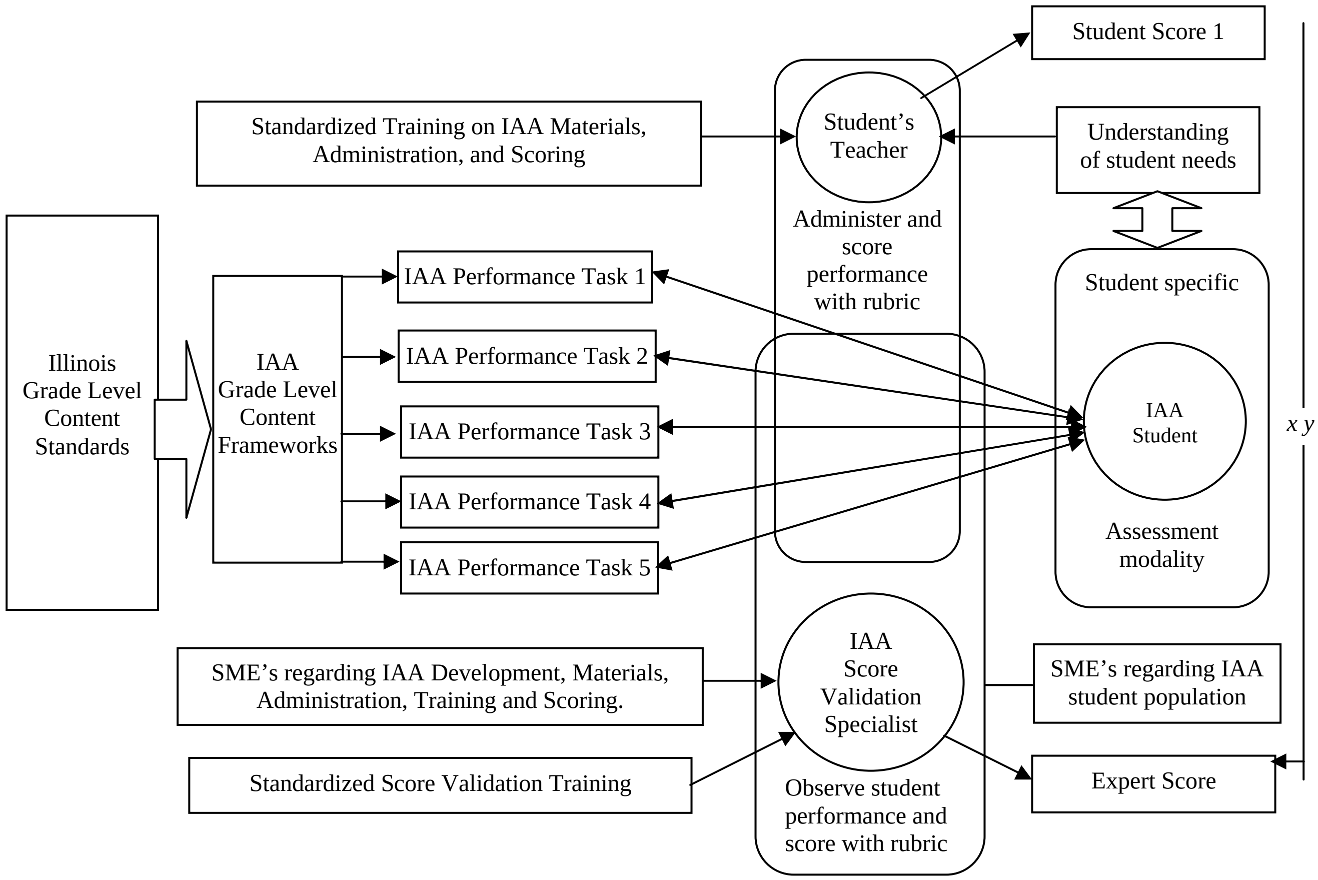


Figure 3.1 IAA performance model with validation component

Content Validity

The content validity of the IAA is established through content standard specification that defines the measurement of actual performance. It is fulfilled through item alignment study, test design, test/item review, and test/item analyses. As described in Chapter I of this report, the IAA measures actual student performance through trained teachers, specified set of content-valid items, test administration that is appropriate to the student's usual communication methods, and a standardized scoring rubric. Evidence of content validity has been detailed in Chapter I, which contains descriptions of the test blueprint, the test construction process, and the decisions made for defining and developing the IAA test. In addition, an alignment study for each subject area was reported in April 2009 by WIDA Consortium (see Appendix E).

Construct Validity

Dimensionality

Dimensionality is a unique aspect of construct validity. Investigation is necessary when item response theory (IRT) is used, because IRT models assume that a test measures only one latent trait (unidimensionality). Although it is generally agreed that unidimensionality is a matter of degree rather than an absolute situation, there is no consensus on what defines dimensionality or on how to evaluate it. Approaches that evaluate dimensionality can be categorized into answer patterns, reliability, components and factor analysis, and latent traits. Components and factor analysis are the most popular methods for dimensionality evaluation (Hattie, 1985; Abedi, 1997).

However, these approaches are best for situations when the score distribution is normal. The IAA scoring method turns the multiple-choice items into polytomous item scores. Distributions of individual item scores and the total scores are often negatively skewed. Additionally, the IAA test length is relative short, between 7 to 16 items. The nature of the IAA data does not fit into those models' normality assumptions. Research on the dimensionality of polytomous items suggested the use of structural equation model or IRT approach. However, mixed results are found and more research is needed on this subject (Thissen & Wainer, 2001; Tennant & Pallant, 2006; Raîche, 2005). Before an appropriate approach is found to deal with the complex data situations of IAA, simple and straightforward approach might provide a better picture of test dimensionality. In this study, the principal component analysis is chosen for its straightforward statistical model in comparison to factor analysis's latent variable approach. Even when normality assumption is violated, the estimation may be degraded but still be worthwhile for investigation purpose (Tabachnick & Fidell, 2007). Additionally, the IRT principal component analysis is conducted to provide supporting evidence for dimensionality.

Principal component analysis (PCA) is a data reduction method. This reduction is achieved by extracting item variances into sets of uncorrelated principal components (i.e., eigenvectors) to discover the dimensionality. Lord (1980) stated that if the ratio of the first to the second eigenvalue is large and the second eigenvalue is close to other eigenvalues, the test is unidimensional. Divgi (1980) expanded Lord's idea and created an index by considering the pattern of the first three factor components (eigenvalues). The Divgi Index examines the ratio of the difference of the first and second eigenvalues over the difference of the second and third eigenvalues. A large ratio indicates a greater difference between the first and second eigenvalues, thus, creating a unidimensional tendency. A cut value of 3 is chosen for the index so that values greater than 3 are considered unidimensional.

Appendix D presents the first ten eigenvalues of the principal component analysis. Table 3.1 lists the Divgi index results by subject and grade. All values are greater than 3, which suggest that all of the IAA tests are unidimensional. Scree plots for the Reading, Mathematics, Science, and Writing Assessment, another reference of dimensionality, are presented in Appendix E, The elbow shaped plots support the unidimensionality conclusion drawn from the Divgi index.

The IRT PCA is estimated through WINSTEPS. Interpretation of IRT PCA is different than previously mentioned PCA because the IRT PCA investigates residuals: the difference between the observed responses and expected estimates, instead of variance. Wright (1996) suggests that if a test is unidimensional, its residuals of extracted components should be at noise level. If residuals are large, the data are multidimensional. In other words, the percent of variance explained by the test or model should be higher than the percent of residuals to acclaim unidimensionality. Table 3.1a presents the IRT PCA variance explained and unexplained by the data. Table 3.1b presents component residuals and the ratio in contrast to the explained variance. Ratios of explained variance over unexplained variance are high for Reading, Mathematics, and Science. Writing has a lower ratio. Component residuals are small for all subjects and grades. These results supporting the PCA in that the IAA tests are unidimensional.

Table 3.1: Divgi Index

Grade	Reading	Mathematics	Science	Writing
3	41.48	24.05	-	18.04
4	187.91	78.83	124.88	-
5	45.66	33.25	-	70.36
6	63.95	52.03	-	117.84
7	69.57	14.79	23.16	-
8	42.54	23.68	-	14.09
11	381.71	80.95	155.57	19.70

Table 3.1a: IRT PCA Variances

Grade	Total Variance	Observed Explained Variance	Observed Unexplained Variance	% Explained Variance	% Unexplained Variance	Explained/ Unexplained
			Mathematics			
3	61	46	15	0.75	0.25	3.07
4	68.2	53.2	15	0.78	0.22	3.55
5	64.1	49.1	15	0.77	0.23	3.27
6	65	50.0	15	0.77	0.23	3.33
7	58.3	43.3	15	0.74	0.26	2.89
8	63.2	48.2	15	0.76	0.24	3.21
11	59.7	44.7	15	0.75	0.25	2.98
			Reading			
3	62.2	48.2	14	0.77	0.23	3.44
4	69.3	55.3	14	0.80	0.20	3.95
5	59.1	45.1	14	0.76	0.24	3.22
6	57	43.0	14	0.75	0.25	3.07
7	51.1	37.1	14	0.73	0.27	2.65
8	65.5	51.5	14	0.79	0.21	3.68
11	39.9	28.9	11	0.72	0.28	2.63
			Science			
4	66.8	51.8	15	0.78	0.22	3.45
7	67.1	51.1	16	0.76	0.24	3.19
11	55.4	40.4	15	0.73	0.27	2.69
			Writing			
3	25.5	18.5	7	0.73	0.27	2.64
5	21.5	14.5	7	0.67	0.33	2.07
6	24.3	17.3	7	0.71	0.29	2.47
8	18.5	11.5	7	0.62	0.38	1.64
11	25.2	18.2	7	0.72	0.28	2.60

Table 3.1b: IRT PCA Residuals

Grade	Component Residual 1	2	3	Ratio of Explained Variance/ Component Residual 1	2	3
				Mathematics		
3	1.9	1.6	1.1	24.21	28.75	41.82
4	1.5	1.3	1.3	35.47	40.92	40.92
5	1.7	1.3	1.2	28.88	37.77	40.92
6	1.5	1.3	1.2	33.33	38.46	41.67
7	2.0	1.3	1.1	21.65	33.31	39.36
8	1.8	1.3	1.1	26.78	37.08	43.82
11	1.4	1.3	1.2	31.93	34.38	37.25
				Reading		
3	1.4	1.3	1.2	34.43	37.08	40.17
4	1.5	1.3	1.2	36.87	42.54	46.08
5	1.4	1.4	1.2	32.21	32.21	37.58
6	1.4	1.3	1.2	30.71	33.08	35.83
7	1.4	1.3	1.2	26.50	28.54	30.92
8	1.5	1.3	1.3	34.33	39.62	39.62
11	1.4	1.3	1.2	20.64	22.23	24.08
				Science		
4	1.5	1.3	1.2	34.53	39.85	43.17
7	2.0	1.2	1.2	25.55	42.58	42.58
11	1.4	1.3	1.3	28.86	31.08	31.08
				Writing		
3	1.6	1.3	1.1	11.56	14.23	16.82
5	1.3	1.2	-	11.15	12.08	-
6	1.4	1.3	1.2	12.36	13.31	14.42
8	1.6	1.2	1.2	7.19	9.58	9.58
11	1.7	1.2	1.2	10.71	15.17	15.17

Internal Construct

The purpose of studying the internal structure of a test is to evaluate the extent to which test components, including subtests and items, relate to one another in theoretically or logically meaningful ways. Methods that are used to provide evidence of the internal structure of a test are usually associated with correlations. Table 3.2 reports the correlation matrices among the IAA Reading, Mathematics, Science, and Writing assessments. The correlations between Reading and Mathematics ranges from .89 in grade 11 to .92 in grade 8; the correlation between Reading and Science ranges from .89 in grade 4 to .91 in grades 11; the correlation between Reading and Writing ranges from .87 in grades 3, 5, 6, and 8 to .90 in grade 11; and the correlation between Mathematics and Science is from .91 in grades 7 and 11 to .92 in grade 4.

In addition, item-total point-biserial correlations were calculated to evaluate the test structure. The corrected point-biserial, in contrast to the uncorrected method, excludes an item from the total score when computing its point-biserial. This method avoids the overestimation issue that commonly occurs in the uncorrected method. Table 3.3 presents the median of the corrected item-total point-biserial correlations for each subject and grade. The median of the corrected item-total point-biserial correlations ranged from 0.65 to 0.79 across subjects and grades.

Table 3.2: Correlation among IAA Assessments

Grade	Test	Reading	Mathematics	Science	Writing
3	Reading	1.00	0.90	-	0.87
	Mathematics	0.90	1.00	-	0.88
	Science	-	-	-	-
	Writing	0.87	0.88	-	1.00
4	Reading	1.00	0.90	0.89	-
	Mathematics	0.90	1.00	0.92	-
	Science	0.89	0.92	1.00	-
	Writing	-	-	-	-
5	Reading	1.00	0.90	-	0.87
	Mathematics	0.90	1.00	-	0.86
	Science	-	-	-	-
	Writing	0.87	0.86	-	1.00
6	Reading	1.00	0.91	-	0.87
	Mathematics	0.91	1.00	-	0.87
	Science	-	-	-	-
	Writing	0.87	0.87		1.00
7	Reading	1.00	0.91	0.90	-
	Mathematics	0.91	1.00	0.91	-
	Science	0.90	0.91	1.00	-
	Writing	-	-	-	-
8	Reading	1.00	0.92	-	0.87
	Mathematics	0.92	1.00	-	0.89
	Science	-	-	-	-
	Writing	0.87	0.89	-	1.00
11	Reading	1.00	0.89	0.91	0.90
	Mathematics	0.89	1.00	0.91	0.89
	Science	0.91	0.91	1.00	0.92
	Writing	0.90	0.89	0.92	1.00

Table 3.3: Median of Item-Total Correlations by Subject and Grade

Grade	Reading	Mathematics	Science	Writing
3	0.71	0.68	.	0.74
4	0.65	0.72	0.71	.
5	0.70	0.71	.	0.69
6	0.70	0.73	.	0.69
7	0.65	0.65	0.70	.
8	0.69	0.75	.	0.69
11	0.79	0.75	0.77	0.76

Criterion-related Validity

In order to examine the criterion-related validity of the IAA, a study was conducted in 2009 where eight scoring monitors provided expert scores of the IAA student performance, and the relationship (i.e., *xy* in Figure 3.1) between expert scores and the teachers' scores was examined. The validation components for the performance model in Figure 3.1 provide the foundation for this study. As can be seen, the correlation between "Student Score 1" and "Expert Score" is presented as a validity coefficient "*xy*". This validation approach is based on the premise that a score given to a student performance by a trained, objective scoring monitor is a true performance score that may be used as an external criterion for estimating criterion validity, if the scoring monitor observes the same student performance as the teacher providing the score. Support for this approach is provided through existing validation research in education and industry (Suen, 1990).

For the 2009 IAA administration, eight scoring monitors were recruited by ISBE to provide secondary scores throughout the state of Illinois. All score monitors had sufficient knowledge of the IAA content, administration, and student population to be described as validation experts and met all pre-determined criteria that defined them as experts in the evaluation of the IAA testing population. The criteria used for selecting the scoring monitors were that they: (1) have more than 10 years of experience as a certified teacher; (2) are familiar with the alternative assessment population, (3) are subject matter experts regarding IAA test design and IAA rubric, and (4) represent different regional locations to get an adequate distribution across the state. The sampling plan was developed with the goal of providing an adequate number of expert scores from a representative sample of IAA students to be able to generalize results to the larger IAA population, while keeping within logistical and resource constraints for the study. With this goal in mind, ISBE solicited nominations and selected from that group eight expert scorers who best met the criteria stated above. Pearson developed a sampling frame of schools from which to solicit participation. ISBE then recruited schools from the representative, purposeful sample developed by Pearson. The sample was based on demographic diversity of students, different subject areas, and grade level diversity within school.

A training program was developed by Pearson to prepare the scoring monitors to be consistent in their approach and scoring for the expert scoring task. In preparation for the training, scoring monitors were asked to review the IAA Implementation Manual, scoring rubric, score sheet, IAA sample items, and the Online User's Guide at ISBE's IAA website. Group training for the eight scoring monitors, conducted by Pearson and ISBE via webex, included review and group discussion of the test materials, test administration, and the monitor protocol. In addition, videos of students being scored were presented to the group of monitors.

The scoring monitors provided an expert score for students' performance using the same materials and protocol as the teacher giving the first and primary score for the student assessment. Expert scores were collected during the spring 2009 IAA operational test window. Coordination of data collection activities among teachers, scoring monitors, and participating schools was a joint effort between ISBE, the scoring monitors, and Pearson. The expert scores were merged with operational test scores for students in the sample. Analyses of the merged data were conducted and results are presented below.

The sample characteristics for the validation study are presented in Table 3.4. As can be seen from the table, the sample for the spring 2009 validation study has comparable percentages of male and female students with the spring 2009 IAA student population.

Table 3.4: Spring 2009 IAA Student Population and Validation Sample Characteristics

	Spring 09 IAA Population	Validation Sample
N	13,620	194
Male	64.12%	64.00%
Female	35.88%	36.00%

Agreement between Teacher Scores and Expert Scores

Since the expert scores are used as the second scores, analysis of agreement between teacher scores and expert scores serves two purposes: inter-rater reliability and score validity. The teacher and expert's scores can be treated as two independent raters and inter-rater reliability of their scores can be computed. On the other hand, the validity evidence for open-ended item scores is commonly provided through the use of expert scores, also referred to as "Validity Papers". In such case, expert scores are considered as the "true" scores and are used to assess validity of the scores given by teachers.

In this analysis, the scores provided by the teachers were compared to those provided by the scoring monitors. The reliability/validity of scoring on various items was defined as the extent to which the items were scored exactly the same by both scorers (i.e., exact agreement) or one point of difference between the two scorers (i.e., adjacent agreement). Table 3.5 provides the mean percentage of exact agreement, the mean percentage of adjacent agreement, and the mean percentage of total agreement (i.e., the mean percentage of exact and adjacent agreement) between the two scorers. The results of these analyses suggest a high degree of agreement. The mean percentage of exact agreement between teacher scores and scoring monitor scores exceeded 93% for all subjects and grades, and the mean percentage of total agreement between teacher scores and scoring monitor scores exceeded 97% for all subjects and grades. The results of rater agreement on each item included in the Reading, Mathematics, Science, and Writing assessment are provided in Appendix F.

Table 3.5: Agreement between Teacher Scores and Expert Scores

Subject	Grade	N of Item	% of Exact Agreement	% of Adjacent Agreement	% of Total Agreement
Reading	3	18	99.60	0.40	100.00
	4	18	93.83	3.49	97.31
	5	18	98.89	0.74	99.63
	6	18	98.61	0.93	99.54
	7	18	98.81	1.19	100.00
	8	18	97.40	1.95	99.35
	11	15	99.51	0.49	100.00
Mathematics	3	19	99.19	0.40	99.60
	4	19	97.04	1.97	99.01
	5	19	97.14	2.86	100.00
	6	19	99.65	0.35	100.00
	7	19	97.74	1.13	98.87
	8	19	99.47	0.53	100.00
	11	19	97.83	1.86	99.69
Science	4	19	96.29	3.19	99.47
	7	20	98.67	1.33	100.00
	11	19	98.36	1.64	100.00
Writing	3	8	99.04	0.00	99.04
	5	8	97.50	1.25	98.75
	6	8	99.22	0.78	100.00
	8	8	100.00	0.00	100.00
	11	8	97.50	1.67	99.17

Correlations between Teacher Scores and Expert Scores

To examine evidence of criterion-related validity based on expert scores, the teachers' scores were correlated with the scoring monitors' scores. The correlations between the teacher scores and the scoring monitor scores were computed. As shown in Table 3.6, these correlations indicate a very strong positive relationship between the sets of scores by subject. The correlation results by grade for Reading, Mathematics, Science, and Writing are shown in Tables 3.6a – 3.6d respectively. Across subjects and grades, a strong positive association was found between the scores given by teachers and scoring monitors. The correlations exceeded .95 for all subjects, and approached unity for most.

Table 3.6 Correlation with Expert Scores by Subject

Subject	Sample Size	Correlation
Reading	134	0.999
Mathematics	103	0.997
Science	52	0.997
Writing	80	0.998

Table 3.6a: Correlation with Expert Scores for Reading

Grade	Sample Size	Correlation
3	14	1.000
4	23	0.987
5	18	1.000
6	14	1.000
7	19	0.998
8	18	0.999
11	28	1.000

Table 3.6b: Correlation with Expert Scores for Mathematics

Grade	Sample Size	Correlation
3	13	0.998
4	16	0.953
5	17	0.999
6	15	1.000
7	14	0.997
8	11	1.000
11	17	0.997

Table 3.6c: Correlation with Expert Scores for Science

Grade	Sample Size	Correlation
4	21	0.995
7	15	0.999
11	16	0.991

Table 3.6d: Correlation with Expert Scores for Writing

Grade	Sample Size	Correlation
3	13	0.987
5	20	0.999
6	16	0.998
8	16	1.000
11	15	0.996

The criterion-related validity evidence from the validation study is clear: the teacher scores on the IAA tests are valid. The validity coefficients based on the correlation between teachers' scores and scoring monitors' scores range from 0.70 to 0.99 by subject. Overall, the validity results based on content-, construct-, and criterion-related evidence suggest that the IAA provides valid assessment of the performance of students in the 1% population.

4. CALIBRATION AND SCALING

The purpose of item calibration and equating is to create a common scale so items developed in different years can be used interchangeably, and student performances can be evaluated across years. The latter is an important aspect for assessing annual progress (AYP) that is mandated by the NCLB Act. Calibration and equating produces item parameter and theta estimates. Theta, the student latent ability, usually ranges from -4 to 4; thus, it is not appropriate for reporting purposes. Therefore, following calibration and equating, the scale is usually transformed to a reporting scale (e.g. scale score) that is easer to interpret and memorize by students, teachers, and other stakeholders.

Calibration

For the calibration of the IAA, the Rasch partial credit model (RPCM) was used because of its flexibility in accommodating a smaller n-count and for its ability to handle polytomous data. The IAA scoring is a one-to-one relationship between theta, raw score (total number of item answer correctly), and scale scores. The RPCM is defined via the following mathematical measurement model where, for a given item involving *m* score categories, the probability of student *j* scoring *x* on item *i*, P_{ijx}, is given by:

$$P_{ijx} = \frac{\exp \sum_{k=0}^{x} (B_j - D_{ik})}{\sum_{h=0}^{m_i} \exp \sum_{k=0}^{h} (B_j - D_{ik})}, \quad x = 0, 1, 2, ..., m_i, \text{ where} \qquad (4.1)$$

$$\sum_{k=0}^{0} (B_j - D_{ik}) \equiv 0 \text{ and } \sum_{k=0}^{h} (B_j - D_{ik}) \equiv \sum_{k=1}^{h} (B_j - D_{ik}). \qquad (4.2)$$

The RPCM has two parameters: the student ability B_j and the step difficulty (D_{ik}). The step difficulty (D_{ik}) is the threshold difficulty that separates students of adjacent scores. All RPCM analyses for the IAA are conducted using the commercially available program WINSTEPS 3.60 (Linacre, 2006).

Scaling

The IAA Reading, Mathematics, Science, and Writing scores are each reported on a continuous score scale that ranges from 300 to 700. The scales are grade-level scale. In other words, scale scores are comparable across years of the same subject and grade, but are not comparable across grades or subjects.

Spring 2008 was the first operational administration of the IAA Mathematics, Reading, Science, and grade 6 Writing tests, while grades 5, 8, and 11 Writing tests were administered first in 2007. As such the base IRT scale was set for grades 5, 8, and 11 Writing in 2007 and all the other tests in 2008. In 2009, however, the IAA test length was increased significantly (see Table 5.2 in Chapter 5 for details) so as to increase content coverage and improve the reliability and validity of the test scores. The increase in test length resulted in more raw score points than the original scale score range of 30-70. Therefore, ISBE decided to set a new IAA scale score range of 300-700, and anchor the Satisfactory cut score at 500. Additionally, the distance between the Mastery scale score cut and Satisfactory scale score cut from 2008 should be maintained relative to the 2009 scale. The new scale transformation constants were then computed for each subject and grade based on these guidelines. Given the change of the scale, the IAA was re-baselined, and 2009 becomes the new base year of all subjects and grades for future administrations.

Due to the increased test length and the standardized administration instructions, a standards validation meeting was held in May 2009, and cut scores for different performance levels were set on the raw score scale. Following the standards validation meeting, the theta value corresponding to the raw score cuts were obtained to compute the scale transformation constants. The equations for computing the slope (*M1*) and intercept (*M2*) of scale transformation are presented in Equations 4.3 and 4.4.

$$M1 = \frac{SSCut_{Mastery08} - SSCut_{Satisfactory08}}{ThetaCut_{Mastery09} - ThetaCut_{Satisfactory09}} \times 10 \tag{4.3}$$

$$M2 = 500_{SSCut\_Satisfactory} - (ThetaCut_{Satisfactory09} \times M1) \tag{4.4}$$

M1 is calculated by first dividing the distance between the 2008 Mastery scale score cut and Satisfactory scale score cut by the distance between theta values associated with the Mastery cut and the Satisfactory cut in 2009. Then this value is multiplied by 10 to reflect the scale change from 30-70 to 300-700. *M2* is calculated by computing the difference between the scale score associated with the Satisfactory cut (500) and the theta associated with this cut in 2009 multiplied by *M1*.

Being the first year of administration, Writing grade 3 doesn't have existing scale score cuts that can be used to calculate the *M1* in equation 1. Therefore, two approaches were investigated: using median scale score cuts of other Writing grades or using Writing grade 5 scale score cuts as base. The latter approach resulted in grade 3 scale score cuts that were more in line with other Writing grades. Thus, the grade 5 scale score cuts from 2008 administration were adopted to calculate Writing grade 3 scale transformation constants.

After scale transformation constants are derived, the scale score (*SS*) and standard error of estimate (*SE*) are computed using the following equations.

$$SS = Theta \times M1 + M2 \tag{4.5}$$

$$SE = Theta \times M1 \qquad (4.6)$$

The raw-score-to-scale-score conversion tables can be found in Appendix B along with the conditional SEM associated with each scale score point.

5. STANDARDS VALIDATION

On May 4th and 5th of 2009, Pearson, under the contract to the Illinois State Board of Education (ISBE), held a standard validation meeting. The purpose of the meeting, as stated to the panelists, was to validate the performance level cut scores on the Illinois Alternate Assessment (IAA) Mathematics tests at grades 3-8 and 11, Reading tests at grades 3-8 and 11, Science tests at grades 4, 7, and 11, and Writing tests at grades 3, 5, 6, 8, and 11.

The cut scores for Writing grades 5, 8, and 11 were established in 2007. Cut scores for all grades in Mathematics, Reading, and Science, along with Writing for grade 6, were established in 2008. In 2009, the IAA test was modified in two respects: (1) the number of items was increased to expand content coverage, and (2) the instructions for test administrators became more specific and prescriptive, with the scoring rubric scripted into the administration instructions. Additionally, the Writing grade 3 test was first administered in 2009. With these modifications, the Illinois State Board of Education (ISBE) recognized the need to reevaluate the cut scores prior to releasing scores for 2009.

The ultimate goal was to provide recommendations to the ISBE on the appropriateness of the cut scores for the Foundational, Satisfactory, and Mastery performance levels on IAA Mathematics, Reading, Science, and Writing tests. The Reasoned Judgment procedure was used for the standards validation. The outcomes of the meeting are described in this chapter.

Panelists

A total of 71 educators participated for a day and a half to determine the appropriateness of the cut scores on 2009 IAA tests. With a joint effort between ISBE and Pearson, the panelists were recruited to be representative of IAA subject matter experts across the content areas. The panelists met in six committees: lower Mathematics (grades 3-5), upper Mathematics (grades 6-8, 11), lower Reading (grades 3-5), upper Reading (grades 6-8, 11), Science (grades 4, 7, 11), and Writing (grades 3, 5, 6, 8, 11). A summary of panelist demographic information is provided in Table 1.

Table 5.1: A summary of Panelist Demographic Information by Committee

Subject	Grade/ Panel	# of Panelists	Special Education Specialty	Male	Female	Ethnic Minority*	Average # of Years Teaching
Mathematics	3, 4, 5	11	9	1	10	0	19
	6, 7, 8, 11	12	9	0	12	2	21
Reading	3, 4, 5	12	9	1	11	1	19
	6 ,7, 8, 11	12	9	2	10	1	14
Science	4, 7, 11	12	10	1	11	2	12
Writing	3, 5, 6, 8, 11	12	9	1	11	3	20

* Some of the panelists did not respond to this question; of these, there was one in upper Mathematics, one in lower Reading, two in Science, and one in Writing.

The IAA Standards Validation Process

To prepare for the standards validation, Pearson linked 2009 and 2008 tests by matching the student score distribution between the years. The resulting raw scores that produced a similar percent of students in each performance level were located and identified as the linked cut scores.

Pearson proposed standards validation through the Reasoned Judgment method to determine the appropriateness of linked cut scores on the 2009 version of the IAA. Reasoned Judgment is one of the popular standard setting/validation procedures used in the alternate assessment context (Roeber, 2002; Perie, 2007). The application of this procedure is similar to the Body of Work component used in the 2008 IAA standard setting meeting. The idea is that the panelists review the range of raw scores made up of a combination of scores, and make judgments about how different combinations fall into each performance level. For example, a student who received a score combination of all 1's on the items would almost certainly be classified as the Entry level. Likewise, a student who received a score combination of 4's on all items would almost certainly be classified as Mastery.

The IAA standards validation went through a similar procedure as other standard setting methods to ensure the validity of the standard setting (Hambleton, 1998). First, the panelists reviewed Illinois Learning Standards, Illinois Assessment Frameworks, and item content maps. Second, they were given adequate time to review the current performance level descriptors so they could fully understand the descriptors and, thus, could evaluate the reasonableness of the linked cut scores in the context of the changes made to the test. Special attention was paid to the threshold students who are barely at the Foundational, Satisfactory, and Mastery performance levels. Key characteristics for these threshold students were identified by the panels for each performance level. These key characteristics were used when the panelists later evaluated the items and determined what score point the threshold students should obtain at a given performance level. Then the panelists were provided materials to familiarize them with the tests, instructions for test administrators, and the scoring rubric.

After the panelists understood the performance level descriptors and were familiar with the testing materials, they worked through the items to estimate what score point threshold students at a given performance level should be able to earn on each item. Once score patterns (counts of 4's, 3's, 2's, and 1's) were developed in this way, the panelist computed their cut scores, and considered other score patterns that could lead to the same cut scores. Next, empirical score patterns were presented to the panelists along with the linked cut scores. Only score patterns observed with at least three students were included. The panelists were asked to evaluate whether the linked cut scores reflected the desired expectations for the threshold students at a given level. If the answer was "Yes", they should keep the linked cut score. If the answer was "No", then they needed to share their content-based rationale for changing the linked cut scores in Round 2. One important aspect of the standards validation procedure was for panelists to understand that their decisions should be based on content expectations only.

In Round 2, Round 1 recommended cut score distributions were presented, which showed the cut score of each panelist along with the median of the raw scores chosen by the group. Discussions were around the range of raw score corresponding to the lowest and highest ratings for each cut score. Next, score patterns around each median cut score were discussed, and content-based rationales were shared for those who changed the linked cut scores. Then the impact data were shown and discussed for the median cut scores and linked cut scores. The impact data indicated the percentage of students in each performance level, if these cut scores were implemented. Following group discussions, the panelists were asked to make their Round 2 recommendations and decide whether to accept the linked cut or recommend a modified cut for each performance level. If any of their Round 2 recommended cut scores were different from the linked cut scores, they were requested to document a content-based rationale on the rationale sheet.

A summary of activities for Round 1 and Round 2 are presented below:

<u>Round 1</u>

- Start with the Entry/Foundational cut score
- Review performance level descriptors if necessary
- Review each item and assign a score point threshold Foundational students should earn
- Compute a raw score cut based on the score pattern
- Consider other patterns leading to that score
- Repeat for Satisfactory and Mastery cut scores
- Repeat for all grades
- Receive score patterns from student data (when at least 3 students obtained the score pattern) and linked cut scores
- Compare the computed cut scores to linked cut scores and patterns and evaluate differences
- Decide if a change is recommended and, if so, provide a content-based rationale.

Round 2

- Present agreement feedback on panelists' ratings
- Discuss patterns around each median cut score
- Discuss rationales from Round 1
- Show impact data for both the median cut scores and the linked cut scores. Discuss whether or not the Round 1 impact data look reasonable
- Make final judgments and document content-based rationales if changes to the linked cut scores are recommended.

Following Round 2, the medians of the cut scores recommended by the panelists were calculated for each performance level by subject and grade. The medians were taken as the recommended cut scores of the standards validation meeting.

Panelist Readiness and Evaluation forms (see Appendix G for an example) were used to collect information regarding the panelists' understanding of their tasks and comfort level with regard to the cut scores recommended in the meeting. Analysis of these forms shows agreement that the participants understood the task, understood the impact data presented, and were prepared to validate the cut scores at each round.

Linked Cut Scores

The linked cut scores were used to facilitate the standards validation due to changes made to the IAA tests in 2009. First, compared with the 2008 tests, the 2009 tests are considerably longer. As shown in Table 2, their length increased from 22% to 167%, with all but one grades at or above 40%. Second, 2009 instructions for administrators are more clear and standardized. The rubric was also scripted into instructions for administrators, which made it much easier for the teachers to use. Other changes include standardized passages for reading, rather than having teachers picking their own passages to test the students as in 2008.

Despite these modifications to the 2009 tests, the learning standards, assessment frameworks, scoring rubric, and the performance level descriptors stayed the same. With the understanding that the achievement of IAA takers is similar across years, the percent of students falling into each performance level should be similar across years as well. As a result, the linked cut scores were derived using a method similar to the equipercentile procedure. This way, the linked cut scores led to similar percent of students falling in each performance level when compared with 2008.

Table 5.2: Comparison of 2008 and 2009 IAA Test Length

Subject	Grade	2008	2009	Percent Increase
Mathematics	3-8, 11	10	15	50 %
Reading	3-8	9	14	56 %
	11	9	11	22 %
Science	4, 11	6	15	150 %
	7	6	16	167 %
Writing	3, 5, 6, 8, 11	5	7	40 %

Grade 3 Writing

As mentioned earlier, the grade 3 Writing test did not have linked cut scores because 2009 was the first year of administration. Considering that the grade 3 Writing test was designed to have expectations for content and difficulty that were parallel to the other Writing grade levels, a standards validation approach was also used. To accomplish this, cut scores on the grade 3 test were estimated by extrapolating the expectations from the remaining Writing tests. The procedure is conceptually similar to the interpolation procedure used in 2008 IAA standard setting. The extrapolation was established through a method similar to the equipercentile procedure. First, the cumulative percentage distribution (CPF) at or above each Writing performance cut of grades 5, 6, 8, and 11 was identified. Next, the median of these CPFs was calculated for each cut score. Last, the cut score for each performance level of grade 3 Writing was established by locating the raw score corresponding to a CPF that is the closest to the median CPF identified earlier.

Recommended Raw Cut Scores

The medians of Round 2 cut scores recommended by the panels are presented by subject and grade in Table 5.3.

Table 5.3: Round 2 Raw Cut Scores by Subject

	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Grade 11
Mathematics							
Foundational	42	39	38	37	39	38	39
Satisfactory	50	50	51	49	48	51	49
Mastery	57	58	59	57	57	58	58
Reading							
Foundational	37	37	39	**30**	38	40	28
Satisfactory	49	47	47	**44**	**46**	47	40
Mastery	55	54	52	54	55	54	44
Science							
Foundational	-	35	-	-	39	-	36
Satisfactory	-	48	-	-	53	-	49
Mastery	-	55	-	-	60	-	57
Writing							
Foundational	14	-	14	16	-	**18**	17
Satisfactory	22	-	21	23	-	24	23
Mastery	26	-	27	**27**	-	27	27

Note. The bolded number represents the Round 2 recommended cut scores that are different from the linked cut score.

Rationale for Changing the Linked Cut Scores

As a result of the standards validation, the panels adopted the linked cut scores across all the subjects and grades except for Reading grades 6 and 7, and Writing grades 6 and 8. For Reading grade 6, the median cut for the Foundational and Satisfactory levels (raw scores 30 and 44) are 4 points lower than the linked cuts (raw 34 and 48) respectively. For Writing, the grade 6 cut for the Mastery level is one point lower than the corresponding linked cut, while the grade 8 cut score for the Foundational level is 2 points lower than the linked cut.

A review of the rationales for Reading grades 6 and 7 revealed that the panelists generally thought the tests were harder than in 2008, and some felt that these tests were harder than the higher grade-level tests. A common theme of the sources of difficulty listed by the panelists included a) difficult vocabulary (e.g., trousers, hibernate, genres, curious, science fiction, discouraged); and b) long passages (e.g., ice skating).

For Writing grade 6, the panelists chose 27 rather than 28 (the perfect score) to be the cut score for the mastery level. They believed that the cut of 27 would allow for the transitional mastery student to have room for one error and allow the mastery level more than just one score. For Writing grade 8, 18 was recommended as the cut score for the Foundational level instead of the linked cut of 20. The typical rationale for this change was that the linked cut score for grade 8 was much higher than the pattern for other grades. In addition, the material and questions in the grade 8 test were considered to be at a much higher difficulty level. The facilitator noted that the committee discussed the need to be consistent in expectations, with regard to score patterns, across grades as the chief rationale for the changes recommended in writing. The two changes align expectations for the Foundational and Mastery levels. The linked cut scores for Satisfactory were judged to be aligned.

Approved Cut Scores

The Round 2 raw cut scores recommended by the panels were presented to ISBE along with the panelists' rationales for changing the linked cut scores. The final raw score cuts approved by ISBE and the Illinois State Testing Review Committee are presented in Table 5.3. Note that the Round 2 reading cut scores that are different from the linked cuts were adjusted, while the writing cut scores resulting from Round 2 were kept the same. Tables 5.4a to 5.4d provide the corresponding scale score range for each performance level.

Table 5.3: IAA Raw Score Cuts by Subject

	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Grade 11
Mathematics							
Foundational	42	39	38	37	39	38	39
Satisfactory	50	50	51	49	48	51	49
Mastery	57	58	59	57	57	58	58
Reading							
Foundational	37	37	39	33	38	40	28
Satisfactory	49	47	47	47	48	47	40
Mastery	55	54	52	54	55	54	44
Science							
Foundational		35			39		36
Satisfactory		48			53		49
Mastery		55			60		57
Writing							
Foundational	14		14	16		18	17
Satisfactory	22		21	23		24	23
Mastery	26		27	27		27	27

Table 5.4a: IAA Reading Scale Score Range for Each Performance Level

	Grade 3		Grade 4		Grade 5		Grade 6		Grade 7		Grade 8		Grade 11	
Performance Level	Low	High	Low	High	Low	High	Low	High	Low	High	Low	High	Low	High
Entry	300	473	300	476	300	466	300	462	300	478	300	476	300	467
Foundational	474	499	477	499	467	499	463	499	479	499	477	499	468	499
Satisfactory	500	544	500	537	500	536	500	545	500	546	500	552	500	557
Mastery	545	576	538	589	537	661	546	608	547	576	553	624	558	558

Table 5.4b: IAA Mathematics Scale Score Range for Each Performance Level

	Grade 3		Grade 4		Grade 5		Grade 6		Grade 7		Grade 8		Grade 11	
Performance Level	Low	High	Low	High	Low	High	Low	High	Low	High	Low	High	Low	High
Entry	300	470	300	469	300	480	300	455	300	480	300	461	300	477
Foundational	471	499	470	499	481	499	456	499	481	499	462	499	478	499
Satisfactory	500	550	500	553	500	540	500	563	500	538	500	554	500	547
Mastery	551	657	554	622	541	566	564	679	539	607	555	628	548	603

Table 5.4c: IAA Science Scale Score Range for Each Performance Level

	Grade 4		Grade 7		Grade 11	
Performance Level	Low	High	Low	High	Low	High
Entry	300	435	300	432	300	468
Foundational	436	499	433	499	469	499
Satisfactory	500	559	500	565	500	542
Mastery	560	700	566	700	543	624

Table 5.4d: IAA Writing Scale Score Range for Each Performance Level

	Grade 3		Grade 5		Grade 6		Grade 8		Grade 11	
Performance Level	Low	High	Low	High	Low	High	Low	High	Low	High
Entry	300	436	300	465	300	442	300	449	300	458
Foundational	437	499	466	499	443	499	450	499	459	499
Satisfactory	500	558	500	558	500	578	500	565	500	576
Mastery	559	660	559	595	579	636	566	625	577	635

Estimated Consequences of the Final Cut Scores

Based on the approved cut scores, IAA students were classified into four performance levels: Entry, Foundational, Satisfactory, and Mastery. The results are presented in Tables 5.5a to 5.5d. Note that the sum of percentages by subject and grade may not add up to 100% due to rounding.

Table 5.5a: Percentages of Students in each Performance Level for Reading

	Reading			
Grade	Entry	Foundational	Satisfactory	Mastery
3	20	24	33	24
4	21	20	35	24
5	24	18	23	36
6	14	18	36	32
7	15	20	42	24
8	18	14	37	32
11	13	18	31	38

Table 5.5b: Percentages of Students in each Performance Level for Mathematics

	Mathematics			
Grade	Entry	Foundational	Satisfactory	Mastery
3	22	17	35	25
4	17	18	35	30
5	16	20	41	23
6	14	15	33	38
7	16	15	41	29
8	12	19	37	31
11	16	14	43	26

Table 5.5c: Percentages of Students in each Performance Level for Science

	Science			
Grade	Entry	Foundational	Satisfactory	Mastery
4	15	18	26	41
7	11	17	29	43
11	12	13	28	47

Table 5.5d: Percentages of Students in each Performance Level for Writing

	Writing			
Grade	Entry	Foundational	Satisfactory	Mastery
3	13	16	30	41
5	11	17	43	29
6	13	20	36	31
8	13	18	30	40
11	12	11	26	50

Panelist Variability

Estimation of panelist variability can be used to evaluate the stability of the cut score recommendations, considering that the standards validation might be replicated using a different collection of panelists. In order to estimate and describe the variability in panelist's judgments, a Generalizability Theory (G-Theory) study was conducted (Lee & Lewis, 2001). For this investigation, the sources of variability of interest were panelists and rounds. For each performance level, the variance associated with each of these sources was estimated using the maximum likelihood SAS VARCOMP procedure. After estimation of the variance components, G-Theory provides a mechanism for describing the variability associated with the panelists' judgments. This is important for determining how similar the cut scores might be if a different set of panelists were asked to validate the cut scores. The result is an estimate of the standard error of the cuts cores for this set of data.

For this study, the number of rounds was treated as a fixed factor, meaning that if the meeting were held again, the same number of rounds would be used. The two rounds of cut scores were used.

The G-Theory standard error was computed using the formula below, and the standard error estimates were adjusted by 1.253 to account for the use of the median.

$$SE_{cut} = \sqrt{\frac{\sigma^2_{Judges}}{N_{Judges}} + \frac{\sigma^2_{Error}}{2 \bullet N_{Judges}}} \,. \tag{5.1}$$

The conditional standard error of measurement (CSEM) for each recommended raw score cut was interpolated based on the Rasch conditional standard error of measurement. It is common for policy-makers to consider the total error associated with cut scores prior to making final decisions. Total error in this case is conceptualized as the sum of the measurement error associated with the instrument and the error associated with the cut score procedures described above. The total error was calculated as follows:

$$SE_{Total} = \sqrt{(CSEM_{Cut})^2 + (SE_{Cut})^2} \,, \tag{5.2}$$

where *CSEM* is the conditional standard error of measurement for the raw score cut, and *SE* is the standard error computed using G-theory. Tables 5.6a, 5.6b, 5.6c, and 5.6d provide the standard errors computed via the two procedures and the total error values for Mathematics, Reading, Science, and Writing respectively.

Table 5.6a: Standard Error Indices for Reading

Grade	Cut	$CSEM_{cut}$	SE_{cut}	SE_{total}
3	Foundational	4.17	1.11	4.32
	Satisfactory	3.61	1.16	3.79
	Mastery	2.15	0.75	2.28
4	Foundational	3.85	1.30	4.06
	Satisfactory	3.54	1.02	3.68
	Mastery	2.37	0.52	2.43
5	Foundational	3.95	1.33	4.17
	Satisfactory	3.64	1.09	3.80
	Mastery	2.89	0.41	2.92
6	Foundational	3.47	1.19	3.67
	Satisfactory	3.80	1.32	4.02
	Mastery	2.32	0.78	2.45
7	Foundational	3.87	1.19	4.05
	Satisfactory	3.62	1.04	3.77
	Mastery	1.91	0.69	2.03
8	Foundational	3.87	1.64	4.20
	Satisfactory	3.51	0.80	3.60
	Mastery	2.26	0.42	2.30
11	Foundational	3.57	0.70	3.64
	Satisfactory	2.97	0.65	3.04
	Mastery	2.25	0.39	2.28

Table 5.6b: Standard Error Indices for Mathematics

Grade	Cut	$CSEM_{cut}$	SE_{cut}	SE_{total}
3	Foundational	4.26	0.99	4.37
	Satisfactory	3.88	0.11	3.88
	Mastery	2.73	0.00	2.73
4	Foundational	4.15	0.54	4.19
	Satisfactory	3.81	0.11	3.81
	Mastery	2.41	0.06	2.41
5	Foundational	4.07	0.51	4.10
	Satisfactory	3.74	0.38	3.76
	Mastery	2.94	0.17	2.94
6	Foundational	3.95	0.89	4.05
	Satisfactory	3.78	0.73	3.85
	Mastery	2.50	0.46	2.54
7	Foundational	4.01	0.92	4.11
	Satisfactory	3.87	0.73	3.94
	Mastery	2.57	0.73	2.67
8	Foundational	4.04	0.60	4.08
	Satisfactory	3.64	0.71	3.71
	Mastery	2.26	0.34	2.28
11	Foundational	3.99	0.88	4.09
	Satisfactory	3.72	0.72	3.79
	Mastery	2.19	0.43	2.23

Table 5.6c: Standard Error Indices for Science

Grade	Cut	$CSEM_{cut}$	SE_{cut}	SE_{total}
4	Foundational	3.87	0.65	3.92
	Satisfactory	3.81	0.69	3.87
	Mastery	2.94	0.32	2.96
7	Foundational	3.96	0.87	4.05
	Satisfactory	3.73	0.97	3.85
	Mastery	2.76	0.58	2.82
11	Foundational	4.04	0.72	4.10
	Satisfactory	3.85	0.58	3.89
	Mastery	2.54	0.42	2.57

Table 5.6d: Standard Error Indices for Writing

Grade	Cut	$CSEM_{cut}$	SE_{cut}	SE_{total}
3	Foundational	2.26	0.16	2.27
	Satisfactory	2.76	0.16	2.76
	Mastery	2.16	0.10	2.16
5	Foundational	2.08	0.14	2.08
	Satisfactory	2.65	0.31	2.67
	Mastery	1.88	0.63	1.98
6	Foundational	2.40	0.25	2.41
	Satisfactory	2.48	0.18	2.49
	Mastery	1.79	0.21	1.80
8	Foundational	2.65	0.84	2.78
	Satisfactory	2.56	0.62	2.63
	Mastery	1.85	0.69	1.98
11	Foundational	2.68	0.26	2.69
	Satisfactory	2.68	0.05	2.68
	Mastery	1.71	0.16	1.72

REFERENCES

Abedi, J. (1997). *Dimensionality of NAEP Subscale Scores in Mathematics*. CSE Technical Report 428. http://www.cse.ucla.edu/CRESST/pages/reports.htm.

Clark, L. A. & Watson, D. (1995). Constructing validity: basic issues in objective scale development. *Psychological Assessment*, *7*(3), 309-319.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Earlbaum Associates, Inc. NJ

Crocker, L. M. & Algina, J. (1986). *Introduction to Classical & Modern Test Theory*. Orlando, FL: Pearson Brace Jovanovich, Inc.

Cronbach, LJ, & Meehl, PE (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281-302.

Dawis, R. (1987). A theory of work adjustment. In B. Bolton (Ed.). *Handbook on the measurement and evaluation in rehabilitation* (2nd ed.) (pp. 207-217). Baltimore: Paul H. Brooks.

Divgi, D. R. (1980). *Dimensionality of Binary Items: Use of a Mixed Model*. Paper presented at the annual meeting of the National Council on Measurement in Education, Boston MA.

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement (3rd Edition)* (pp. 105-146). New York: Macmillan.

Glaser, R. (1963). Instructional technology and the measurement of learning outcomes: Some questions. *American Psychologist, 18,* 519-521.

Hambleton, R. (1998). Setting Performance Standards on Achievement Tests: Meeting the Requirements of Title I. *Handbook for the Development of Performance Standards*. Washington, DC: U.S. Department of Education and the Council of Chief State School Officers.

Hattie, J. (1985). Methodology review: assessing unidimensionality of tests and items. *Applied Psychological Measurement, 9*(2), 139-164.

Individuals with Disabilities Education Act (1990). 20 U.S.C. § 1400 et seq. (1990) (amended 1997, 2004)

Lee, G. & Lewis, D. M. (2001). *A generalizability theory approach toward estimating standard errors of cutscores set using the bookmark standard setting procedure*. Paper presented at the annual meeting of the national council on measurement in education, Seattle, WA.

Linacre, J. M. (2006). *WINSTEPS: Rasch measurement*, Version 3.61 [Computer Software]. Chicago, IL: WINSTEPS.

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

Messick, S (1989). Validity. In R. L. Linn(Ed.), *Educational Measurement* (3rd ed., pp. 13-103). New York: Macmillan.

Messick, S. (1995). Validity of psychological assessment. *American Psychologist, 50*(9),741-749.

Naylor, J. C., & Ilgen, D. R. (1984). Goal setting: A theoretical analysis of a motivational technology. *Research in Organizational Behavior, 6*, 95-141.

No Child Left Behind Act (2001). 20 U.S.C. § 6301 et seq (PL 107-110).

Perie, M. (2007). *Setting alternate achievement standards*. National Center for the Improvement of Educational Assessment. Dover, NH: NCIEA.

Raîche, G. (2005). Critical Eigenvalue sizes in standardized residual principal components analysis. *Rasch Measurement Transactions*, 19:1, 1012

Roeber, E. (2002). *Setting standards on alternate assessments* (NCEO Synthesis Report 42). Minneapolis: University of Minnesota, National Center on Educational Outcomes.

Rosenthal, R. & Rosnow, R. L. (1991). *Essentials of behavioral research: Methods and data analysis* (2nd ed.). New York: McGraw Hill.

Rudner, L. M. (2005). Expected classification accuracy. *Practical Assessment, Research & Evaluation 10*(13). Available online http://pareonline.net/pdf/v10n13.pdf

Suen, H. K. (1990). *Principles of test theories*. Lawrence Erlbaum Associates, Inc. NJ

Tabachnick, B. G. & Fidell, L. S. (2007). Using Multivariate Statistics (fifth ed.). Pearson Education, Inc.

Tennant, A. & Pallant, J.F. (2006) Unidimensionality Matters! (A Tale of Two Smiths?). *Rasch Measurement Transactions*, 20:1 p. 1048-51

Thissen, D. & Wainer, H. (2001). *Test Scoring*. Lawrence Erlbaum Associates, Inc. NJ

U.S. Department of Education. (2005). *Alternate Achievement Standards for Students with the most Significant Cognitive Disabilities: Nonregulatory Guidance.* Available online http://www.ed.gov/policy/elsec/guid/altguidance.doc

Wright, B. D. (1996) Local dependency, correlations and principal components. *Rasch Measurement Transactions*, 10:3, 509-511

APPENDIX A: IAA Scoring Rubric

IAA Performance-Based Task Scoring Rubric

Level 4:	***Level 3:***	***Level 2:***	***Level 1:***
The student correctly performs the task without assistance or with a single repetition of instructions or refocusing.	**The student correctly performs the task with general prompts.**	**The student correctly performs the task with specific prompts.**	**The student does not perform the task at Level 2 or provides an incorrect response despite Level 2 support.**
• The student responds correctly to the task when presented as it is written in the instructions with the necessary materials. • If the student does not respond independently or responds incorrectly to the initial presentation of the task when given adequate wait time, the teacher repeats the instructions and/or refocuses the student's attention.	• If the student responds incorrectly to the task at Level 4 when given adequate wait time, the teacher provides additional information or adds prompts about the expected response from the student such as: o Elaborating or providing additional clarifying information on directions or expected response. o Demonstrating a like response such as, "This is a picture of a dog. Show me a picture of a cat." o Providing examples but not modeling the correct response.	• If the student responds incorrectly to the task at Level 3 when given adequate wait time, the teacher provides specific prompts to direct the student's correct response such as: o Modeling exact response, "This is a picture of a dog, what is this?" (Show a picture of a dog). o After physically guiding the student to the correct response such as using hand over hand, the student then indicates the correct answer in his/her mode of communication.	
The student then responds correctly.	***The student then responds correctly.***	***The student responds correctly after being given the correct answer.***	***The student does not respond or does not respond correctly. Teacher demonstrates response and moves on to the next task.***

Illinois State Board of Education has adapted this rubric from the Colorado Student Assessment Program Alternate Level of Independence Performance Rubric. ISBE August 31, 2006

APPENDIX B: Conditional Standard Errors of Measurement Associated With IAA Scale Scores

Reading

Raw Score	Grade 3		Grade 4		Grade 5		Grade 6		Grade 7		Grade 8		Grade 11	
	Scale Score	SE	Scale Score	SE	Scale Score	SE	Scale Score	SE	Scale Score	SE	Scale Score	SE	Scale Score	SE
1	300	53	300	54	300	97	300	64	300	46	300	72	300	47
2	300	53	300	54	300	97	300	64	300	46	300	72	300	47
3	300	53	300	54	300	97	300	64	300	46	300	72	300	47
4	300	53	300	54	300	97	300	64	300	46	300	72	300	47
5	300	53	300	54	300	97	300	64	300	46	300	72	300	47
6	300	53	300	54	300	97	300	64	300	46	300	72	300	47
7	300	53	300	54	300	97	300	64	300	46	300	72	300	47
8	300	53	300	54	300	97	300	64	300	46	300	72	300	47
9	300	53	300	54	300	97	300	64	300	46	300	72	300	47
10	300	53	300	54	300	97	300	64	300	46	300	72	300	47
11	300	53	300	54	300	97	300	64	300	46	300	72	367	47
12	300	53	300	54	300	97	300	64	300	46	300	72	397	26
13	300	53	300	54	300	97	300	64	300	46	300	72	415	18
14	357	53	345	54	300	97	315	64	364	46	300	72	425	14
15	391	29	381	30	300	53	358	35	394	25	342	39	432	12
16	410	20	401	21	325	38	382	25	412	18	369	28	437	11
17	421	16	413	17	347	31	397	20	422	15	385	22	441	10
18	428	13	422	15	362	26	407	18	430	13	396	19	445	9
19	434	12	428	13	374	23	415	16	435	11	405	17	448	9
20	438	11	433	12	383	21	421	14	440	10	411	16	451	8
21	442	10	438	11	391	20	426	13	444	9	417	14	453	8
22	445	9	441	10	398	18	431	12	447	9	422	14	456	8
23	448	9	445	10	404	17	435	12	450	8	426	13	458	8
24	450	8	448	9	409	17	439	11	453	8	430	12	460	7
25	452	8	451	9	415	16	442	11	455	8	434	12	462	7
26	455	8	453	9	419	15	445	10	457	7	438	11	464	7
27	457	8	456	8	424	15	448	10	459	7	441	11	466	7
28	459	7	458	8	428	15	451	10	461	7	444	11	468	7
29	461	7	460	8	432	14	454	10	463	7	447	11	470	7
30	462	7	462	8	436	14	456	9	465	7	450	11	472	7
31	464	7	464	8	439	14	459	9	467	7	452	10	474	7
32	466	7	466	8	443	14	461	9	469	7	455	10	476	7
33	468	7	469	8	446	14	463	9	470	7	458	10	478	8
34	469	7	471	8	450	14	466	9	472	7	460	10	480	8
35	471	7	473	8	453	13	468	9	474	7	463	10	483	8
36	473	7	475	8	457	13	470	9	476	7	466	10	486	8
37	474	7	477	8	460	13	473	9	477	7	468	10	488	9
38	476	7	479	8	464	14	475	9	479	7	471	10	492	9
39	478	7	481	8	467	14	478	9	481	7	474	10	495	10
40	479	7	483	8	471	14	480	9	482	7	477	11	500	12
41	481	7	485	8	474	14	482	9	484	7	480	11	506	13
42	483	7	487	8	478	14	485	10	486	7	482	11	514	17
43	485	8	489	8	482	15	488	10	488	7	486	11	530	24
44	487	8	492	9	486	15	490	10	490	7	489	11	558	46
45	489	8	494	9	490	15	493	10	492	8	492	12		
46	492	8	497	9	495	16	497	11	495	8	496	12		
47	494	9	500	9	500	17	500	11	497	8	500	13		
48	497	9	503	10	505	17	504	12	500	9	504	13		
49	500	10	507	10	512	19	508	12	503	9	509	14		
50	503	10	511	11	518	20	513	13	507	10	515	15		

Raw Score	Grade 3		Grade 4		Grade 5		Grade 6		Grade 7		Grade 8		Grade 11	
	Scale Score	SE	Scale Score	SE	Scale Score	SE	Scale Score	SE	Scale Score	SE	Scale Score	SE	Scale Score	SE
51	508	11	515	12	527	22	518	15	511	11	521	17		
52	512	13	521	14	537	24	525	16	516	12	529	19		
53	519	15	528	16	549	28	534	19	522	14	539	21		
54	528	18	538	19	568	35	546	23	531	17	553	27		
55	545	27	556	28	601	51	568	34	547	24	579	38		
56	576	52	589	53	661	95	608	63	576	45	624	71		

Mathematics

Raw Score	Grade 3		Grade 4		Grade 5		Grade 6		Grade 7		Grade 8		Grade 11	
	Scale Score	SE	Scale Score	SE	Scale Score	SE	Scale Score	SE	Scale Score	SE	Scale Score	SE	Scale Score	SE
1	300	95	300	73	300	39	300	98	300	57	300	74	300	55
2	300	95	300	73	300	39	300	98	300	57	300	74	300	55
3	300	95	300	73	300	39	300	98	300	57	300	74	300	55
4	300	95	300	73	300	39	300	98	300	57	300	74	300	55
5	300	95	300	73	300	39	300	98	300	57	300	74	300	55
6	300	95	300	73	300	39	300	98	300	57	300	74	300	55
7	300	95	300	73	300	39	300	98	300	57	300	74	300	55
8	300	95	300	73	300	39	300	98	300	57	300	74	300	55
9	300	95	300	73	300	39	300	98	300	57	300	74	300	55
10	300	95	300	73	300	39	300	98	300	57	300	74	300	55
11	300	95	300	73	300	39	300	98	300	57	300	74	300	55
12	300	95	300	73	300	39	300	98	300	57	300	74	300	55
13	300	95	300	73	300	39	300	98	300	57	300	74	300	55
14	300	95	300	73	300	39	300	98	300	57	300	74	300	55
15	300	95	300	73	388	39	300	98	338	57	300	74	342	55
16	308	51	344	40	414	21	300	54	376	32	333	41	379	30
17	343	36	372	28	429	15	323	38	398	22	362	29	400	21
18	363	29	388	23	437	12	345	31	411	18	378	23	412	17
19	376	24	399	20	443	11	361	27	420	16	390	20	421	15
20	386	22	408	17	448	9	373	24	428	14	399	18	427	13
21	394	20	414	16	451	9	382	22	433	13	406	16	433	12
22	401	18	420	14	454	8	390	20	438	12	412	15	437	11
23	407	17	425	14	457	7	397	19	442	11	417	14	441	11
24	412	16	429	13	459	7	404	18	446	10	422	13	445	10
25	417	15	433	12	462	7	409	17	449	10	426	13	448	10
26	421	15	436	12	464	6	414	16	452	10	429	12	451	9
27	425	14	440	11	465	6	419	16	455	9	433	12	454	9
28	429	14	443	11	467	6	423	15	458	9	436	11	456	9
29	432	13	446	11	469	6	428	15	460	9	439	11	458	8
30	435	13	448	10	470	6	432	14	463	8	442	11	461	8
31	439	13	451	10	472	6	435	14	465	8	445	10	463	8
32	442	13	454	10	473	5	439	14	467	8	447	10	465	8
33	445	13	456	10	474	5	442	14	469	8	450	10	467	8
34	448	12	458	10	476	5	446	13	471	8	453	10	469	8
35	451	12	461	10	477	5	449	13	473	8	455	10	471	8
36	454	12	463	10	478	5	453	13	475	8	458	10	473	8
37	457	12	465	10	480	5	456	13	477	8	460	10	475	8
38	460	12	468	10	481	5	459	13	479	8	462	10	476	8
39	462	12	470	10	482	5	462	13	481	8	465	10	478	8
40	465	12	472	10	483	5	466	13	483	8	467	10	480	8
41	468	12	475	10	485	5	469	13	485	8	470	10	482	8
42	471	13	477	10	486	5	473	14	487	8	472	10	484	8
43	474	13	480	10	487	5	476	14	489	8	475	10	486	8
44	478	13	482	10	489	5	480	14	491	8	477	10	488	8
45	481	13	485	10	490	6	483	14	493	8	480	11	490	8
46	484	13	488	11	491	6	487	15	495	8	483	11	493	8
47	488	14	490	11	493	6	491	15	498	9	486	11	495	8
48	492	14	493	11	495	6	495	15	500	9	489	11	497	9
49	496	15	497	11	496	6	500	16	503	9	492	12	500	9
50	500	15	500	12	498	6	505	17	505	10	496	12	503	9
51	505	16	504	12	500	7	510	17	508	10	500	13	506	10
52	510	17	508	13	502	7	516	18	512	11	504	14	509	10
53	515	18	512	14	505	7	523	19	516	11	509	14	513	11

Raw Score	Grade 3		Grade 4		Grade 5		Grade 6		Grade 7		Grade 8		Grade 11	
	Scale Score	SE	Scale Score	SE	Scale Score	SE	Scale Score	SE	Scale Score	SE	Scale Score	SE	Scale Score	SE
54	522	19	517	15	507	8	530	21	520	12	515	16	517	12
55	530	21	523	16	511	9	539	23	525	13	521	17	523	13
56	539	23	531	18	515	10	550	26	531	15	529	19	529	14
57	551	27	540	21	520	12	564	30	539	17	540	22	537	17
58	569	34	554	26	528	14	584	36	551	21	555	27	548	21
59	600	49	577	38	541	21	618	51	571	30	581	39	568	29
60	657	93	622	71	566	39	679	96	607	57	628	73	603	54

Science

Raw Score	Grade 4		Grade 7		Grade 11	
	Scale Score	SE	Scale Score	SE	Scale Score	SE
1	300	134	300	129	300	67
2	300	134	300	129	300	67
3	300	134	300	129	300	67
4	300	134	300	129	300	67
5	300	134	300	129	300	67
6	300	134	300	129	300	67
7	300	134	300	129	300	67
8	300	134	300	129	300	67
9	300	134	300	129	300	67
10	300	134	300	129	300	67
11	300	134	300	129	300	67
12	300	134	300	129	300	67
13	300	134	300	129	300	67
14	300	134	300	129	300	67
15	300	134	300	129	318	67
16	300	73	300	129	362	36
17	300	52	300	71	387	25
18	300	42	300	50	401	21
19	318	36	300	41	411	18
20	334	32	301	35	418	16
21	347	29	316	32	424	14
22	358	27	329	29	429	13
23	367	25	340	27	434	12
24	375	24	350	25	438	12
25	382	23	358	24	441	11
26	389	22	366	22	444	11
27	395	21	373	22	447	10
28	401	20	379	21	450	10
29	407	20	385	20	453	10
30	412	19	390	20	455	9
31	417	19	396	19	458	9
32	422	19	401	19	460	9
33	427	19	406	18	462	9
34	431	18	410	18	464	9
35	436	18	415	18	467	9
36	441	18	419	18	469	9
37	445	18	424	18	471	9
38	450	18	428	17	473	9
39	454	18	433	17	475	9
40	459	18	437	17	477	9
41	463	18	441	17	480	9
42	468	19	445	17	482	9
43	473	19	450	17	484	9
44	478	19	454	18	486	9
45	483	20	459	18	489	10
46	488	20	463	18	491	10
47	494	21	468	18	494	10
48	500	21	473	19	497	10
49	506	22	478	19	500	11
50	513	23	483	19	503	11
51	521	24	488	20	507	12
52	529	25	494	20	511	12

Raw Score	Grade 4		Grade 7		Grade 11	
	Scale Score	SE	Scale Score	SE	Scale Score	SE
53	538	27	500	21	515	13
54	548	29	507	22	520	14
55	560	31	514	23	526	16
56	575	35	521	24	534	18
57	594	40	530	26	543	20
58	621	49	540	27	557	25
59	667	70	552	30	581	36
60	700	131	566	33	624	67
61			584	39		
62			610	48		
63			656	68		
64			700	127		

Writing

Raw Score	Grade 3		Grade 5		Grade 6		Grade 8		Grade 11	
	Scale Score	SE	Scale Score	SE	Scale Score	SE	Scale Score	SE	Scale Score	SE
1	300	103	300	59	300	91	300	93	300	84
2	300	103	300	59	300	91	300	93	300	84
3	300	103	300	59	300	91	300	93	300	84
4	300	103	300	59	300	91	300	93	300	84
5	300	103	300	59	300	91	300	93	300	84
6	300	103	300	59	300	91	300	93	300	84
7	300	103	358	59	300	91	300	93	302	84
8	333	56	398	33	324	51	319	51	356	45
9	371	39	421	23	360	36	355	36	387	31
10	392	31	435	19	381	29	376	30	404	25
11	407	27	445	17	396	26	391	26	416	22
12	419	24	453	15	408	23	403	23	425	20
13	429	23	460	14	418	21	413	21	433	19
14	437	22	466	13	427	20	422	20	440	18
15	445	21	471	13	435	20	429	19	446	17
16	453	20	476	12	443	19	437	19	453	17
17	460	20	481	12	450	19	444	19	459	17
18	467	20	486	12	457	19	450	19	465	17
19	475	21	490	12	465	19	457	19	471	17
20	483	21	495	12	473	20	464	19	477	17
21	491	22	500	13	481	21	472	20	484	18
22	500	23	505	13	490	22	480	21	491	19
23	510	25	511	14	500	23	489	22	500	21
24	522	28	518	16	512	25	500	25	510	23
25	538	31	527	18	526	29	514	28	524	27
26	559	38	539	21	546	34	533	35	543	33
27	595	55	559	31	579	48	566	49	577	48
28	660	102	595	57	636	89	625	91	635	86

APPENDIX C: Classification Consistency

Reading

R3

Level	R1	R2	R3	R4	True
R1	18.0	2.9	0.1	0.4	21.3
R2	1.7	18.3	6.7	1.1	27.8
R3	0.0	2.8	24.1	7.1	33.9
R4	0.0	0.0	1.8	15.3	17.0
Ex	19.7	23.9	32.6	23.8	100.0

R4

Level	R1	R2	R3	R4	True
R1	19.2	3.0	0.1	0.2	22.4
R2	1.4	14.5	6.1	0.6	22.7
R3	0.0	2.6	25.9	6.6	35.1
R4	0.0	0.0	2.6	17.3	19.9
Ex	20.6	20.1	34.6	24.6	100.0

R5

Level	R1	R2	R3	R4	True
R1	22.1	3.1	0.2	0.2	25.6
R2	1.7	10.8	4.8	1.1	18.4
R3	0.0	3.0	12.7	6.6	22.3
R4	0.0	0.0	3.7	29.9	33.7
Ex	23.8	16.9	21.4	37.8	100.0

R6

Level	R1	R2	R3	R4	True
R1	22.1	3.1	0.2	0.2	25.6
R2	1.7	10.8	4.8	1.1	18.4
R3	0.0	3.0	12.7	6.6	22.3
R4	0.0	0.0	3.7	29.9	33.7
Ex	23.8	16.9	21.4	37.8	100.0

R7

Level	R1	R2	R3	R4	True
R1	22.1	3.1	0.2	0.2	25.6
R2	1.7	10.8	4.8	1.1	18.4
R3	0.0	3.0	12.7	6.6	22.3
R4	0.0	0.0	3.7	29.9	33.7
Ex	23.8	16.9	21.4	37.8	100.0

R8

Level	R1	R2	R3	R4	True
R1	16.5	2.9	0.3	0.3	19.9
R2	1.4	8.7	5.7	0.6	16.3
R3	0.0	2.5	27.5	8.5	38.5
R4	0.0	0.0	3.0	22.2	25.3
Ex	17.9	14.0	36.5	31.6	100.0

R11

Level	R1	R2	R3	R4	True
R1	12.2	1.8	0.1	1.0	15.1
R2	1.2	14.2	6.5	3.0	24.8
R3	0.0	1.5	22.4	15.3	39.2
R4	0.0	0.0	1.7	19.2	20.9
Ex	13.4	17.5	30.7	38.4	100.0

Mathematics

M3

Level	M1	M2	M3	M4	True
M1	20.4	3.5	0.3	0.2	24.4
M2	1.8	10.8	6.8	0.6	20.0
M3	0.0	2.9	24.5	6.5	34.0
M4	0.0	0.0	3.3	18.3	21.6
Ex	22.3	17.3	34.9	25.5	100.0

M4

Level	M1	M2	M3	M4	True
M1	15.8	2.3	0.1	0.2	18.3
M2	1.1	13.1	5.7	0.6	20.5
M3	0.0	2.6	27.2	7.8	37.7
M4	0.0	0.0	2.5	21.1	23.5
Ex	16.8	18.0	35.4	29.8	100.0

M5

Level	M1	M2	M3	M4	True
M1	15.1	2.1	0.0	0.2	17.4
M2	1.1	16.8	6.1	0.6	24.7
M3	0.0	2.8	33.0	7.6	43.5
M4	0.0	0.0	1.4	13.0	14.4
Ex	16.2	21.7	40.6	21.4	100.0

M6

Level	M1	M2	M3	M4	True
M1	14.2	1.8	0.0	0.2	16.3
M2	1.1	12.9	6.3	0.8	21.0
M3	0.0	2.0	23.9	10.0	35.8
M4	0.0	0.0	1.6	25.2	26.9
Ex	15.3	16.7	31.9	36.1	100.0

M7

Level	M1	M2	M3	M4	True
M1	15.8	2.4	0.1	0.2	18.6
M2	1.0	8.5	6.1	0.6	16.2
M3	0.0	1.7	26.4	9.7	37.8
M4	0.0	0.0	1.8	25.6	27.5
Ex	16.8	12.6	34.5	36.1	100.0

M8

Level	M1	M2	M3	M4	True
M1	11.4	2.0	0.0	0.1	13.5
M2	0.8	14.9	6.1	0.7	22.5
M3	0.0	2.4	27.8	8.5	38.8
M4	0.0	0.0	3.3	21.9	25.2
Ex	12.3	19.3	37.2	31.2	100.0

M11

Level	M1	M2	M3	M4	True
M1	12.2	2.0	0.1	0.2	14.4
M2	1.0	9.3	6.1	0.5	16.9
M3	0.0	1.7	34.0	9.8	45.5
M4	0.0	0.0	2.4	20.8	23.2
Ex	13.2	12.9	42.6	31.2	100.0

Science

S4

Level	S1	S2	S3	S4	True
S1	9.6	1.3	0.0	0.1	11.0
S2	0.8	10.7	4.8	0.8	17.2
S3	0.0	1.6	15.2	9.8	26.6
S4	0.0	0.0	2.2	42.9	45.1
Ex	10.4	13.7	22.3	53.7	100.0

S7

Level	S1	S2	S3	S4	True
S1	11.0	1.2	0.0	0.1	12.3
S2	1.3	16.3	3.9	0.5	21.9
S3	0.0	4.9	22.7	1.9	29.6
S4	0.0	0.0	17.1	19.1	36.2
Ex	12.3	22.4	43.7	21.6	100.0

S11

Level	S1	S2	S3	S4	True
S1	10.1	1.4	0.0	0.1	11.7
S2	0.9	11.5	4.6	0.6	17.4
S3	0.0	2.3	24.9	9.5	36.7
S4	0.0	0.0	3.5	30.7	34.2
Ex	11.0	15.2	32.9	40.9	100.0

Writing

W3

Level	W1	W2	W3	W4	True
W1	11.8	2.1	0.0	0.2	14.2
W2	1.3	12.2	7.6	2.0	23.1
W3	0.1	2.0	18.6	9.9	30.6
W4	0.0	0.0	3.5	28.7	32.2
Ex	13.2	16.3	29.7	40.8	100.0

W5

Level	W1	W2	W3	W4	True
W1	10.1	2.1	0.0	0.2	12.5
W2	0.9	12.0	7.0	0.9	20.9
W3	0.0	2.5	33.9	9.2	45.6
W4	0.0	0.0	2.2	18.8	21.0
Ex	11.0	16.6	43.1	29.2	100.0

W6

Level	W1	W2	W3	W4	True
W1	10.1	2.1	0.0	0.2	12.5
W2	0.9	12.0	7.0	0.9	20.9
W3	0.0	2.5	33.9	9.2	45.6
W4	0.0	0.0	2.2	18.8	21.0
Ex	11.0	16.6	43.1	29.2	100.0

W8

Level	W1	W2	W3	W4	True
W1	16.1	4.2	0.3	0.6	21.1
W2	1.7	16.8	7.4	1.8	27.7
W3	0.0	3.9	18.0	7.7	29.6
W4	0.0	0.0	2.4	19.2	21.6
Ex	17.8	24.8	28.2	29.2	100.0

W11

Level	W1	W2	W3	W4	True
W1	13.9	2.9	0.2	0.4	17.4
W2	1.6	14.0	5.7	1.2	22.5
W3	0.0	4.9	25.6	9.1	39.6
W4	0.0	0.0	2.0	18.5	20.5
Ex	15.6	21.8	33.4	29.2	100.0

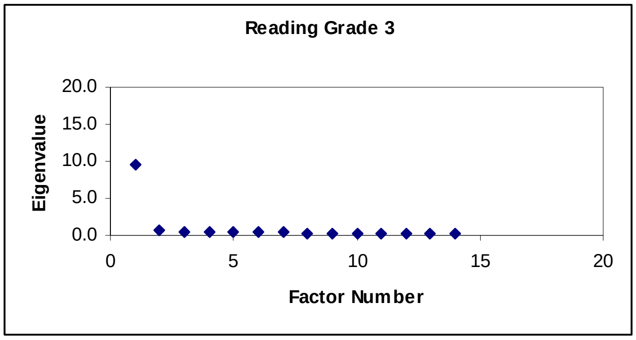
APPENDIX D: First Ten Eigenvalues from the Principal Component Analysis

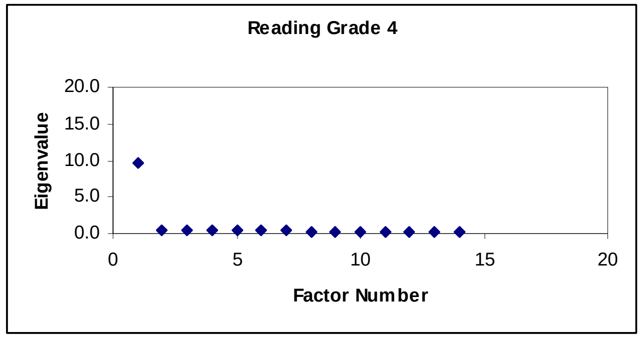
Grade	Reading		Mathematics		Science		Writing	
	Number	Eigenvalue	Number	Eigenvalue	Number	Eigenvalue	Number	Eigenvalue
3	1	9.509	1	9.444			1	5.063
	2	0.677	2	1.045			2	0.641
	3	0.464	3	0.696			3	0.396
	4	0.453	4	0.510			4	0.338
	5	0.402	5	0.448			5	0.245
	6	0.379	6	0.389			6	0.181
	7	0.361	7	0.354			7	0.136
	8	0.327	8	0.350				
	9	0.293	9	0.310				
	10	0.276	10	0.289				
4	1	9.671	1	10.655	1	10.234		
	2	0.552	2	0.651	2	0.598		
	3	0.503	3	0.525	3	0.521		
	4	0.495	4	0.456	4	0.481		
	5	0.414	5	0.381	5	0.405		
	6	0.374	6	0.331	6	0.390		
	7	0.366	7	0.309	7	0.365		
	8	0.331	8	0.296	8	0.337		
	9	0.279	9	0.260	9	0.331		
	10	0.260	10	0.237	10	0.290		
5	1	9.165	1	10.387			1	4.922
	2	0.739	2	0.758			2	0.437
	3	0.554	3	0.469			3	0.373
	4	0.491	4	0.396			4	0.361
	5	0.424	5	0.386			5	0.320
	6	0.406	6	0.334			6	0.308
	7	0.363	7	0.314			7	0.280
	8	0.353	8	0.295				
	9	0.322	9	0.283				
	10	0.276	10	0.277				
6	1	9.681	1	10.545			1	5.041
	2	0.629	2	0.692			2	0.451
	3	0.488	3	0.503			3	0.412
	4	0.437	4	0.402			4	0.369
	5	0.403	5	0.363			5	0.289
	6	0.363	6	0.323			6	0.236
	7	0.330	7	0.318			7	0.201
	8	0.283	8	0.290				
	9	0.271	9	0.281				
	10	0.258	10	0.262				
7	1	9.342	1	9.471	1	10.709		
	2	0.638	2	1.114	2	0.965		
	3	0.512	3	0.549	3	0.545		

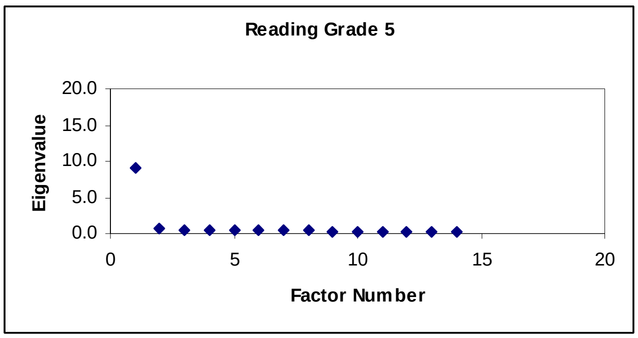
Grade	Reading		Mathematics		Science		Writing	
	Number	Eigenvalue	Number	Eigenvalue	Number	Eigenvalue	Number	Eigenvalue
	4	0.455	4	0.494	4	0.448		
	5	0.434	5	0.465	5	0.422		
	6	0.417	6	0.428	6	0.409		
	7	0.361	7	0.365	7	0.384		
	8	0.352	8	0.332	8	0.353		
	9	0.305	9	0.324	9	0.282		
	10	0.299	10	0.300	10	0.279		
8	1	9.664	1	10.417			1	4.969
	2	0.709	2	0.885			2	0.674
	3	0.498	3	0.483			3	0.369
	4	0.458	4	0.445			4	0.318
	5	0.422	5	0.397			5	0.277
	6	0.365	6	0.349			6	0.231
	7	0.332	7	0.323			7	0.163
	8	0.290	8	0.282				
	9	0.280	9	0.246				
	10	0.253	10	0.228				
11	1	8.927	1	10.603	1	11.082	1	5.568
	2	0.340	2	0.623	2	0.546	2	0.528
	3	0.318	3	0.500	3	0.479	3	0.273
	4	0.267	4	0.451	4	0.382	4	0.230
	5	0.250	5	0.391	5	0.366	5	0.167
	6	0.208	6	0.355	6	0.321	6	0.137
	7	0.178	7	0.326	7	0.288	7	0.098
	8	0.154	8	0.307	8	0.263		
	9	0.147	9	0.256	9	0.241		
	10	0.111	10	0.244	10	0.224		

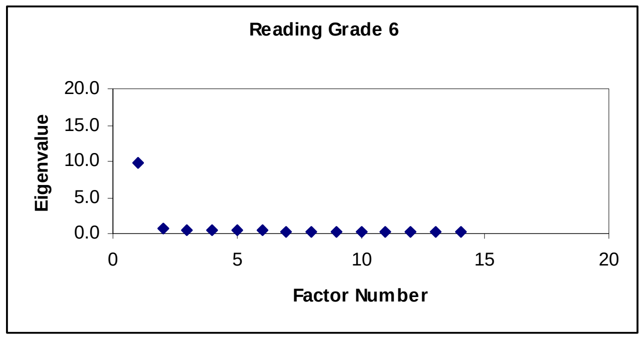
APPENDIX E: Principal Component Analysis Scree Plot

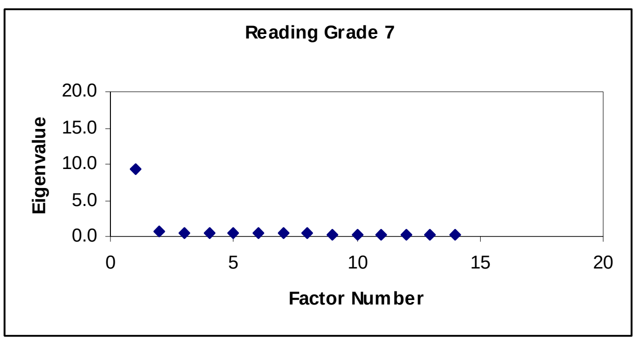
Reading

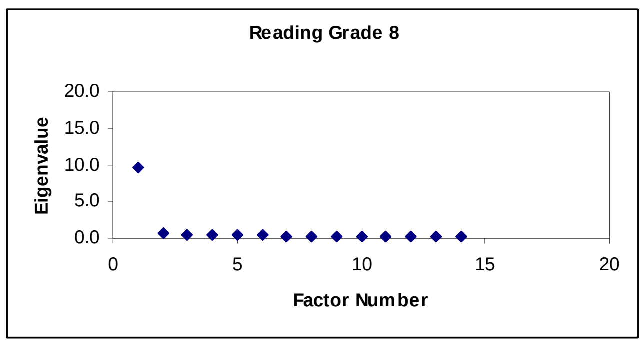


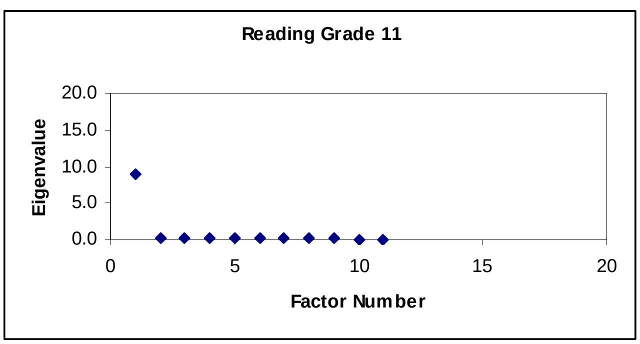




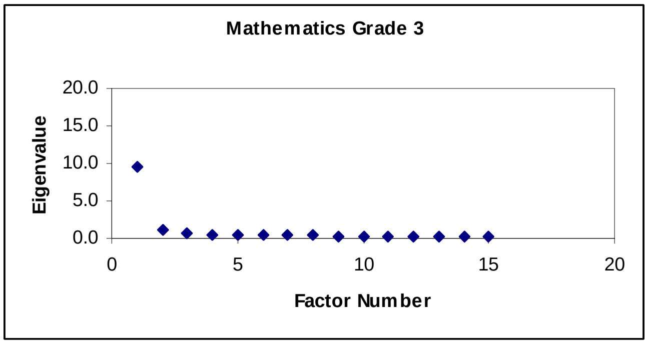


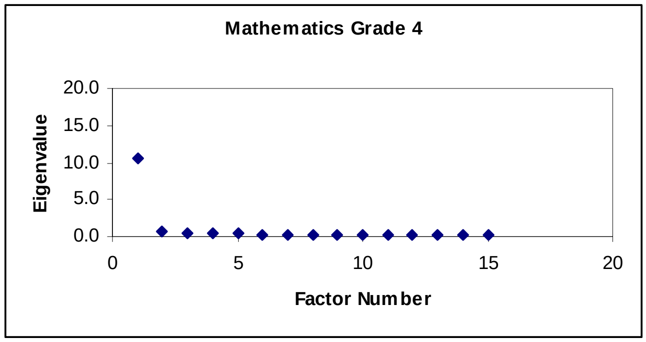


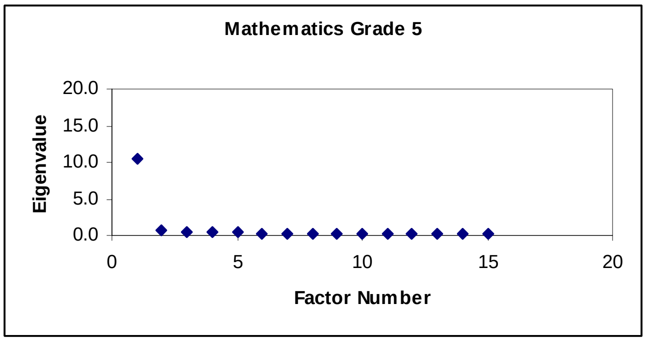


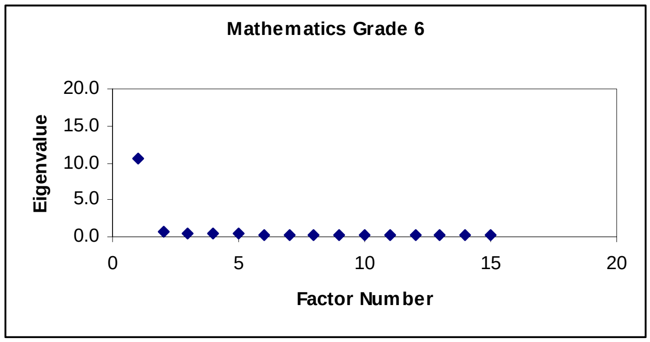


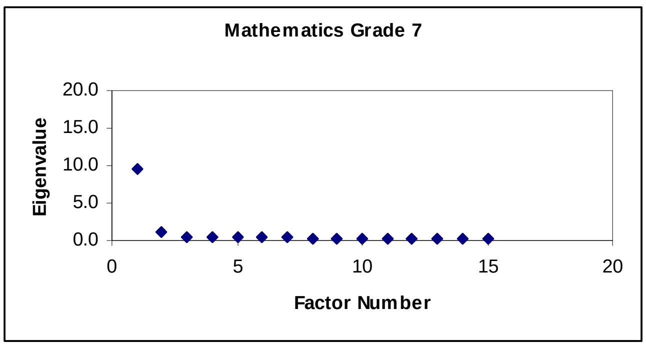
Mathematics

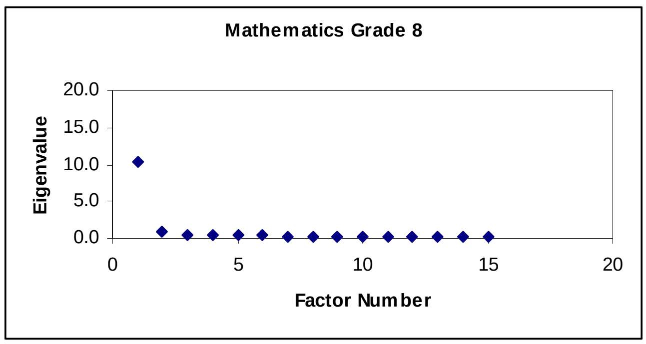


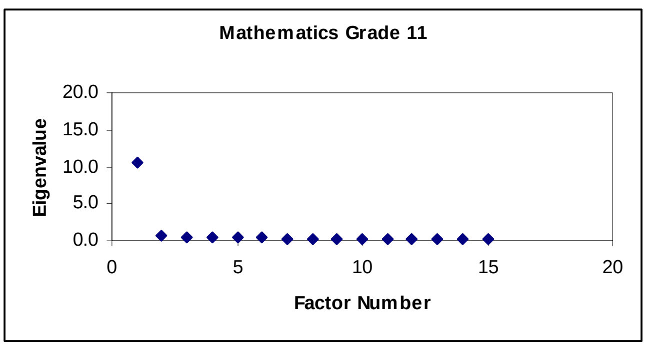




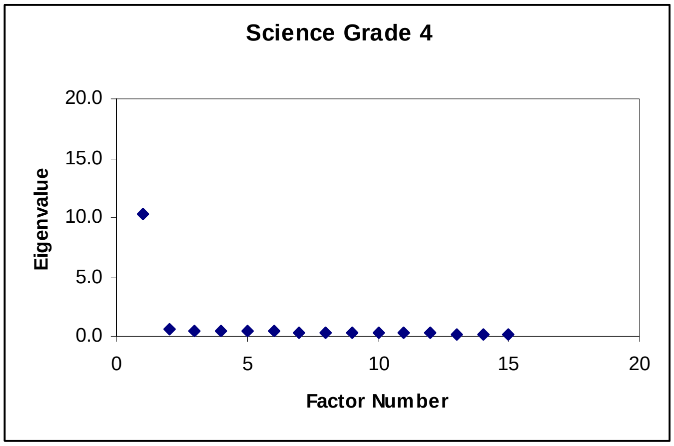




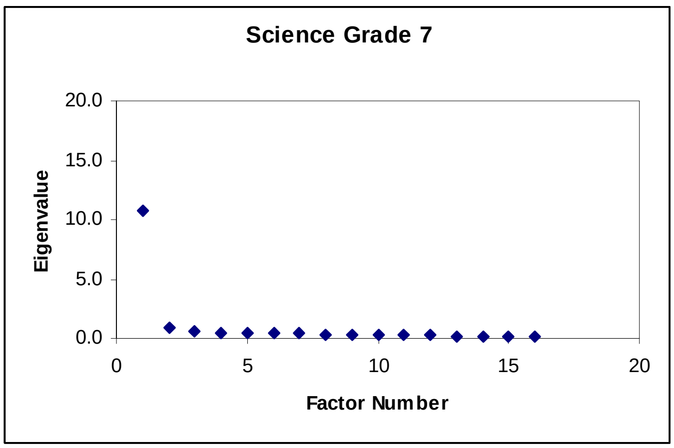




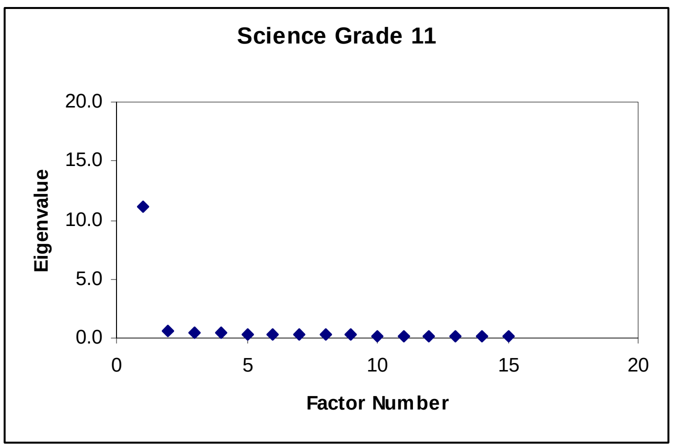
Science



Science Grade 4
Eigenvalue
20.0
15.0
10.0
5.0
0.0
0
5
10
15
20
Factor Number

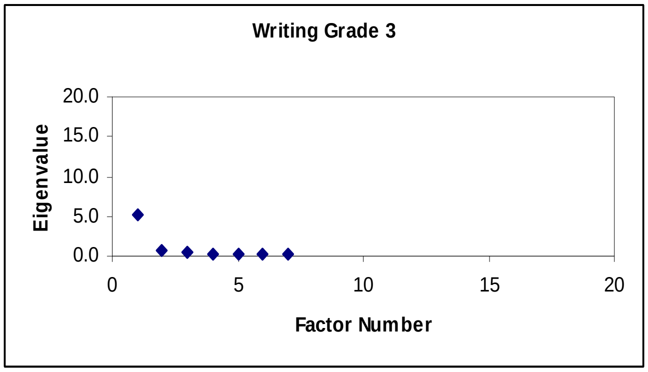


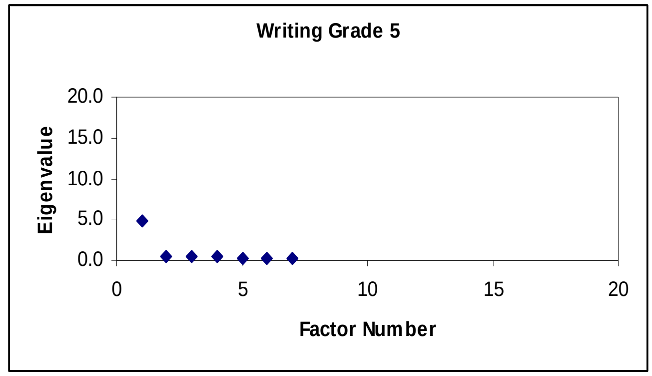
Science Grade 7
Eigenvalue
20.0
15.0
10.0
5.0
0.0
0
5
10
15
20
Factor Number

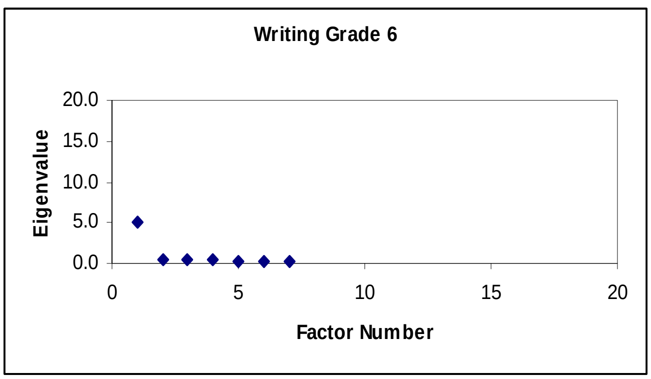


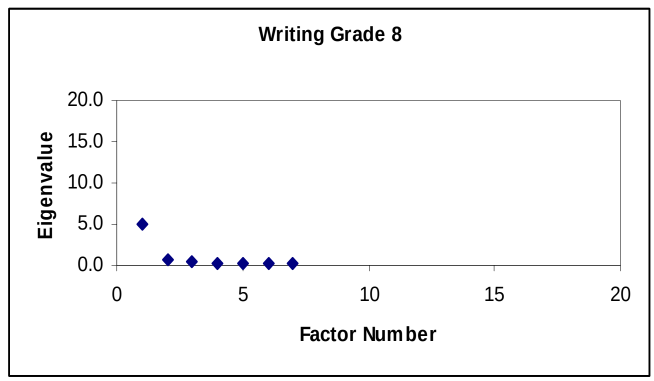
Science Grade 11
Eigenvalue
20.0
15.0
10.0
5.0
0.0
0
5
10
15
20
Factor Number

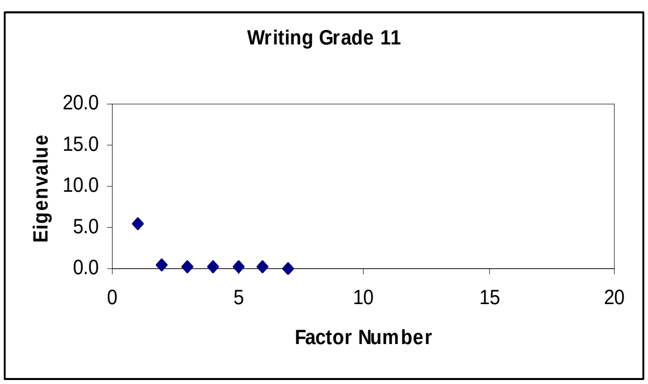
Writing


Writing Grade 3
Eigenvalue
20.0
15.0
10.0
5.0
0.0
0
5
10
15
20
Factor Number


Writing Grade 5
Eigenvalue
20.0
15.0
10.0
5.0
0.0
0
5
10
15
20
Factor Number


Writing Grade 6
Eigenvalue
20.0
15.0
10.0
5.0
0.0
0
5
10
15
20
Factor Number


Writing Grade 8
Eigenvalue
20.0
15.0
10.0
5.0
0.0
0
5
10
15
20
Factor Number


Writing Grade 11
Eigenvalue
20.0
15.0
10.0
5.0
0.0
0
5
10
15
20
Factor Number

APPENDIX F: Rater Agreement by Item

Reading

Grade	Item	*N*	% of exact agreement	% of adjacent agreement	% of total agreement
3	1	14	100.00	0.00	100.00
	2	14	100.00	0.00	100.00
	3	14	100.00	0.00	100.00
	4	14	100.00	0.00	100.00
	5	14	100.00	0.00	100.00
	6	14	100.00	0.00	100.00
	7	14	100.00	0.00	100.00
	8	14	100.00	0.00	100.00
	9	14	100.00	0.00	100.00
	10	14	100.00	0.00	100.00
	11	14	100.00	0.00	100.00
	12	14	100.00	0.00	100.00
	13	14	92.86	7.14	100.00
	14	14	100.00	0.00	100.00
	15	14	100.00	0.00	100.00
	16	14	100.00	0.00	100.00
	17	14	100.00	0.00	100.00
	18	14	100.00	0.00	100.00
4	1	21	90.91	4.55	95.46
	2	22	95.65	0.00	95.65
	3	23	95.65	4.35	100.00
	4	22	86.96	8.70	95.66
	5	23	95.65	4.35	100.00
	6	22	95.65	0.00	95.65
	7	21	91.30	0.00	91.30
	8	23	100.00	0.00	100.00
	9	23	100.00	0.00	100.00
	10	22	95.65	0.00	95.65
	11	21	86.96	4.35	91.31
	12	23	91.30	8.70	100.00
	13	21	90.91	4.55	95.46
	14	22	90.91	9.09	100.00
	15	22	90.91	9.09	100.00
	16	21	95.45	0.00	95.45
	17	21	100.00	0.00	100.00
	18	20	95.00	5.00	100.00
5	1	16	100.00	0.00	100.00
	2	16	100.00	0.00	100.00
	3	16	100.00	0.00	100.00
	4	16	100.00	0.00	100.00
	5	14	93.33	0.00	93.33
	6	14	100.00	0.00	100.00
	7	13	100.00	0.00	100.00

Grade	Item	*N*	% of exact agreement	% of adjacent agreement	% of total agreement
	8	14	100.00	0.00	100.00
	9	13	100.00	0.00	100.00
	10	16	100.00	0.00	100.00
	11	16	100.00	0.00	100.00
	12	16	100.00	0.00	100.00
	13	16	100.00	0.00	100.00
	14	15	100.00	0.00	100.00
	15	16	100.00	0.00	100.00
	16	15	100.00	0.00	100.00
	17	15	100.00	0.00	100.00
	18	15	86.67	13.33	100.00
6	1	11	91.67	0.00	91.67
	2	12	100.00	0.00	100.00
	3	12	100.00	0.00	100.00
	4	12	100.00	0.00	100.00
	5	12	91.67	8.33	100.00
	6	12	100.00	0.00	100.00
	7	12	100.00	0.00	100.00
	8	12	100.00	0.00	100.00
	9	12	100.00	0.00	100.00
	10	14	100.00	0.00	100.00
	11	13	100.00	0.00	100.00
	12	13	100.00	0.00	100.00
	13	12	100.00	0.00	100.00
	14	12	91.67	8.33	100.00
	15	12	100.00	0.00	100.00
	16	12	100.00	0.00	100.00
	17	12	100.00	0.00	100.00
	18	12	100.00	0.00	100.00
7	1	18	100.00	0.00	100.00
	2	19	94.74	5.26	100.00
	3	19	100.00	0.00	100.00
	4	19	100.00	0.00	100.00
	5	19	100.00	0.00	100.00
	6	19	100.00	0.00	100.00
	7	19	100.00	0.00	100.00
	8	19	100.00	0.00	100.00
	9	19	100.00	0.00	100.00
	10	19	100.00	0.00	100.00
	11	19	100.00	0.00	100.00
	12	19	100.00	0.00	100.00
	13	19	100.00	0.00	100.00
	14	19	100.00	0.00	100.00
	15	19	94.74	5.26	100.00
	16	19	94.74	5.26	100.00
	17	18	94.44	5.56	100.00
	18	19	100.00	0.00	100.00
8	1	15	100.00	0.00	100.00

Grade	Item	*N*	% of exact agreement	% of adjacent agreement	% of total agreement
	2	15	100.00	0.00	100.00
	3	14	100.00	0.00	100.00
	4	15	100.00	0.00	100.00
	5	15	100.00	0.00	100.00
	6	16	100.00	0.00	100.00
	7	17	100.00	0.00	100.00
	8	16	94.12	0.00	94.12
	9	16	100.00	0.00	100.00
	10	18	88.89	11.11	100.00
	11	18	100.00	0.00	100.00
	12	18	94.44	5.56	100.00
	13	16	94.12	0.00	94.12
	14	17	100.00	0.00	100.00
	15	17	100.00	0.00	100.00
	16	17	94.12	5.88	100.00
	17	16	100.00	0.00	100.00
	18	16	87.50	12.50	100.00
11	1	27	100.00	0.00	100.00
	2	27	100.00	0.00	100.00
	3	27	100.00	0.00	100.00
	4	27	100.00	0.00	100.00
	5	27	100.00	0.00	100.00
	6	26	100.00	0.00	100.00
	7	27	96.30	3.70	100.00
	8	26	100.00	0.00	100.00
	9	27	100.00	0.00	100.00
	10	27	100.00	0.00	100.00
	11	27	96.30	3.70	100.00
	12	28	100.00	0.00	100.00
	13	28	100.00	0.00	100.00
	14	28	100.00	0.00	100.00
	15	28	100.00	0.00	100.00

Mathematics

Grade	Item	*N*	% of exact agreement	% of adjacent agreement	% of total agreement
3	1	13	100.00	0.00	100.00
	2	13	100.00	0.00	100.00
	3	12	92.31	0.00	92.31
	4	13	100.00	0.00	100.00
	5	13	92.31	7.69	100.00
	6	13	100.00	0.00	100.00
	7	13	100.00	0.00	100.00
	8	13	100.00	0.00	100.00
	9	13	100.00	0.00	100.00
	10	13	100.00	0.00	100.00
	11	13	100.00	0.00	100.00
	12	13	100.00	0.00	100.00
	13	13	100.00	0.00	100.00
	14	13	100.00	0.00	100.00
	15	13	100.00	0.00	100.00
	16	13	100.00	0.00	100.00
	17	13	100.00	0.00	100.00
	18	13	100.00	0.00	100.00
	19	13	100.00	0.00	100.00
4	1	16	93.75	6.25	100.00
	2	16	93.75	6.25	100.00
	3	16	100.00	0.00	100.00
	4	16	100.00	0.00	100.00
	5	15	93.75	0.00	93.75
	6	15	93.75	0.00	93.75
	7	16	100.00	0.00	100.00
	8	16	100.00	0.00	100.00
	9	16	100.00	0.00	100.00
	10	16	93.75	6.25	100.00
	11	16	100.00	0.00	100.00
	12	16	100.00	0.00	100.00
	13	16	93.75	6.25	100.00
	14	16	100.00	0.00	100.00
	15	16	100.00	0.00	100.00
	16	16	93.75	6.25	100.00
	17	16	93.75	6.25	100.00
	18	16	100.00	0.00	100.00
	19	15	93.75	0.00	93.75
5	1	17	100.00	0.00	100.00
	2	17	94.12	5.88	100.00
	3	17	94.12	5.88	100.00
	4	17	100.00	0.00	100.00
	5	17	94.12	5.88	100.00
	6	17	100.00	0.00	100.00
	7	17	94.12	5.88	100.00
	8	17	94.12	5.88	100.00

Grade	Item	*N*	% of exact agreement	% of adjacent agreement	% of total agreement
	9	17	100.00	0.00	100.00
	10	17	100.00	0.00	100.00
	11	16	100.00	0.00	100.00
	12	16	93.75	6.25	100.00
	13	16	93.75	6.25	100.00
	14	16	100.00	0.00	100.00
	15	16	100.00	0.00	100.00
	16	16	100.00	0.00	100.00
	17	16	93.75	6.25	100.00
	18	16	100.00	0.00	100.00
	19	16	93.75	6.25	100.00
6	1	15	100.00	0.00	100.00
	2	15	93.33	6.67	100.00
	3	15	100.00	0.00	100.00
	4	15	100.00	0.00	100.00
	5	15	100.00	0.00	100.00
	6	15	100.00	0.00	100.00
	7	15	100.00	0.00	100.00
	8	15	100.00	0.00	100.00
	9	15	100.00	0.00	100.00
	10	15	100.00	0.00	100.00
	11	15	100.00	0.00	100.00
	12	15	100.00	0.00	100.00
	13	15	100.00	0.00	100.00
	14	15	100.00	0.00	100.00
	15	14	100.00	0.00	100.00
	16	15	100.00	0.00	100.00
	17	15	100.00	0.00	100.00
	18	15	100.00	0.00	100.00
	19	15	100.00	0.00	100.00
7	1	14	100.00	0.00	100.00
	2	14	100.00	0.00	100.00
	3	13	92.86	0.00	92.86
	4	13	85.71	7.14	92.85
	5	13	92.86	0.00	92.86
	6	14	100.00	0.00	100.00
	7	14	100.00	0.00	100.00
	8	14	100.00	0.00	100.00
	9	14	100.00	0.00	100.00
	10	14	100.00	0.00	100.00
	11	14	100.00	0.00	100.00
	12	14	100.00	0.00	100.00
	13	14	100.00	0.00	100.00
	14	14	100.00	0.00	100.00
	15	14	100.00	0.00	100.00
	16	14	100.00	0.00	100.00
	17	14	85.71	14.29	100.00
	18	14	100.00	0.00	100.00

Grade	Item	*N*	% of exact agreement	% of adjacent agreement	% of total agreement
	19	13	100.00	0.00	100.00
8	1	9	100.00	0.00	100.00
	2	9	100.00	0.00	100.00
	3	9	100.00	0.00	100.00
	4	10	100.00	0.00	100.00
	5	10	100.00	0.00	100.00
	6	11	100.00	0.00	100.00
	7	11	100.00	0.00	100.00
	8	11	100.00	0.00	100.00
	9	10	100.00	0.00	100.00
	10	11	100.00	0.00	100.00
	11	11	100.00	0.00	100.00
	12	11	100.00	0.00	100.00
	13	11	100.00	0.00	100.00
	14	11	100.00	0.00	100.00
	15	11	100.00	0.00	100.00
	16	11	100.00	0.00	100.00
	17	10	100.00	0.00	100.00
	18	10	100.00	0.00	100.00
	19	10	90.00	10.00	100.00
11	1	16	100.00	0.00	100.00
	2	15	100.00	0.00	100.00
	3	16	100.00	0.00	100.00
	4	16	100.00	0.00	100.00
	5	16	100.00	0.00	100.00
	6	17	100.00	0.00	100.00
	7	17	100.00	0.00	100.00
	8	17	100.00	0.00	100.00
	9	17	88.24	11.76	100.00
	10	17	94.12	5.88	100.00
	11	16	94.12	0.00	94.12
	12	17	100.00	0.00	100.00
	13	17	100.00	0.00	100.00
	14	17	88.24	11.76	100.00
	15	17	100.00	0.00	100.00
	16	17	100.00	0.00	100.00
	17	17	100.00	0.00	100.00
	18	17	100.00	0.00	100.00
	19	17	94.12	5.88	100.00

Science

Grade	Item	*N*	% of exact agreement	% of adjacent agreement	% of total agreement
4	1	19	94.74	5.26	100.00
	2	19	100.00	0.00	100.00
	3	19	95.00	0.00	95.00
	4	20	100.00	0.00	100.00
	5	20	100.00	0.00	100.00
	6	20	90.00	10.00	100.00
	7	18	100.00	0.00	100.00
	8	19	94.74	5.26	100.00
	9	19	100.00	0.00	100.00
	10	19	100.00	0.00	100.00
	11	19	100.00	0.00	100.00
	12	20	100.00	0.00	100.00
	13	20	100.00	0.00	100.00
	14	20	95.00	5.00	100.00
	15	20	90.00	10.00	100.00
	16	20	95.00	5.00	100.00
	17	19	90.00	5.00	95.00
	18	20	90.00	10.00	100.00
	19	20	95.00	5.00	100.00
7	1	15	100.00	0.00	100.00
	2	15	100.00	0.00	100.00
	3	15	100.00	0.00	100.00
	4	15	100.00	0.00	100.00
	5	15	93.33	6.67	100.00
	6	15	100.00	0.00	100.00
	7	15	100.00	0.00	100.00
	8	15	100.00	0.00	100.00
	9	15	100.00	0.00	100.00
	10	15	100.00	0.00	100.00
	11	15	100.00	0.00	100.00
	12	15	93.33	6.67	100.00
	13	15	86.67	13.33	100.00
	14	15	100.00	0.00	100.00
	15	15	100.00	0.00	100.00
	16	15	100.00	0.00	100.00
	17	15	100.00	0.00	100.00
	18	15	100.00	0.00	100.00
	19	15	100.00	0.00	100.00
	20	15	100.00	0.00	100.00
11	1	16	100.00	0.00	100.00
	2	16	100.00	0.00	100.00
	3	16	100.00	0.00	100.00
	4	16	100.00	0.00	100.00
	5	16	100.00	0.00	100.00
	6	16	100.00	0.00	100.00
	7	16	100.00	0.00	100.00

Grade	Item	*N*	% of exact agreement	% of adjacent agreement	% of total agreement
	8	16	100.00	0.00	100.00
	9	16	100.00	0.00	100.00
	10	16	100.00	0.00	100.00
	11	16	100.00	0.00	100.00
	12	16	100.00	0.00	100.00
	13	16	87.50	12.50	100.00
	14	16	100.00	0.00	100.00
	15	16	100.00	0.00	100.00
	16	16	100.00	0.00	100.00
	17	16	100.00	0.00	100.00
	18	16	81.25	18.75	100.00
	19	16	100.00	0.00	100.00

Writing

Grade	Item	*N*	% of exact agreement	% of adjacent agreement	% of total agreement
3	1	13	100.00	0.00	100.00
	2	13	100.00	0.00	100.00
	3	13	100.00	0.00	100.00
	4	12	92.31	0.00	92.31
	5	13	100.00	0.00	100.00
	6	13	100.00	0.00	100.00
	7	13	100.00	0.00	100.00
	8	13	100.00	0.00	100.00
5	1	20	95.00	5.00	100.00
	2	19	95.00	0.00	95.00
	3	19	95.00	0.00	95.00
	4	20	95.00	5.00	100.00
	5	20	100.00	0.00	100.00
	6	20	100.00	0.00	100.00
	7	20	100.00	0.00	100.00
	8	19	100.00	0.00	100.00
6	1	16	100.00	0.00	100.00
	2	16	100.00	0.00	100.00
	3	16	93.75	6.25	100.00
	4	16	100.00	0.00	100.00
	5	16	100.00	0.00	100.00
	6	16	100.00	0.00	100.00
	7	16	100.00	0.00	100.00
	8	16	100.00	0.00	100.00
8	1	16	100.00	0.00	100.00
	2	16	100.00	0.00	100.00
	3	16	100.00	0.00	100.00
	4	16	100.00	0.00	100.00
	5	16	100.00	0.00	100.00
	6	16	100.00	0.00	100.00
	7	16	100.00	0.00	100.00
	8	16	100.00	0.00	100.00
11	1	14	93.33	0.00	93.33
	2	15	93.33	6.67	100.00
	3	15	100.00	0.00	100.00
	4	15	93.33	6.67	100.00
	5	15	100.00	0.00	100.00
	6	15	100.00	0.00	100.00
	7	15	100.00	0.00	100.00
	8	15	100.00	0.00	100.00

APPENDIX G: IAA Standard Validation Evaluation form

The purpose of this evaluation is to obtain your feedback about the standards validation process. Your feedback will provide a basis for evaluating the training, methods, and materials in the standards validation process.

Please complete the information below. Do not put your name on the form as we want your feedback to be anonymous.

Panel: Please place an X in one box that describes the panel you served

☐ Reading (Grades 3-5)
☐ Reading (Grades 6-8, & 11)
☐ Mathematics (Grades 3-5)
☐ Mathematics (Grades 6-8, & 11)
☐ Science (Grades 4, 7, & 11)
☐ Writing (Grades 3, 5, 6, 8, & 11)

1. Please read each of the following statements carefully. Place an X in one box for each statement to indicate the degree to which you agree with each statement.

		Strongly Agree	Agree	Disagree	Strongly disagree
a.	I understood the purpose of this standards validation workshop.	☐	☐	☐	☐
b.	The description of the performance level descriptors was clear and understandable.	☐	☐	☐	☐
c.	The description of the reasoned judgment process was clear and understandable.	☐	☐	☐	☐
d.	The description of the feedback data was clear and understandable.	☐	☐	☐	☐

2. Please rate the usefulness of the following materials or procedures in completing the standards validation process. Place an X in one box for each statement to indicate the degree to which you agree with each statement.

		Very useful	Somewhat useful	Not at all useful
a.	Reviewing test materials	☐	☐	☐
b.	Training	☐	☐	☐
c.	Group discussions	☐	☐	☐

3. How important was each of the following factors in your validation of the cut scores? Place an X in one box for each statement to indicate the degree to which you agree with each statement.

		Very important	Somewhat important	Not important
a.	The description of performance level descriptors	☐	☐	☐
b.	Your perception of the importance of particular score patterns	☐	☐	☐
c.	Your experiences with students	☐	☐	☐
d.	Group discussions	☐	☐	☐
e.	Agreement on rater location data	☐	☐	☐
f.	Impact data	☐	☐	☐

4. Were any materials or procedures especially influential in your evaluation of the cut scores? If so, which ones? In what ways were they especially influential?

____

____

____

____

____

5. How appropriate was the amount of time you were given to complete the different components of the standards validation process? Place an X in one box for each statement to indicate the degree to which you agree with each statement.

		Too much	About right	Too little
a.	Training on the standards validation process	☐	☐	☐
b.	Group discussions	☐	☐	☐

6. What suggestions do you have to improve the standard validation process and the training? Please use the reverse side if necessary.

____

____

____

____

APPENDIX H: Alignment Study