

Illinois Alternate Assessment
2011 Technical Manual

Illinois State Board of Education
Division of Assessment

Table of Contents

1. PURPOSE AND DESIGN OF THE IAA TESTING PROGRAM	3
NCLB Requirements	3
Excerpts from the August 2005 Non-Regulatory Guidance	4
Test Development and Test Blueprint	6
Item Development	10
Item Development Cycle.....	10
Item Specifications.....	11
Test Administration Training.....	13
Test Implementation Manual	13
Test Booklets	14
Student Score Sheets	14
Online Test Platform	14
Teacher Training	14
Bias Review	15
Differential Item Functioning.....	15
Internal Consistency	20
Standard Error of Measurement	21
IRT Test Information Function	24
IRT Conditional SEM.....	27
Classification Accuracy	27
3. VALIDITY	30
Performance-Based Measurement	30
Content-related Validity	32
Construct Validity	33
Dimensionality	33
Internal Construct	36
Criterion-related Validity	38
Agreement between Teacher Scores and Expert Scores	39
Correlations between Teacher Scores and Expert Scores	40
Validity Related to Comparison to Typical Performance.....	42
Familiarity with Students.....	42
Comparison to Typical Performance.....	43
4. CALIBRATION AND SCALING.....	45
Calibration	45
Stability Check.....	45
Scaling.....	46
Apply Scale Transformation Constants and Define Scale Score Cuts	46
5. RESULTS	48
Performance Relative to Illinois Alternate Assessment Frameworks	48
REFERENCES	51
APPENDIX A: IAA Scoring Rubric	53
APPENDIX B: Conditional Standard Errors of Measurement Associated with IAA Scale Scores.....	54
APPENDIX C: Classification Consistency	61
APPENDIX D: First Ten Eigenvalues from the Principal Component Analysis.....	65
APPENDIX E: Scree Plots for All Components	67
APPENDIX F: Agreement between Teacher and Expert Scores by Item	70
APPENDIX G: IAA Performance Theta Cuts and Transformation Constants.....	79
APPENDIX H: Item Statistics Summary.....	80

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 later amended to meet the requirements of the No Child Left Behind Act (NCLB) of 2001. These laws mandated that an alternate assessment be in place for those students with the most 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 were developed and piloted for the IAA in fall 2007, and operationally administered in spring 2008.

This technical manual provides technical information on 2011 IAA tests. In particular, this manual addresses test development, implementation, scoring, and technical attributes of the IAA.

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 align with the State's academic content standards; 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 that states could use to develop alternate assessments, but offers 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 based on alternate achievement standards. It specifically states that there is no typical format and suggests 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 of measuring 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://www.cehd.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 for 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 Adequate Yearly Progress (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 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.

Summary of Program Changes for 2011

There were some major changes between the 2010 administration and the current 2011 administration. Two points need to be noted about the selection of 2011 anchor sets. First, there were changes to the way 2011 IAA Reading tests were administered. Before 2011, test administrators were required to read the passages, questions, and answer options aloud to all the students. In 2011, the anchor items of Reading were administered the same way as in 2010, but for other operational Reading items (depending on the Assessment Objective), the test administrators read the passages aloud only to students who qualified to use read aloud as an accommodation within their IEP program. Test administrators still read the questions and answer options aloud to all the students. Second, as per ISBE, the distinction of 2-page and 4-page items was not considered in selecting the anchor set for 2011.

Starting with 2011 pilot items, ISBE changed the rubric for the score point of 4. Before 2011, if a student does not answer correctly after the task is presented once, the test administrator would repeat the task. If the student gets the correct answer this time, the student would still get a score point of 4. Starting with 2011 pilot items, if a student did not get the correct answer after the task is presented once, the test administrator would move on to the rubric of score point 3, instead of repeating the task again.

There will be a 3-year phase in concerning the change to the 4pt. score of the rubric (dropping repeat).

1. 2011 (yr1) - change applied to FT items only.
2. 2012 (yr2) – operational items (field tested in 2011) will reflect the change; linking items will not
3. 2013 (yr3) – operational items and new linking items will reflect this change

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 was 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 the 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 would 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 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 tests. 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, which was 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 starting 2008.

In 2009, the IAA was further improved in two respects: a standardized test administration procedure and increased test length. Standardization of IAA administration was achieved by: (1) incorporating supplemental testing materials into the test booklet, (2) using a prescriptive scoring rubric to increase consistency in scoring, and (3) including the rubric in the booklet for convenience in the administration process. A comparison of test lengths for the 2008 and 2009 administrations can be found in Table 1.1. In light of these changes and the establishment of a new scale in 2009, it was decided that only item statistics from 2008 field test and item statistics from 2009 operational tests and thereafter would be included in the item bank.

Table 1.1: Comparison of 2008 and 2009 IAA Test Length

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

For 2011 and 2010, the test length of the IAA stayed the same as 2009 for all subjects and grades. For 2011, the Writing test was only administered to grade 11. The 2011 blueprint of census items for each subject is presented in Tables 1.2a through 1.2d.

Table 1.2a: Reading Blueprint

Grade	Goal	Number of Items	Percent of Items
3	1	11	79
3	2	3	21
4	1	10	71
4	2	4	29
5	1	9	64
5	2	5	36
6	1	10	71
6	2	4	29
7	1	10	71
7	2	4	29
8	1	8	57
8	2	6	43
11	1	11	100

Table 1.2b: Mathematics Blueprint

Grade	Goal	Number of Items	Percent of Items
3	6	7	47
3	7	2	13
3	8	2	13
3	9	2	13
3	10	2	13
4	6	7	47
4	7	2	13
4	8	2	13
4	9	2	13
4	10	2	13
5	6	6	40
5	7	2	13
5	8	3	20
5	9	2	13
5	10	2	13
6	6	4	27
6	7	2	13
6	8	4	27
6	9	3	20
6	10	2	13
7	6	4	27
7	7	3	20
7	8	3	20
7	9	3	20
7	10	2	13
8	6	4	27
8	7	3	20
8	8	3	20
8	9	2	13
8	10	3	20
11	6	5	33
11	7	3	20
11	8	2	13
11	9	4	27
11	10	1	7

Table 1.2c: Science Blueprint

Grade	Goal	Number of Items	Percent of Items
4	11	2	13
4	12	10	67
4	13	3	20
7	11	2	13
7	12	11	69
7	13	3	19
11	11	2	13
11	12	11	73
11	13	2	13

Table 1.2d: Writing Blueprint

Grade	Goal	Number of Items	Percent of Items
11	3	7	100

Note: Previously, Writing was also administered in grades 3, 5, 6, & 8.

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. Before a new item is used on a test, it is evaluated by content experts and teacher panels through qualitative and quantitative approaches. 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 Development** – Author; review and edit items 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.
4. **Customer Preview** – Review by and feedback from ISBE staff on all items developed for each subject to check for a common understanding of ISBE expectations for quality and for content and cognitive mapping.

5. **Committee Reviews** – Review of passages and items by Illinois stakeholders for content and bias/sensitivity with Pearson staff. Items that are suspected of bias are not used in the test.
6. **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.
7. **Pilot Test Administration** – Test embedded pilot items along with census items. The IAA is tested annually between February and March.
8. **Data Review** – Perform different item analyses on the pilot test items after test administration. The analysis 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.
9. **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 to select census items. Psychometric review of item and test statistics is implemented to add to the quality of the tests.
10. **Census Test Administration** – Test census items along with pilot items. The IAA is tested annually between February and March.

Item Specifications

A general description of the Illinois student population being assessed by the IAA was used as context for item development purposes. The IAA students have, or function as if they have, the most 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 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 3, 2009, from <http://www.cehd.umn.edu/NCEO/OnlinePubs/Synthesis44.html>

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 maximizes 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 1.3.

Table 1.3: 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.

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/administrators to the same level of understanding. Training materials are developed and presented by Pearson in collaboration with ISBE via web-based sessions.

Test Implementation Manual

The *IAA Test Implementation Manual* was developed by Pearson and ISBE using input from best practices in 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 materials 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

Each IAA 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.

Student Score Sheets

The IAA Student Score Sheet has been developed by Pearson and ISBE to be user friendly, efficient means of data capture. The score sheet is located in the test booklet, the Implementation Manual and posted online. Teachers record the student's scores, accommodations listed on the IEP, and the accommodations used during testing on the score sheet and then transfer the scores and accommodations 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.
- Provide information about the receipt, verification and return of secure test materials.
- Demonstrate capabilities of the online scoring tool.

Training Logistics

- In December 2010 and January 2011, Pearson and ISBE staff hosted (3) webinar training sessions, which were attended by nearly 500 Illinois IAA Coordinators and educators.

Training Facilitators

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

Training Materials

- All materials in support of the IAA Trainings and spring 2011 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 2011 packaging and distribution requirements
- Training materials included a PowerPoint presentation, IAA rubric, student video clips, and IAA Student Score Sheet to acquaint participants with data fields that were required for the spring 2011 administration.
- Test administration resources included the IAA Frameworks, the 30-page *Test Implementation Manual*, *Online User Guides for Teachers, Coordinators and Scoring Monitors*, and sample items.

Bias Review

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

Differential Item Functioning

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 subgroups of students after matching on their ability. It is important to note that DIF might be the result of actual differences in relevant knowledge of individual item or statistical Type 1 error. As a result, DIF statistics should only be 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.

Any IAA pilot items that were flagged as showing DIF were subjected to further examination. For each of these items, the data review committee was asked to judge whether the differential difficulty of the item was unfairly related to group membership. If the differential difficulty of the item was considered to be related to group membership, and the difference was deemed unfair, then the item should not be used at all. Otherwise, the item should only be used if there is no other item matching the test blueprint.

DIF analyses for IAA were conducted between male and female, white and black, and white and Hispanics. Male and white are usually referred to as the reference group, and the others as the focal group. The Educational Testing Service (ETS) DIF procedure for polytomous items was adopted, which uses the Mantel Chi-square (Mantel χ^2) in conjunction with the standardized mean difference (SMD).

Mantel Statistic

The Mantel χ^2 is a conditional mean comparison of the ordered response categories for reference and focal groups combined over levels of the matching variable score. *Ordered* means that a response of “2” on an item is better than “1”, and “3” is better than “2.” *Conditional* refers to the comparison of members from the two groups who are matched on the total test score.

Table 1.4 shows a $2 \times T \times K$ contingency table, where T is the number of response categories and K is the number of levels of the matching variable. The values, y_1, y_2, \dots, y_T are the T scores that can be gained on the item. The values, n_{Fik} and n_{Rik} , represent the numbers of focal and reference groups who are at the k^{th} level of the matching variable and gain an item score of y_i . The “+” indicates total number over a particular index (Zwick, Donoghue, & Grima, 1993).

Table 1.4 $2 \times T$ Contingency Table at the k^{th} level

Group	Item Score				Total
	y_1	y_2	...	y_T	
Reference	n_{R1k}	n_{R2k}	...	n_{RTk}	n_{R+k}
Focal	n_{F1k}	n_{F2k}	...	n_{FTk}	n_{F+k}
Total	n_{+1k}	n_{+2k}	...	n_{+Tk}	n_{++k}

Note. This table was cited from Zwick, et al. (1993)

The Mantel statistic is defined as follows:

$$\text{Mantel } \chi^2 = \frac{(\sum_k F_k - \sum_k E(F_k))^2}{\sum_k \text{Var}(F_k)}$$

where

F_k = the sum of scores for the focal group at the k^{th} level of the matching variable and is defined as follows:

$$F_k = \sum_t y_t n_{Ftk}$$

The expectation of F_k under the null hypothesis is

$$E(F_k) = \frac{n_{F+k}}{n_{++k}} \sum_t y_t n_{+tk}$$

And, the variance of F_k under the null hypothesis is as follows:

$$Var(F_k) = \frac{n_{R+k} n_{F+k}}{n_{++k}^2 (n_{++k} - 1)} \left[(n_{++k} \sum_t y_t^2 n_{+tk}) - (\sum_t y_t n_{+tk})^2 \right].$$

Under the null hypothesis (H_0), the Mantel statistic has a chi-square distribution with one degree of freedom. In DIF applications, rejecting H_0 suggests that the students of the reference and focal groups who are similar in overall test performance tend to differ in their mean performance.

Standardized Mean Difference (SMD)

A summary statistic to accompany the Mantel approach is the standardized mean difference (SMD) between the reference and focal groups proposed by Dorans and Schmitt (1991). This statistic compares the means of the reference and focal groups, adjusting for differences in the distribution of the reference and focal group members across the levels of the matching variable.

$$SMD = \sum_k p_{Fk} m_{Fk} - \sum_k p_{Fk} m_{Rk}$$

where

$$p_{Fk} = \frac{n_{F+k}}{n_{F++}}, \text{ the proportion of the focal group members who are at the } k^{th}$$

level of the matching variable,

$$m_{Rk} = \frac{1}{n_{F+k}} \times (\sum_t y_t n_{Ftk}), \text{ the mean item score of the focal group members at}$$

the k^{th} level, and

$$m_{Rk} = \text{the analogous value for the reference group.}$$

As can be seen from the equation above, the SMD is the difference between the unweighted item mean of the focal group and the weighted item mean of the reference group. The weights for the reference group are applied to make the weighted number of the reference group students the same as in the focal group within the same ability. A negative SMD value implies that the focal group has a lower mean item score than the reference group, conditional on the matching variable.

DIF classification for IAA items

The SMD is divided by the total group item standard deviation (SD) to obtain an effect-size value for the SMD. This effect-size SMD is then examined in conjunction with the Mantel χ^2 to obtain DIF classifications that are depicted in Table 1.5 below.

Table 1.5 DIF Classifications for IAA Items

Category	Description	Criterion
A	No/Negligible DIF	Non-significant Mantel χ^2 or Significant Mantel χ^2 and $ SMD/SD \leq 0.17$
B	Moderate DIF	Significant Mantel χ^2 and $0.17 < SMD/SD \leq 0.25$
C	Large DIF	Significant Mantel χ^2 and $.25 < SMD/SD $

Note. SD is the total group standard deviation of the item score in its original metric.

Table 1.6 summarizes the number of items selected as 2011 census items by DIF category. Note that items from Category A are the first chosen for test construction. When items from Category A do not adequately fulfill the blueprint, items from Category B are selected. If the blueprint is still incomplete after exhausting the pool of Category B items, then items from Category C are considered, unless the differential difficulty of the item between the subgroups is determined in the data review to be unfair. For the 2011 administration none of the census items showed Category C DIF.

Table 1.6: DIF between Male/Female, White/Black, and White/Hispanic

Subject	Grade	Male/Female			White/Black			White/Hispanics		
		A	B	C	A	B	C	A	B	C
Reading	3	14	0	0	14	0	0	14	0	0
	4	14	0	0	13	1	0	14	0	0
	5	14	0	0	14	0	0	14	0	0
	6	14	0	0	13	1	0	14	0	0
	7	14	0	0	14	0	0	13	1	0
	8	14	0	0	14	0	0	14	0	0
	11	11	0	0	11	0	0	11	0	0
Mathematics	3	15	0	0	15	0	0	15	0	0
	4	15	0	0	15	0	0	15	0	0
	5	15	0	0	15	0	0	15	0	0
	6	15	0	0	15	0	0	15	0	0
	7	15	0	0	15	0	0	15	0	0
	8	15	0	0	15	0	0	15	0	0
	11	15	0	0	15	0	0	15	0	0
Science	4	15	0	0	14	1	0	15	0	0
	7	16	0	0	16	0	0	16	0	0
	11	15	0	0	15	0	0	15	0	0
Writing	11	7	0	0	7	0	0	7	0	0

Note. A = no or negligible DIF, B = moderate DIF, C = large DIF

2. RELIABILITY AND GENERALIZABILITY

The reliability of a test refers to its accuracy and the extent to which it yields consistent results across situations (Anastasi & Urbina, 1997). Classical test theory assumes that an observed score (X) consists of a student's true score (T) and some amount of error (E), as represented below:

$$X = T + E.$$

The difference between a student's observed test score and true score is measurement error. As reliability increases, the measurement error decreases, and the precision of the observed test score increases. The reliability of a test should always be taken into account when interpreting the observed test scores and differences between test scores obtained over multiple occasions. 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

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:

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

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 IAA tests administered in spring 2011. Included with the coefficient alpha in the table are the number of students responding to each 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. As presented in Tables 2.2a to 2.2c, the alpha coefficients by ethnicity, Limited English Proficiency (LEP), and income level are also high.

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 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:

$$\text{CTT SEM} = SD_x \sqrt{1 - r_{xx}}, \quad (2.2)$$

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

The SEM can be helpful in quantifying the extent of measurement 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	1,843	43.72	11.54	0.93	3.15
	4	1,907	44.71	11.41	0.93	2.99
	5	1,967	45.67	10.76	0.93	2.89
	6	1,896	46.09	11.01	0.94	2.69
	7	1,928	47.06	10.36	0.94	2.61
	8	1,879	47.45	10.11	0.93	2.66
	11	2,062	37.64	9.06	0.95	2.10
Mathematics	3	1,842	47.36	12.45	0.94	3.16
	4	1,903	49.76	11.82	0.94	2.91
	5	1,964	49.81	11.38	0.94	2.89
	6	1,893	49.79	11.71	0.94	2.78
	7	1,926	49.2	10.86	0.93	2.82
	8	1,877	50.61	11.31	0.94	2.69
	11	2,060	49.02	11.93	0.94	2.83
Science	4	1,903	47.93	12.26	0.94	3.06
	7	1,926	53.9	11.27	0.94	2.75
	11	2,057	50.56	12.68	0.96	2.43
Writing	11	2,057	23.45	6.00	0.92	1.74

Table 2.2a: Reliability Estimates by Ethnicity

Grade	Subgroup	Reading	Mathematics	Science	Writing
3	Asian	0.91	0.93	--	--
	Black	0.94	0.95	--	--
	Hispanic	0.93	0.94	--	--
	White	0.92	0.93	--	--
4	Asian	0.92	0.94	0.92	--
	Black	0.94	0.95	0.95	--
	Hispanic	0.93	0.95	0.94	--
	White	0.92	0.93	0.93	--
5	Asian	0.93	0.94	--	--
	Black	0.94	0.94	--	--
	Hispanic	0.93	0.94	--	--
	White	0.92	0.93	--	--
6	Asian	0.91	0.94	--	--
	Black	0.95	0.95	--	--
	Hispanic	0.95	0.95	--	--
	White	0.93	0.93	--	--
7	Asian	0.93	0.93	0.95	--
	Black	0.94	0.93	0.94	--
	Hispanic	0.92	0.93	0.93	--
	White	0.94	0.93	0.94	--
8	Asian	0.95	0.96	--	--
	Black	0.93	0.95	--	--
	Hispanic	0.93	0.95	--	--
	White	0.93	0.94	--	--
11	Asian	0.97	0.97	0.97	0.94
	Black	0.94	0.94	0.96	0.92
	Hispanic	0.96	0.96	0.97	0.93
	White	0.94	0.94	0.96	0.91

Table 2.2b: Reliability Estimates by LEP

Grade	Subgroup	Reading	Mathematics	Science	Writing
3	LEP	0.93	0.94	--	--
	Non-LEP	0.93	0.94	--	--
4	LEP	0.94	0.95	0.94	--
	Non-LEP	0.93	0.94	0.94	--
5	LEP	0.92	0.94	--	--
	Non-LEP	0.93	0.94	--	--
6	LEP	0.95	0.95	--	--
	Non-LEP	0.94	0.94	--	--
7	LEP	0.92	0.93	0.94	--
	Non-LEP	0.94	0.93	0.94	--
8	LEP	0.94	0.96	--	--
	Non-LEP	0.93	0.94	--	--
11	LEP	0.96	0.96	0.97	0.93
	Non-LEP	0.95	0.94	0.96	0.92

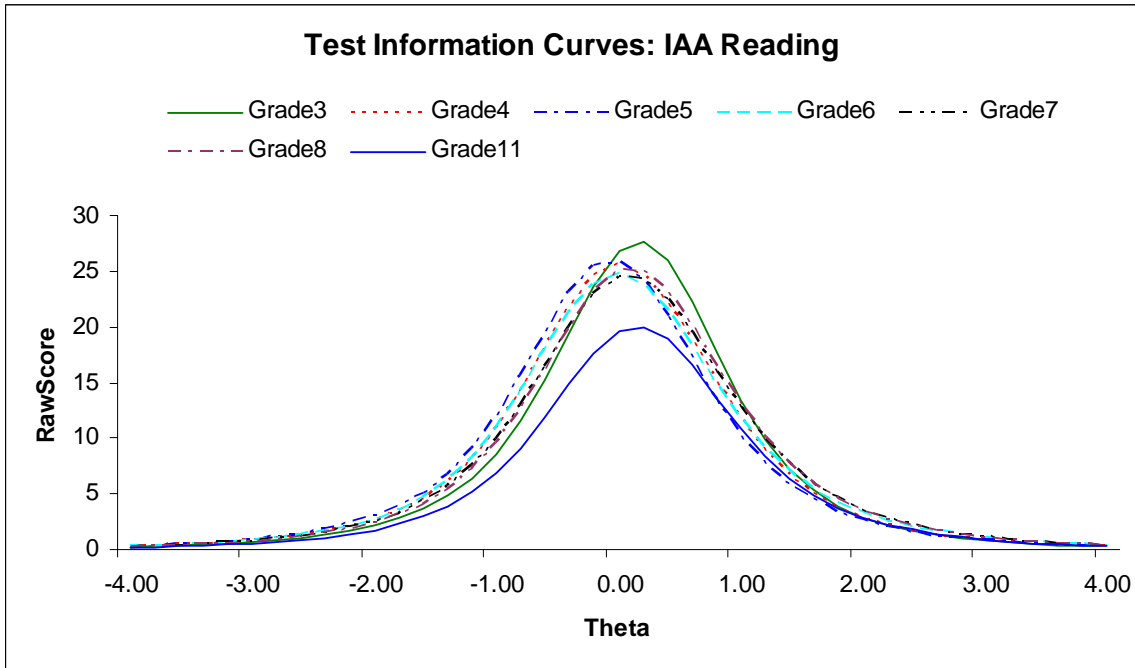
Table 2.2c: Reliability Estimates by Income

Grade	Subgroup	Reading	Mathematics	Science	Writing
3	Low-Income	0.93	0.94	--	--
	Non-Low-Income	0.92	0.93	--	--
4	Low-Income	0.93	0.94	0.94	--
	Non-Low-Income	0.93	0.94	0.93	--
5	Low-Income	0.93	0.94	--	--
	Non-Low-Income	0.92	0.93	--	--
6	Low-Income	0.94	0.94	--	--
	Non-Low-Income	0.94	0.94	--	--
7	Low-Income	0.92	0.92	0.93	--
	Non-Low-Income	0.94	0.94	0.95	--
8	Low-Income	0.94	0.95	--	--
	Non-Low-Income	0.92	0.94	--	--
11	Low-Income	0.94	0.94	0.96	0.91
	Non-Low-Income	0.95	0.95	0.96	0.92

IRT Test Information Function

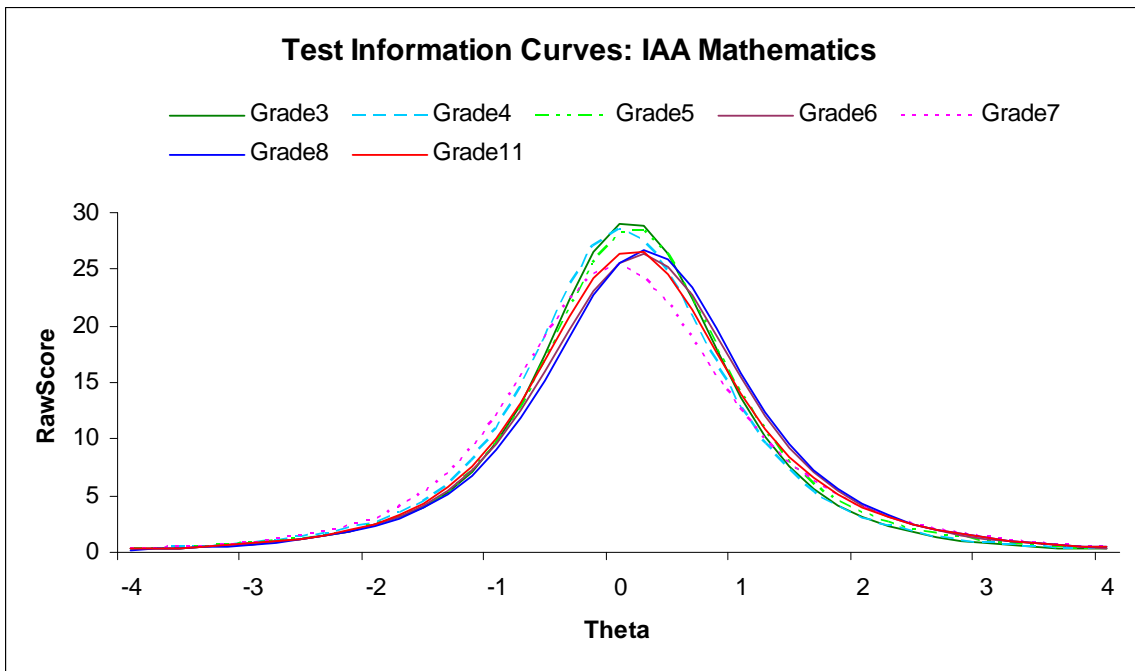
The reliability coefficients reported above were derived within the context of classical test theory and provide a single measure of precision for the entire test. With the Item Response Theory (IRT), it is possible to measure the relative precision of the test at different points on the scale. The amount 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



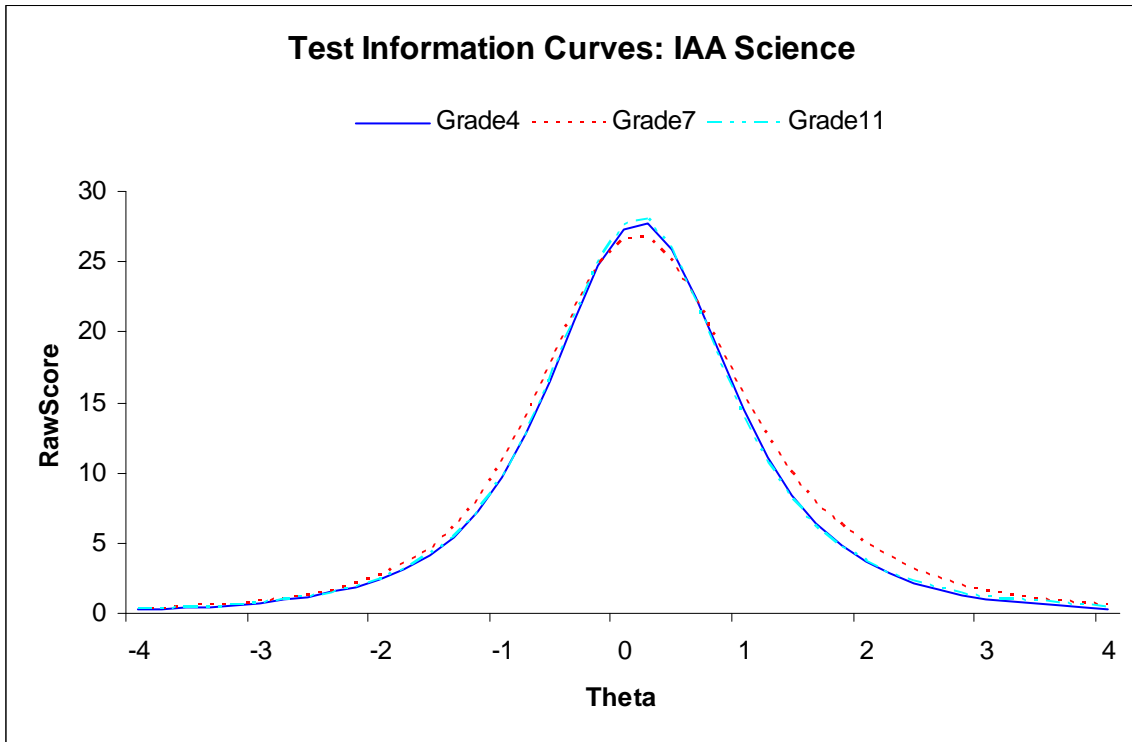
Note: Grades 3-8 have 14 items and grade 11 has 11 items.

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



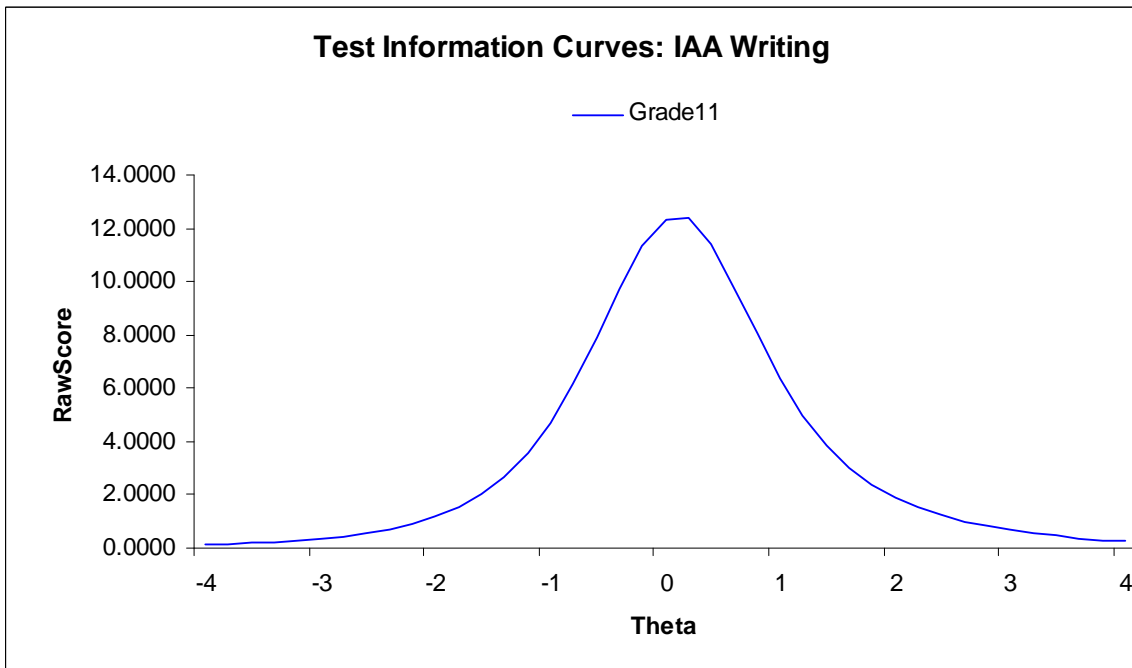
Note: Mathematics has 15 items for all grades.

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 Grade 11 Test Information Functions



Note: Writing has 7 items for grade 11.

IRT Conditional SEM

The standard error of measurement (SEM) reflects the degree of measurement 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:

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

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.

For the IAA, the SEM was first estimated on a theta scale by subject and grade. When reporting with scale scores, the SEM was transformed onto the IAA scale by applying a scaling slope (see Appendix B).

Classification Accuracy

Proficiency classification accuracy refers to the extent to which decisions based on test scores match decisions that would have been made if the scores did not contain any measurement error (Hambleton & Novick, 1973). Every test administration will result in some error in classifying examinees. The concept of the standard error of measurement (SEM) has an impact on how to explain the cut scores used to classify students into different achievement levels. For example, some students may have a true achievement level greater than a cut score. However, due to random variations (measurement error), the observed test score may be below the cut score. As a result, the students may be classified as having a lower achievement level. As discussed in the section on the SEM, a student's observed score is most likely to fall into a standard error band around his or her true score. Thus, the classification of students into different achievement levels can be imperfect, especially for the borderline students whose true scores lie close to achievement level cut scores.

For the IAA, there are four levels of achievement: Entry, Foundational, Satisfactory, and Mastery. An analysis of the consistency in classification is described below.

True level of achievement, which is based on the student's true score, cannot be observed, and therefore classification accuracy cannot be directly determined. It is possible, however, to estimate classification accuracy based on prediction from the Item Response Theory (IRT) model.

The method followed is based on the work of Rudner (2005). An assumption is made that for a given (true) ability score θ , the observed score $\hat{\theta}$ is normally distributed with a mean of θ and a standard deviation of $SE(\theta)$ (i.e., the CSEM at θ). Using this information, the expected proportion of students with true scores in any particular achievement level (bounded by cut scores c and d) who are classified into an performance level category (bounded by cut scores a and b) can be obtained by

$$P(\text{Level}_k) = \sum_{\theta=c}^d \left(\phi \left(\frac{b-\theta}{SE(\theta)} \right) - \phi \left(\frac{a-\theta}{SE(\theta)} \right) \right) f(\theta), \quad (2.4)$$

where a and b are theta scale points representing the score boundaries for the observed level, d and c are the theta scale points representing score boundaries for the true level, ϕ is the normal cumulative distribution function and $f(\theta)$ is the density function associated with the true score. For the IAA, the observed probability distribution of student theta estimates is used to estimate the $f(\theta)$ and to 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.

To compute classification consistency, the proportions are computed for all cells of a K by K classification table. The sum of the diagonal entries represents the decision consistency of classification for the test.

An example classification table is presented in Table 2.3. The rows represent the theoretical true score percentages of examinees in each performance level, while the columns represent the observed percentages. The R1 through R4 refer to the performance levels of Entry through Mastery respectively. The diagonal entries within the table represent the agreement between true and observed percentages of classified examinees. For example, 17.8 is the accuracy of Entry level and 21.4 is the accuracy of Foundational level. The sum of the diagonal values, 17.8, 23.4, 28.7, and 8.2, is the overall test classification accuracy (78.1). The overall test classification is presented in Table 2.4 by subject and grade. Classification accuracy tables, similar to Table 2.3, for all subjects and grades can be found in Appendix C.

Table 2.3: Reading Grade 3 Classification Accuracy

Level	R1	R2	R3	R4	True
R1	17.8	3.4	0.0	0.2	21.4
R2	2.0	23.4	6.9	0.5	32.7
R3	0.0	3.0	28.7	4.2	36.0
R4	0.0	0.0	1.6	8.2	9.8
Observed	19.8	29.8	37.2	13.1	100.0

Table 2.4: IAA Classification Accuracy

Grade	Reading	Mathematics	Science	Writing
3	78.1	75.3	--	--
4	76.9	77.9	77.4	--
5	73.4	78.3	--	--
6	76.5	78.6	--	--
7	76.9	77.3	78.5	--
8	74.6	77.3	--	--
11	72.6	79.5	79.6	71.5

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-related validity, construct validity, and criterion-related validity. Content-related 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 can be examined through analyses of a test's internal constructs that 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. In short, the construct 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-related 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 d) 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 standards, 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 scoring monitors 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 evidence are presented: content-related validity, construct validity, and criterion-related validity.

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. In addition, 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 suggests 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; Raïche, 2005). Before an approach is established to adequately deal with the complex data situations of IAA, simple and straightforward methods might provide some useful evidence for test dimensionality. In this study, the principal component analysis was chosen for its straightforward statistical model in comparison to factor analysis's latent variable approach. 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 was conducted to provide supporting evidence for unidimensionality.

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. The item level polychoric correlation matrix computed with SAS was subjected to PCA. Lord (1980) suggested 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 along with the percent of variance explained by each component. As can be seen, the first eigenvalues are considerably larger than the rest of eigenvalues. The percent of

variance explained by the first eigenvalue ranges from 61.0% to 78.0% across subjects and grades.

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. Graphical representations of the eigenvalues, known as scree plots, can be found in Appendix E for the IAA Reading, Mathematics, Science, and Writing assessments. The elbow shaped plots support the unidimensionality conclusion drawn from the Divgi index.

Table 3.1: Divgi Indices

Grade	Reading	Mathematics	Science	Writing
3	45.42	71.86	--	--
4	149.39	48.19	22.35	--
5	59.73	64.82	--	--
6	67.66	30.89	--	--
7	28.19	138.82	17.71	--
8	18.10	15.14	--	--
11	66.06	20.41	121.14	14.19

The IRT PCA was estimated through WINSTEPS. Interpretation of IRT PCA is different from the previously mentioned PCA in that the IRT PCA investigates residuals: the parts of observations not explained by the Rasch dimension. PCA of response residuals among items can reveal the presence of unexpected secondary dimensions contained in item content. Wright (1996) suggests that if a test is unidimensional, its response residuals should be random and show no structure. To claim unidimensionality, the percent of variance explained by the Rasch measures should be higher than that explained by the residuals. Table 3.2a presents the total variance as well as variance explained and unexplained by the Rasch measures. The ratios of the explained over unexplained variance are large across subjects and grades. Table 3.2b presents the first three residual components and the ratio between the variance explained by Rasch measures and the first three residual components. The variance explained by the three residual components is small for all subjects and grades. These results support the traditional PCA results in that the IAA tests are unidimensional.

Table 3.2a: IRT PCA Variances

Grade	Total Variance	Observed Explained Variance	Observed Unexplained Variance	% Explained Variance	% Unexplained Variance	Ratio Explained/Unexplained
Reading						
3	46.2	32.2	14	70	30	2.30
4	51.6	37.6	14	73	27	2.69
5	47.4	33.4	14	70	30	2.39
6	58.1	44.1	14	76	24	3.15
7	50.7	36.7	14	72	28	2.62
8	47.0	33.0	14	70	30	2.36
11	51.3	40.3	11	79	21	3.66
Mathematics						
3	60.9	45.9	15	75	25	3.06
4	54.0	39.0	15	72	28	2.60
5	58.1	43.1	15	74	26	2.87
6	66.6	51.6	15	77	23	3.44
7	54.2	39.2	15	72	28	2.61
8	53.0	38.0	15	72	28	2.53
11	58.2	43.2	15	72	26	2.88
Science						
4	49.7	34.7	15	70	30	2.31
7	63.0	47.0	16	75	25	2.94
11	72.3	57.3	15	79	21	3.82
Writing						
11	24.6	17.6	7	72	28	2.51

Table 3.2b: First Three IRT PCA Residual Components

Grade	Variance Explained by Residual Component			Ratio Explained Variance/ Residual Component		
	1	2	3	1	2	3
Reading						
3	1.5	1.3	1.2	24.77	26.83	2.30
4	1.4	1.4	1.2	26.86	31.33	2.69
5	1.5	1.3	1.2	25.69	27.83	2.39
6	1.4	1.4	1.2	31.50	36.75	3.15
7	1.5	1.2	1.2	30.58	30.58	2.62
8	1.7	1.3	1.2	25.38	27.50	2.36
11	1.5	1.2	1.1	33.58	36.64	3.66
Mathematics						
3	1.5	1.4	1.2	32.79	38.25	3.06
4	1.6	1.5	1.2	26.00	32.50	2.60
5	1.7	1.4	1.2	30.79	35.92	2.87
6	1.5	1.3	1.2	39.69	43.00	3.44
7	1.5	1.3	1.2	30.15	32.67	2.61
8	1.8	1.4	1.2	27.14	31.67	2.53
11	1.8	1.3	1.2	33.23	36.00	2.88
Science						
4	1.7	1.3	1.2	26.69	28.92	2.31
7	1.7	1.4	1.2	33.57	39.17	2.94
11	1.4	1.3	1.2	44.08	47.75	3.82
Writing						
11	1.7	1.2	1.1	14.67	16.00	2.51

Internal Construct

The purpose of examining 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 3.3 reports the correlation matrices among the IAA Reading, Mathematics, Science, and Writing assessments. The correlation between Reading and Mathematics scores ranges from .89 to .92; the correlation between Reading and Science scores also ranges from .90 to .92; the correlation between Reading and Writing is .90; the correlation between Mathematics and Writing is .88; and the correlation between Mathematics and Science is from .90 to .91.

In addition, item-total correlations were calculated to evaluate the test structure. The corrected item-total correlation, in contrast to the uncorrected method, excludes the item score from the total score when computing its item-total correlation. This method avoids the overestimation issue that commonly occurs in the uncorrected method. Table 3.4 presents the median of the corrected item-total correlations for each subject and grade. The median of the corrected item-total correlations ranges from 0.67 to 0.80 across subjects and grades.

Table 3.3: Correlation among IAA Assessments

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

Table 3.4: Median of Item-Total Correlations by Subject and Grade

Grade	Reading	Mathematics	Science	Writing
3	0.66	0.68	--	--
4	0.67	0.72	0.69	--
5	0.65	0.67	--	--
6	0.71	0.72	--	--
7	0.72	0.67	0.68	--
8	0.68	0.73	--	--
11	0.78	0.72	0.80	0.78

Criterion-related Validity

In order to examine the criterion-related validity of the IAA, a study was conducted in 2011 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 by Teacher" and "Student Score by Scoring Monitor" is presented as a validity coefficient " xy ." This validation approach is based on the premise that a score given to a student response 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 2011 IAA administration, eight scoring monitors were recruited by ISBE to provide secondary scores throughout the state of Illinois. All scoring 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 designs and IAA rubrics, 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 selected eight expert scorers who best met the criteria stated above. The monitors were instructed to base their student sample 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 2011 IAA operational test window. Coordination of 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 3.5. As can be seen from the table, the sample for the spring 2011 validation study has comparable percentages of male and female students with the spring 2011 IAA student population

Table 3.5: Spring 2011 IAA Student Population and Validation Sample

	<i>N</i>	Percentage	
		Male	Female
IAA Population	13,479	65.0	35.0
Validation Sample	151	70.9	29.1

Note: Students with missing gender information were not counted in the calculation of percentages.

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 agreement 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 agreement of scores on an item was defined as the extent to which the items were scored by both scorers with *exact agreement* or with one point of difference between the two scorers (i.e., *adjacent agreement*). Table 3.6 provides the average percentage of exact agreement, the average percentage of adjacent agreement, and the average percentage of total agreement (i.e., sum of the average percentages of exact and adjacent agreement) between the two scorers across items by subject and grade. The average number of students used to analyze each item was also presented by subject and grade. The results of these analyses suggest a high degree of agreement. The average percentage of exact agreement between teacher scores and scoring monitor scores exceeded 92% for all subjects and grades, and the average percentage of total agreement exceeded 95% for all subjects and grades. Appendix F presents the results of rater agreement on each item for students with complete pairs of ratings for Reading, Mathematics, Science, and Writing assessments.

Table 3.6: Average Agreement between Teacher and Expert Scores

Subject	Grade	Number of Items	Number of Students	% Exact Agreement	% Adjacent Agreement	% Total Agreement
Reading	3	14	12	94.7	4.1	98.9
	4	14	10	91.9	2.7	94.5
	5	14	19	97.1	2.5	99.7
	6	14	11	96.1	2.6	98.7
	7	14	15	95.7	4.3	100.0
	8	14	20	95.8	3.9	99.7
	11	11	23	97.2	2.8	100.0
Mathematics	3	15	13	96.7	3.3	100.0
	4	15	14	94.7	4.8	99.5
	5	15	11	99.4	0.6	100.0
	6	15	14	98.1	1.9	100.0
	7	15	20	96.5	2.9	99.4
	8	15	11	92.2	4.8	97.1
	11	15	22	97.3	1.8	99.1
Science	4	15	15	98.7	0.9	99.6
	7	16	18	95.2	4.4	99.6
	11	15	18	98.5	0.4	98.9
Writing	11	7	22	96.8	2.0	98.8

Correlations between Teacher Scores and Expert Scores

To examine evidence of criterion-related validity based on expert scores, the correlations were calculated between students' total scores based on teacher ratings and their total scores based on ratings from the scoring monitors. The correlation results for Reading, Mathematics, Science, and Writing are shown in Tables 3.7a through 3.7d respectively. Across subjects and grades, a strong positive association was found between the scores given by teachers and scoring monitors. The correlations exceeded 0.95 for all subjects and grades, and approached unity for most.

Table 3.7a: Correlations between Teacher and Expert Scores: Reading

Grade	Number of Students	Mean		Std Deviation		Minimum		Maximum		<i>r</i>
		Teacher	Expert	Teacher	Expert	Teacher	Expert	Teacher	Expert	
3	13	44.08	44.23	13.97	13.95	10	10	56	56	0.997
4	11	44.73	44.82	10.34	10.28	20	20	56	56	0.999
5	21	47.48	47.48	11.32	11.67	15	13	56	56	0.999
6	11	47.91	47.91	10.12	10.34	27	27	56	56	0.999
7	16	46.88	46.44	14.94	14.87	7	7	56	56	0.999
8	23	42.13	42.26	14.94	14.76	13	14	56	56	0.998
11	23	36.65	36.43	9.27	9.24	12	12	44	44	0.999

Table 3.7b: Correlations between Teacher and Expert Scores: Mathematics

Grade	Number of Students	Mean		Std Deviation		Minimum		Maximum		<i>r</i>
		Teacher	Expert	Teacher	Expert	Teacher	Expert	Teacher	Expert	
3	13	54.23	54.23	8.05	8.01	31	31	60	60	0.999
4	14	51.29	51.14	8.83	8.83	36	35	60	60	0.996
5	11	53.27	53.36	13.50	13.42	16	16	60	60	1.000
6	14	56.50	56.50	1.65	1.65	54	54	59	59	1.000
7	21	48.29	48.43	11.91	11.78	20	20	59	60	0.999
8	12	46.50	46.42	16.89	16.26	6	9	59	59	0.997
11	22	49.18	48.73	10.03	10.76	22	18	60	60	0.996

Table 3.7c: Correlations between Teacher and Expert Scores: Science

Grade	Number of Students	Mean		Std Deviation		Minimum		Maximum		<i>r</i>
		Teacher	Expert	Teacher	Expert	Teacher	Expert	Teacher	Expert	
4	16	48.69	48.81	16.32	16.24	8	8	60	60	0.999
7	18	59.11	59.33	3.80	3.48	52	54	64	64	0.953
11	18	50.06	50.28	12.80	12.45	19	19	60	60	0.998

Table 3.7d: Correlations between Teacher and Expert Scores: Writing

Grade	Number of Students	Mean		Std Deviation		Minimum		Maximum		<i>r</i>
		Teacher	Expert	Teacher	Expert	Teacher	Expert	Teacher	Expert	
11	23	22.17	22.17	6.52	6.80	7	7	28	28	0.994

Validity Related to Comparison to Typical Performance

To provide further evidence for the validity of the IAA scores, test administrators/teachers were asked to compare the student's performance on the test to his/her typical performance in the classroom. As shown below, on the IAA Student Score Sheet that was used to record student scores, the test administrator/teachers were also required to enter information about their familiarity with the student along with their comparison of student performance on the test and his/her typical classroom performance on similar tasks.

TEACHER FAMILIARITY WITH STUDENT PERFORMANCE - Teacher

Instructions:

Please indicate familiarity with student performance. This applies to the person administering the test.

Very Familiar <input type="checkbox"/>	Familiar <input type="checkbox"/>	Somewhat Familiar <input type="checkbox"/>	Not At All <input type="checkbox"/>
---	--------------------------------------	---	--

COMPARISON TO TYPICAL PERFORMANCE- Teacher Instructions:

How did the student perform on this test, compared to his/her typical classroom performance on similar tasks? (Please answer to the best of your ability.)

	READING	MATHEMATICS	SCIENCE	WRITING
Much better than average	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Better than average	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Average	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Worse than average	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Much worse than average	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Prefer not to answer	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Familiarity with Students

Table 3.8 presents the distribution of teachers' ratings of their familiarity with the students being assessed. In general, the teachers rated themselves *very familiar* with the great majority of the students, ranging from 76.1% to 84.7% across the grades and subjects. About 12.3% to 19.3% of the students were rated *familiar* by the teachers. *Somewhat familiar* and *not at all* made up only 2.6% to 6.2% of the students in total.

Table 3.8 Distribution of Teacher Familiarity with Students

Subject	Grade	% Very Familiar	% Familiar	% Somewhat Familiar	% Not at All	% Very Familiar + Familiar
Reading	3	82.4	13.8	3.0	0.9	96.2
	4	82.6	14.2	2.1	1.1	96.8
	5	81.5	14.4	3.2	0.9	95.9
	6	76.0	19.8	3.5	0.8	95.8
	7	81.0	14.6	3.8	0.6	95.6
	8	83.8	12.7	2.6	0.8	96.5
	11	75.2	18.3	4.3	2.2	93.5
Mathematics	3	82.4	13.8	3.0	0.9	96.2
	4	82.6	14.2	2.1	1.1	96.8
	5	81.5	14.4	3.2	0.9	95.9
	6	76.0	19.8	3.5	0.8	95.8
	7	81.0	14.6	3.8	0.6	95.6
	8	83.8	12.7	2.6	0.8	96.5
	11	75.2	18.3	4.3	2.2	93.5
Science	4	82.6	14.2	2.1	1.1	96.8
	7	81.0	14.6	3.8	0.6	95.6
	11	75.2	18.3	4.3	2.2	93.5
Writing	11	75.2	18.3	4.3	2.2	93.5

Comparison to Typical Performance

As mentioned above, data were also collected for each student on the teacher’s comparison of the student’s test performance with his/her typical classroom performance. For the purpose of analyzing typical performance, not all the ratings are valid depending on the degree the teacher is familiar with the student being assessed. At least a rating of *familiar* is deemed necessary to be included for the analysis. As show in Table 3.8, the ratings of *very familiar* and *familiar* add up to well over 90% (ranging from 93.5% to 96.8%), which provide a sufficient number of students for the analysis of typical performance.

Table 3.9 presents the percentage of students whose performance on the IAA was rated from *much better than average* to *much worse than average* and *prefer not to answer*. The results of these analyses suggest a relatively high degree of agreement between student’s performance on the test and in the classroom. More than half of the students (from 54.7% to 66.3% across grades and subjects) got a rating of *average*, which means their performance on the test was considered the same as their typical classroom performance on similar tasks. Another 19.2% to 26.8% of students were considered *better than average*, and 6.1% to 11.4% of the students were considered *much better than average*.

Table 3.9 Comparison between Student Performance on the IAA and Typical Classroom Performance

Subject	Grade	Much Better than Average	Better than average	Average	Worse than Average	Much Worse than Average	Prefer Not to Answer
Reading	3	9.2	24.2	56.3	5.0	1.3	4.1
	4	8.5	24.2	59.4	3.0	1.2	3.6
	5	6.1	25.7	61.1	2.8	1.0	3.3
	6	8.2	24.3	60.5	3.8	1.1	2.2
	7	9.5	26.0	58.4	2.9	0.7	2.5
	8	8.5	23.6	61.5	2.9	0.6	2.8
	11	7.6	23.7	62.2	2.3	0.8	3.3
Mathematics	3	10.8	24.6	54.7	4.9	1.1	3.9
	4	9.7	25.7	57.6	2.7	0.8	3.6
	5	9.6	26.2	56.9	3.2	0.9	3.1
	6	9.1	24.1	60.7	2.9	0.8	2.4
	7	9.3	22.9	60.8	3.5	0.6	2.9
	8	11.4	22.8	59.5	2.5	0.7	3.1
	11	6.3	19.2	66.3	4.1	0.6	3.5
Science	4	9.6	25.6	56.6	3.5	0.9	3.9
	7	10.6	26.8	56.1	2.7	0.6	3.1
	11	7.5	22.1	63.8	2.1	0.8	3.7
Writing	11	6.7	22.2	65.1	2.0	0.6	3.5

Above all, with around 60% of the students considered to have performed as well on the IAA as in the classroom, the results of the typical performance analysis provided supporting evidence for the validity of IAA scores.

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 that the scores resulting from different years and test forms can be used interchangeably, and student performance can be evaluated across years. The latter is an important aspect for assessing Adequate Yearly Progress (AYP) that is mandated by the NCLB. Calibration and equating produce 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 easier for students, teachers, and other stakeholders to remember and interpret.

Calibration

For the calibration of the IAA, the Rasch partial credit model (RPCM) was used because of its flexibility in accommodating a smaller sample size and for its ability to handle polytomous data. The IAA calibration and equating result in a one-to-one relationship between raw score (total number of items answered correctly), theta, 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:

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

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

The RPCM has two sets of 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 were conducted using the commercially available program WINSTEPS 3.60 (Linacre, 2006).

Stability Check

For 2011 IAA, the constrained calibration (anchor item difficulty and step difficulty parameters) was performed using WINSTEPS to bring the current year item/theta parameter estimates to the 2009 baseline scale. There are four linking items for each IAA test and the linking items were examined as follows:

- Items with absolute displacement values greater than 0.50 logit are flagged.
- The flagged items are presented to ISBE for review.

- If decision is made to remove the items, they are excluded from the anchoring set. Then a new calibration is conducted using remaining common items.

When evaluating the stability of common items, two factors need to be taken into account. IAA tests are generally short, ranging from 7 to 16 items. Each assessment has 4 linking items, which means eliminating any items from the linking set might undermine the linkage between the two administrations. Besides maintaining sufficient number of common items, the content representation of the common items should be evaluated as well. Given the brevity of the IAA tests, content coverage should be carefully considered when eliminating items. An item should be retained if its elimination would result in a non-representative common item set.

As a result of the stability check, no linking items were flagged for spring 2011 IAA, and therefore, all linking items were retained for estimation of the equating constant.

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 1.1 in Chapter 1 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 (slope and intercepts) were then computed for each subject and grade based on these guidelines (see Appendix G). 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.

Apply Scale Transformation Constants and Define Scale Score Cuts

The scale scores on 2011 IAA were obtained by applying the scale transformation constants derived in 2009 to the theta estimates from WINSTEPS outputs to the fourth decimal point. When determining the performance level, the thetas were compared to theta cut scores of four decimal points (see Appendix G). A performance

level is defined as at or above the cut point; in other words, any theta that is equal to or higher than the theta cut is categorized into the higher level.

The computed scale scores are rounded to the nearest integer. However, when a theta is below the cut yet its rounded scale score is equal to or above the scale score cut, the scale score should be adjusted downward to ensure it is below the cut. For example, a panel-recommended theta cut for the satisfactory level equals .63 and transforms to a scale score cut of 500. The actual test has attainable thetas of .62 and .65, both of which transform to a rounded scale score of 500. Then the .62 should be rounded down to 499 so that these students who have not reached the recommended theta cut of .63 will be placed into the lower performance level. The same procedure will be followed for the mastery and foundational levels.

The IAA scale scores range from 300 to 700 for all subject and grades. The Satisfactory cut is set at 500. The scale score (SS) and standard error of estimate (SE) are computed using the following equations:

$$\begin{aligned} \text{SS} &= \text{Theta} * \text{slope} + \text{intercept} \\ \text{SE} &= \text{Theta} * \text{slope} \end{aligned}$$

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.

The IAA equating and scaling were independently replicated and cross-checked for accuracy. Equating results were reviewed by a third-party research scientist and approved through manager process review prior to submission to the ISBE. A summary of item statistics can be found in Appendix H.

5. RESULTS

Performance Relative to Illinois Alternate Assessment Frameworks

Following a standards validation meeting in 2009, the cut scores for the Foundational, Satisfactory, and Mastery performance levels on IAA Mathematics, Reading, Science, and Writing tests were established on raw score and theta scale (see Appendix G for the theta cuts). Details on the standards validation procedures can be found in *Illinois Alternate Assessment 2009 Technical Manual*. The corresponding scale score range for the four performance levels for 2011 is presented in Tables 5.1a to 5.1d.

Table 5.1a: IAA Reading Scale Score Range by Performance Level

Performance Level	Grade 3		Grade 4		Grade 5		Grade 6		Grade 7		Grade 8		Grade 11	
	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	581	538	586	537	655	546	612	547	579	553	624	558	568

Table 5.1b: IAA Mathematics Scale Score Range by Performance Level

Performance Level	Grade 3		Grade 4		Grade 5		Grade 6		Grade 7		Grade 8		Grade 11	
	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	656	554	619	541	570	567	684	539	613	555	634	548	606

Table 5.1c: IAA Science Scale Score Range by Performance Level

Performance Level	Grade 4		Grade 7		Grade 11	
	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	633

Table 5.1d: IAA Writing Scale Score Range by Performance Level

Performance Level	Grade 11	
	Low	High
Entry	300	458
Foundational	459	499
Satisfactory	500	576
Mastery	577	641

Based on the scale score cuts presented above, IAA students were classified into four performance levels: Entry, Foundational, Satisfactory, and Mastery. The results for 2011 are presented in Tables 5.2a to 5.2d along with the results for 2009 and 2010. Note that the sum of percentages by subject and grade may not add up to 100% due to rounding.

Table 5.2a: Percentage of Students by Performance Level for Reading

Grade	Year	Entry	Foundational	Satisfactory	Mastery
3	2009	20	24	33	24
	2010	19	28	36	18
	2011	20	30	37	13
4	2009	21	20	35	24
	2010	18	22	41	20
	2011	21	25	38	16
5	2009	24	18	23	36
	2010	23	18	20	40
	2011	24	21	22	33
6	2009	14	18	36	32
	2010	11	21	32	37
	2011	14	21	34	31
7	2009	15	20	42	24
	2010	15	15	36	33
	2011	15	21	39	25
8	2009	18	14	37	32
	2010	16	14	38	32
	2011	15	15	35	34
11	2009	13	18	31	38
	2010	12	17	34	36
	2011	13	19	39	29

Table 5.2b: Percentage of Students by Performance Level for Mathematics

Grade	Year	Entry	Foundational	Satisfactory	Mastery
3	2009	22	17	35	25
	2010	19	20	31	30
	2011	23	20	29	28
4	2009	17	18	35	30
	2010	15	18	43	24
	2011	15	18	50	17
5	2009	16	20	41	23
	2010	12	22	43	23
	2011	13	23	44	20
6	2009	14	15	33	38
	2010	11	14	31	44
	2011	12	15	36	37
7	2009	16	15	41	29
	2010	14	13	40	32
	2011	15	12	46	27
8	2009	12	19	37	31
	2010	9	18	40	33
	2011	11	18	37	34
11	2009	16	14	43	26
	2010	14	13	45	28
	2011	14	14	50	22

Table 5.2c: Percentage of Students by Performance Level for Science

Grade	Year	Entry	Foundational	Satisfactory	Mastery
4	2009	15	18	26	41
	2010	12	18	30	41
	2011	13	19	28	41
7	2009	11	17	29	43
	2010	10	13	28	49
	2011	8	16	29	47
11	2009	12	13	28	47
	2010	11	12	30	47
	2011	13	12	25	50

Table 5.2d: Percentage of Students by Performance Level for Writing

Grade	Year	Entry	Foundational	Satisfactory	Mastery
11	2009	12	11	26	50
	2010	11	10	28	51
	2011	13	10	32	44

REFERENCES

- Abedi, J. (1997). Dimensionality of NAEP subscale scores in mathematics (CSE Technical Report 428). Retrieved July 26, 2009, from <http://www.cse.ucla.edu/products/Reports/TECH428.pdf>
- Anastasi, A., & Urbina S. (1997). *Psychological testing* (7th ed). New Jersey: Prentice Hall.
- Clark, L. A., & Watson, D. (1995). Constructing validity: basic issues in objective scale development. *Psychological Assessment*, 7(3), 309-319.
- Crocker, L. M., & Algina, J. (1986). *Introduction to classical & modern test theory*. Orlando, FL: Harcourt Brace Jovanovich.
- Cronbach, L. J., & Meehl, P. E. (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.
- Dorans, N. J., & Schmitt, A. P. (1991). *Constructed Response and Differential Item Functioning: A Pragmatic Approach*. (ETS Research Report 91-47.) Princeton, NJ: Educational Testing Service.
- Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 105-146). New York: Macmillan.
- Hattie, J. (1985). Methodology review: assessing unidimensionality of tests and items. *Applied Psychological Measurement*, 9(2), 139-164.
- Hambleton, R.K., & Novick, M. R. (1973). Toward an integration of theory and method for criterion- referenced tests. *Journal of Educational Measurement*, 10(3), 159--170.
- Illinois Alternate Assessment 2008 Technical Manual. (2008). Pearson. http://www.isbe.net/assessment/pdfs/IAA_Tech_Manual_08.pdf
- Illinois Alternate Assessment 2009 Technical Manual. (2009). Pearson. http://www.isbe.net/assessment/pdfs/IAA_Tech_Manual_09.pdf
- Individuals with Disabilities Education Act (1990). 20 U.S.C. § 1400 et seq (P.L. 101-476). (Amended in 1997, 2004).
- Linacre, J. M. (2006). *WINSTEPS: Rasch measurement* (Version 3.60) [Computer Software]. Chicago, IL: WINSTEPS.
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. New York: Erlbaum.

- 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).
- Raïche, G. (2005). Critical eigenvalue sizes in standardized residual principal components analysis. *Rasch Measurement Transactions*, *19*(1), 1012.
- 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). Retrieved July 26, 2009, from <http://pareonline.net/getvn.asp?v=10&n=13>
- Suen, H. K. (1990). *Principles of test theories*. Hillsdale, NJ: Earlbaum.
- Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). Boston: Pearson Education.
- Tennant, A., & Pallant, J.F. (2006) Unidimensionality matters! (A tale of two Smiths?). *Rasch Measurement Transactions*, *20*(1), 1048-51.
- Thissen, D., & Wainer, H. (2001). *Test scoring*. Hillsdale, NJ: Earlbaum.
- US Department of Education (2005). *Alternate achievement standards for students with the most significant cognitive disabilities: Non-regulatory guidance*. Retrieved July 26, 2009, from <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.
- Zwick, R., Donoghue, J. R., & Grima, A. (1993). Assessment of differential item functioning for performance tasks. *Journal of Educational Measurement*, *30*(3), 233-251

APPENDIX A: IAA Scoring Rubric
Illinois Performance-based Task Assessment
 IAA Performance Rubric

<i>Level 4:</i>	<i>Level 3:</i>	<i>Level 2:</i>	<i>Level 1:</i>
<p>The student correctly performs the task without assistance or with a single repetition of instructions or refocusing.</p> <ul style="list-style-type: none"> • 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. <p><i>The student then responds correctly.</i></p>	<p>The student correctly performs the task with a general prompt.</p> <ul style="list-style-type: none"> • If the student responds incorrectly to the task at Level 4 when given adequate wait time, the teacher directs the student to the section of the test booklet containing additional information or a prompt about the expected response from the student such as: <ul style="list-style-type: none"> 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. <p><i>The student then responds correctly.</i></p>	<p>The student correctly performs the task with specific prompts.</p> <ul style="list-style-type: none"> • 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: <ul style="list-style-type: none"> 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. <p><i>The student responds correctly after being given the correct answer.</i></p>	<p>The student does not perform the task at Level 2 or provides an incorrect response despite Level 2 support.</p> <p><i>The student does not respond or does not respond correctly. Teacher demonstrates response and moves on to the next task.</i></p>

Illinois State Board of Education has adapted this rubric from the Colorado Student Assessment Program Alternate Level of Independence Performance Rubric. ISBE October 31, 2008.

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	370	47
12	300	53	300	54	300	97	300	64	300	46	300	72	401	26
13	300	53	300	54	300	97	300	64	300	46	300	72	419	18
14	361	53	353	54	300	97	321	64	369	46	304	72	428	14
15	395	29	389	29	300	53	364	35	400	25	351	39	435	12
16	414	20	408	20	331	37	388	25	417	18	378	27	441	11
17	425	16	419	16	351	30	401	20	427	14	394	22	445	10
18	432	13	427	14	365	25	411	17	434	12	404	19	448	9
19	437	12	433	12	376	23	418	15	439	11	412	17	452	9
20	442	11	437	11	385	20	424	14	444	10	419	15	454	8
21	446	10	441	10	392	19	429	13	448	9	425	14	457	8
22	449	9	445	10	398	18	433	12	451	9	429	13	459	8
23	452	9	448	9	404	17	437	11	454	8	434	13	462	8
24	454	8	450	9	409	16	441	11	456	8	438	12	464	7
25	457	8	453	8	413	15	444	10	459	8	441	12	466	7
26	459	8	455	8	418	15	447	10	461	7	444	11	468	7
27	461	8	457	8	422	14	450	10	463	7	447	11	470	7
28	463	7	460	8	426	14	452	10	465	7	450	11	472	7
29	465	7	462	8	429	14	455	9	467	7	453	10	474	7
30	466	7	463	8	433	14	457	9	468	7	456	10	476	7
31	468	7	465	7	436	13	460	9	470	7	458	10	479	7
32	470	7	467	7	440	13	462	9	472	7	461	10	481	8
33	471	7	469	7	443	13	465	9	474	6	464	10	483	8
34	473	7	471	7	446	13	467	9	475	6	466	10	485	8
35	475	7	473	7	449	13	469	9	477	6	468	10	488	8
36	476	7	474	7	453	13	471	9	478	6	471	10	491	9
37	478	7	476	7	456	13	474	9	480	6	473	10	494	9
38	480	7	478	7	459	13	476	9	482	7	476	10	498	10
39	481	7	480	7	462	13	478	9	483	7	478	10	502	11
40	483	7	482	8	466	14	481	9	485	7	481	10	507	12
41	485	7	484	8	469	14	483	9	487	7	484	10	513	14
42	487	7	486	8	473	14	486	10	489	7	487	11	523	17
43	489	8	488	8	477	14	489	10	491	7	489	11	539	25
44	491	8	490	8	481	15	491	10	493	7	493	11	568	46
45	493	8	493	8	485	15	494	10	495	7	496	11		
46	495	8	495	9	489	16	498	11	497	8	499	12		
47	497	9	498	9	494	16	501	11	499	8	503	12		
48	500	9	501	10	499	17	505	12	502	8	507	13		
49	503	10	504	10	506	18	509	13	505	9	512	14		
50	506	10	508	11	512	20	514	14	509	10	517	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	511	11	512	12	520	22	520	15	513	11	523	16		
52	516	13	518	13	530	24	527	17	518	12	531	18		
53	522	15	524	16	543	28	536	19	524	14	540	21		
54	532	19	535	19	562	35	548	24	534	17	554	26		
55	549	27	552	28	594	51	571	34	550	24	579	38		
56	581	52	586	53	655	95	612	64	579	46	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	308	73	394	39	300	98	340	57	300	74	350	55
16	319	51	356	39	420	21	303	54	378	31	344	41	386	30
17	353	35	382	27	435	15	340	38	400	22	373	29	407	21
18	372	28	397	22	443	12	361	30	413	18	389	23	419	17
19	385	24	406	18	448	10	376	26	422	16	400	20	428	15
20	395	21	414	16	453	9	387	23	429	14	409	18	434	13
21	402	19	420	15	456	8	396	21	434	13	416	16	439	12
22	409	18	425	14	459	8	404	19	439	12	422	15	444	11
23	414	16	429	13	461	7	410	18	443	11	427	14	448	10
24	419	16	433	12	464	7	416	17	447	10	432	13	451	10
25	424	15	436	11	465	6	422	16	450	10	436	13	454	9
26	428	14	440	11	467	6	426	16	453	10	439	12	457	9
27	431	14	442	11	469	6	431	15	456	9	443	12	459	9
28	435	13	445	10	470	6	435	15	459	9	446	11	462	8
29	438	13	448	10	472	6	439	14	461	9	449	11	464	8
30	441	13	450	10	473	5	443	14	464	9	452	11	466	8
31	444	12	453	10	475	5	447	14	466	8	455	10	468	8
32	447	12	455	10	476	5	450	14	468	8	457	10	470	8
33	450	12	457	9	477	5	454	13	470	8	460	10	472	8
34	453	12	460	9	478	5	457	13	472	8	461	10	474	7
35	456	12	462	9	480	5	460	13	475	8	465	10	476	7
36	459	12	464	9	480	5	463	13	477	8	467	10	477	7
37	461	12	466	9	482	5	467	13	479	8	470	10	479	7
38	464	12	468	9	483	5	470	13	480	8	472	10	481	7
39	467	12	470	9	484	5	473	13	483	8	474	10	483	7
40	470	12	473	9	486	5	476	13	485	8	477	10	485	7
41	472	12	475	10	487	5	480	13	487	8	479	10	487	7
42	475	12	477	10	488	5	483	13	489	8	482	10	489	8
43	478	12	480	10	489	5	486	14	491	8	484	10	490	8
44	481	13	482	10	491	5	490	14	493	8	487	10	492	8
45	484	13	484	10	492	5	493	14	496	9	489	10	495	8
46	488	13	487	10	493	6	497	14	498	9	492	11	497	8
47	491	13	490	11	495	6	501	15	500	9	495	11	499	8
48	495	14	493	11	496	6	505	15	503	9	498	11	501	9
49	498	14	496	11	498	6	509	15	506	9	501	12	504	9
50	502	15	499	12	499	6	514	16	509	10	505	12	507	9
51	507	16	502	12	502	7	519	17	512	10	509	13	509	10
52	512	16	506	13	504	7	524	18	516	11	513	13	513	10
53	517	17	511	14	506	7	531	19	520	12	517	14	516	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	523	19	516	15	509	8	538	20	524	12	523	15	521	12
55	531	20	521	16	512	9	546	22	529	14	529	17	526	13
56	540	23	528	18	516	10	556	25	536	15	537	19	532	14
57	552	27	538	21	521	11	569	29	544	17	547	22	540	17
58	568	33	551	26	529	14	588	36	556	21	562	27	551	21
59	599	49	575	38	542	21	622	52	577	31	587	39	571	29
60	656	92	619	71	567	39	684	97	613	57	634	73	606	54

Science

Raw Score	Grade 4		Grade 7		Grade 11	
	Scale Score	SE	Scale Score	SE	Scale Score	SE
1	300	133	300	128	300	67
2	300	133	300	128	300	67
3	300	133	300	128	300	67
4	300	133	300	128	300	67
5	300	133	300	128	300	67
6	300	133	300	128	300	67
7	300	133	300	128	300	67
8	300	133	300	128	300	67
9	300	133	300	128	300	67
10	300	133	300	128	300	67
11	300	133	300	128	300	67
12	300	133	300	128	300	67
13	300	133	300	128	300	67
14	300	133	300	128	300	67
15	300	133	300	128	323	67
16	300	73	300	128	367	37
17	300	51	300	70	392	26
18	318	41	300	49	407	21
19	338	35	300	40	417	18
20	353	31	319	35	424	16
21	365	28	334	31	430	14
22	375	26	347	28	436	13
23	384	24	357	26	440	12
24	392	23	366	24	444	12
25	399	22	374	23	448	11
26	405	21	381	22	451	11
27	411	20	388	21	454	10
28	416	20	394	20	457	10
29	422	19	400	20	459	10
30	426	19	405	19	462	9
31	431	18	410	19	464	9
32	435	18	415	18	466	9
33	440	18	420	18	468	9
34	444	18	424	18	471	9
35	449	18	429	18	473	9
36	453	17	433	18	475	9
37	457	17	437	17	477	9
38	461	17	442	17	479	9
39	465	17	446	17	481	9
40	469	17	450	17	483	9
41	474	18	455	17	485	9
42	478	18	459	17	488	9
43	482	18	463	18	490	9
44	487	18	468	18	492	9
45	491	19	472	18	495	9
46	496	19	477	18	497	10
47	501	19	481	18	499	10

Raw Score	Grade 4		Grade 7		Grade 11	
	Scale Score	SE	Scale Score	SE	Scale Score	SE
48	507	20	486	19	502	10
49	512	21	491	19	505	11
50	518	21	497	19	509	11
51	525	22	502	20	512	12
52	532	24	508	21	516	12
53	540	25	514	21	520	13
54	549	27	521	22	525	14
55	560	30	528	23	532	16
56	574	33	536	24	539	18
57	591	39	545	26	549	21
58	616	48	556	28	563	26
59	661	70	568	30	589	37
60	700	131	582	34	633	67
61			601	39		
62			628	48		
63			675	69		
64			700	127		

Writing

Raw Score	Grade 11	
	Scale Score	SE
1	300	85
2	300	85
3	300	85
4	300	85
5	300	85
6	300	85
7	303	85
8	359	46
9	391	33
10	410	27
11	423	23
12	433	21
13	442	19
14	449	18
15	456	17
16	463	17
17	469	17
18	475	17
19	481	17
20	487	17
21	494	18
22	502	19
23	510	21
24	521	23
25	534	26
26	552	33
27	585	47
28	641	85

APPENDIX C: Classification Consistency

Reading

R3

Level	R1	R2	R3	R4	True
R1	17.8	3.4	0.0	0.2	21.4
R2	2.0	23.4	6.9	0.5	32.7
R3	0.0	3.0	28.7	4.2	36.0
R4	0.0	0.0	1.6	8.2	9.8
Observed	19.8	29.8	37.2	13.1	100.0

R7

Level	R1	R2	R3	R4	True
R1	13.7	2.4	0.1	0.2	16.4
R2	1.3	15.7	6.1	0.7	23.8
R3	0.0	2.5	31.4	8.0	42.0
R4	0.0	0.0	1.7	16.1	17.9
Observed	15.0	20.6	39.3	25.1	100.0

R4

Level	R1	R2	R3	R4	True
R1	19.1	3.5	0.1	0.2	22.9
R2	1.7	17.9	6.4	0.5	26.6
R3	0.0	3.2	30.1	5.5	38.8
R4	0.0	0.0	1.7	9.7	11.4
Observed	20.9	24.6	38.3	15.9	100.0

R8

Level	R1	R2	R3	R4	True
R1	14.1	2.8	0.2	0.3	17.4
R2	1.3	9.7	5.6	0.7	17.4
R3	0.0	2.4	26.7	9.1	38.2
R4	0.0	0.0	2.9	24.1	27.0
Observed	15.4	14.8	35.5	34.2	100.0

R5

Level	R1	R2	R3	R4	True
R1	21.7	4.3	0.3	0.4	26.7
R2	1.9	14.2	6.2	1.5	23.9
R3	0.0	2.7	12.2	5.9	20.8
R4	0.0	0.0	3.2	25.3	28.5
Observed	23.7	21.3	21.8	33.1	100.0

R11

Level	R1	R2	R3	R4	True
R1	11.5	1.6	0.0	0.4	13.6
R2	1.0	15.8	6.8	1.7	25.4
R3	0.0	1.7	30.8	12.3	44.7
R4	0.0	0.0	1.6	14.5	16.1
Observed	12.5	19.1	39.3	29.0	100.0

R6

Level	R1	R2	R3	R4	True
R1	12.6	1.9	0.0	0.1	14.6
R2	1.2	16.7	6.0	0.7	24.6
R3	0.0	2.4	25.9	8.6	36.9
R4	0.0	0.0	2.6	21.3	23.9
Observed	13.8	20.9	34.5	30.7	100.0

Mathematics

M3

Level	M1	M2	M3	M4	True
M1	21.3	4.0	0.2	0.2	25.8
M2	1.8	13.1	5.4	0.7	20.9
M3	0.0	2.9	20.6	6.4	29.9
M4	0.0	0.0	3.1	20.4	23.4
Observed	23.1	20.0	29.2	27.7	100.0

M7

Level	M1	M2	M3	M4	True
M1	13.9	2.1	0.1	0.0	16.2
M2	1.2	7.6	4.8	0.2	13.8
M3	0.0	2.1	36.9	8.4	47.5
M4	0.0	0.0	3.7	18.9	22.6
Observed	15.1	11.8	45.6	27.5	100.0

M4

Level	M1	M2	M3	M4	True
M1	13.5	2.3	0.1	0.2	16.1
M2	1.1	13.8	7.2	0.5	22.6
M3	0.0	2.4	40.5	5.9	48.8
M4	0.0	0.0	2.3	10.0	12.3
Observed	14.7	18.5	50.1	16.6	100.0

M8

Level	M1	M2	M3	M4	True
M1	9.7	1.9	0.0	0.1	11.7
M2	0.9	13.6	4.4	0.6	19.5
M3	0.0	2.0	29.9	9.7	41.6
M4	0.0	0.0	3.1	24.0	27.1
Observed	10.5	17.5	37.5	34.4	100.0

M5

Level	M1	M2	M3	M4	True
M1	12.2	2.1	0.0	0.2	14.5
M2	1.1	18.1	6.4	0.6	26.1
M3	0.0	2.5	35.2	6.9	44.6
M4	0.0	0.0	1.8	12.8	14.7
Observed	13.3	22.7	43.5	20.4	100.0

M11

Level	M1	M2	M3	M4	True
M1	13.2	1.9	0.0	0.1	15.1
M2	0.8	10.0	5.7	0.2	16.8
M3	0.0	1.8	41.1	7.0	49.9
M4	0.0	0.0	2.8	15.3	18.1
Observed	14.0	13.6	49.7	22.5	100.0

M6

Level	M1	M2	M3	M4	True
M1	10.8	1.6	0.0	0.1	12.5
M2	0.9	11.0	4.0	0.4	16.3
M3	0.0	1.9	28.9	8.9	39.8
M4	0.0	0.0	3.5	27.9	31.4
Observed	11.7	14.5	36.4	37.3	100.0

Science

S4

Level	S1	S2	S3	S4	True
S1	11.7	1.9	0.0	0.1	13.6
S2	1.0	14.6	5.0	0.6	21.1
S3	0.0	2.5	19.3	8.0	29.7
S4	0.0	0.0	3.4	31.9	35.3
Observed	12.7	19.0	27.6	40.5	100.0

S7

Level	S1	S2	S3	S4	True
S1	7.7	1.4	0.0	0.0	9.2
S2	0.5	12.5	4.5	0.4	17.9
S3	0.0	1.8	20.7	9.2	31.6
S4	0.0	0.0	3.7	37.6	41.3
Observed	8.2	15.6	28.9	47.3	100.0

S11

Level	S1	S2	S3	S4	True
S1	11.7	1.2	0.0	0.1	13.1
S2	0.8	9.6	3.1	0.6	14.1
S3	0.0	1.4	19.6	10.3	31.3
S4	0.0	0.0	2.6	38.7	41.3
Observed	12.5	12.3	25.3	49.7	100.0

Writing

W11

Level	W1	W2	W3	W4	True
W1	12.0	1.9	0.2	0.5	14.6
W2	1.2	6.6	5.3	1.5	14.6
W3	0.1	1.6	24.3	13.9	39.8
W4	0.0	0.0	2.2	28.6	30.8
Observed	13.3	10.1	32.0	44.5	100.0

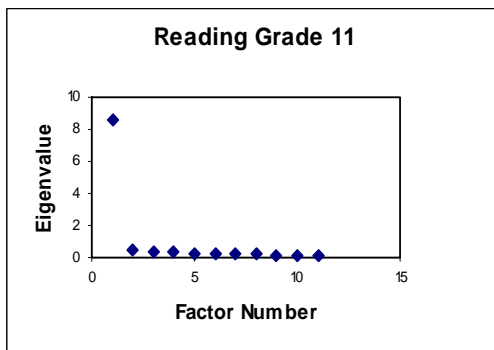
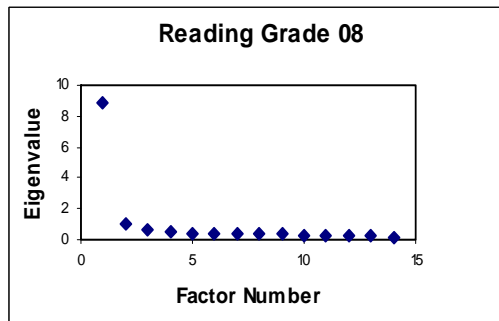
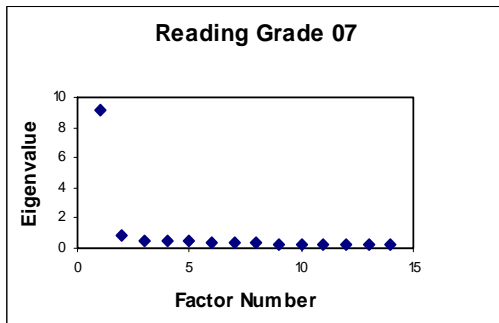
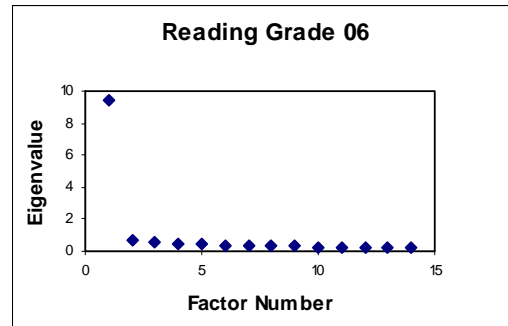
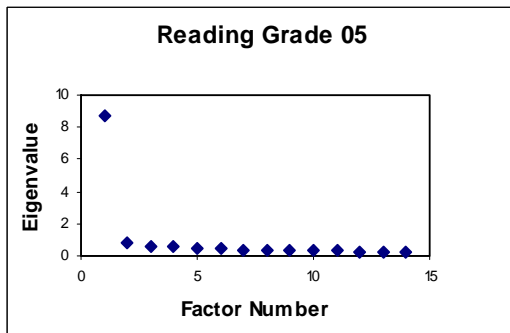
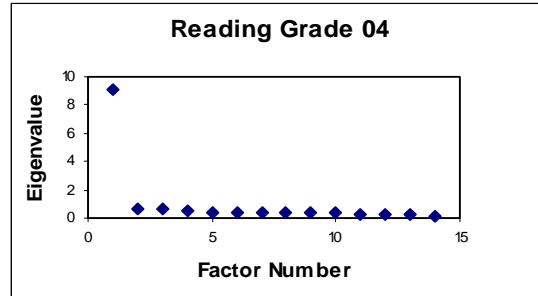
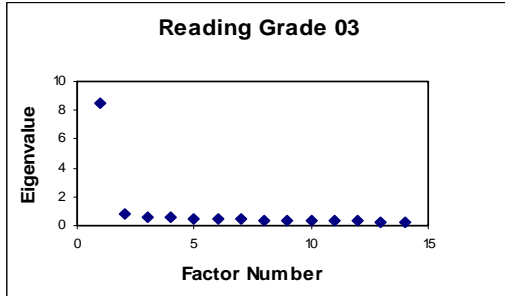
APPENDIX D: First Ten Eigenvalues from the Principal Component Analysis

Grade	Number	Reading		Mathematics		Science		Writing	
		Eigenvalue	% Variance Explained	Eigenvalue	% Variance Explained	Eigenvalue	% Variance Explained	Eigenvalue	% Variance Explained
3	1	8.546	61.0	9.766	65.1				
	2	0.776	5.5	0.751	5.0				
	3	0.605	4.3	0.626	4.2				
	4	0.545	3.9	0.478	3.2				
	5	0.491	3.5	0.445	3.0				
	6	0.444	3.2	0.414	2.8				
	7	0.412	2.9	0.384	2.6				
	8	0.397	2.8	0.368	2.5				
	9	0.364	2.6	0.312	2.1				
	10	0.345	2.5	0.291	1.9				
4	1	9.029	64.5	10.158	67.7	9.549	63.7		
	2	0.660	4.7	0.844	5.6	0.915	6.1		
	3	0.604	4.3	0.651	4.3	0.528	3.5		
	4	0.518	3.7	0.419	2.8	0.486	3.2		
	5	0.436	3.1	0.401	2.7	0.452	3.0		
	6	0.407	2.9	0.391	2.6	0.418	2.8		
	7	0.384	2.7	0.348	2.3	0.398	2.7		
	8	0.349	2.5	0.320	2.1	0.354	2.4		
	9	0.331	2.4	0.288	1.9	0.349	2.3		
	10	0.327	2.3	0.269	1.8	0.313	2.1		
5	1	8.688	62.1	9.685	64.6				
	2	0.763	5.5	0.815	5.4				
	3	0.631	4.5	0.678	4.5				
	4	0.528	3.8	0.494	3.3				
	5	0.461	3.3	0.439	2.9				
	6	0.420	3.0	0.396	2.6				
	7	0.390	2.8	0.384	2.6				
	8	0.372	2.7	0.344	2.3				
	9	0.352	2.5	0.327	2.2				
	10	0.332	2.4	0.308	2.1				
6	1	9.385	67.0	10.164	67.8				
	2	0.670	4.8	0.816	5.4				
	3	0.542	3.9	0.513	3.4				
	4	0.428	3.1	0.451	3.0				
	5	0.404	2.9	0.413	2.8				
	6	0.357	2.6	0.363	2.4				
	7	0.353	2.5	0.347	2.3				
	8	0.327	2.3	0.299	2.0				
	9	0.316	2.3	0.288	1.9				
	10	0.271	1.9	0.271	1.8				
7	1	9.198	65.7	9.148	61.0	10.161	63.5		
	2	0.802	5.7	0.724	4.8	1.056	6.6		
	3	0.504	3.6	0.664	4.4	0.542	3.4		

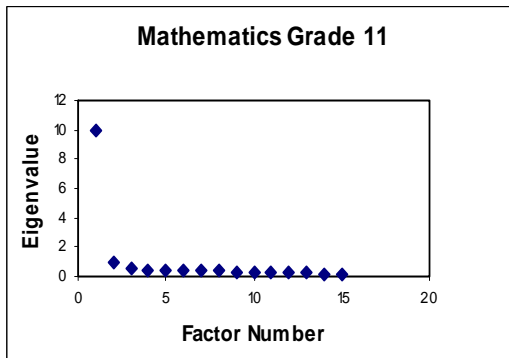
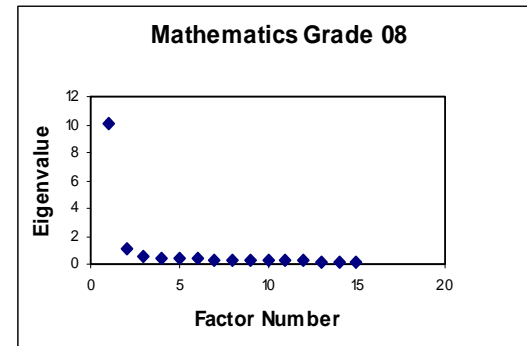
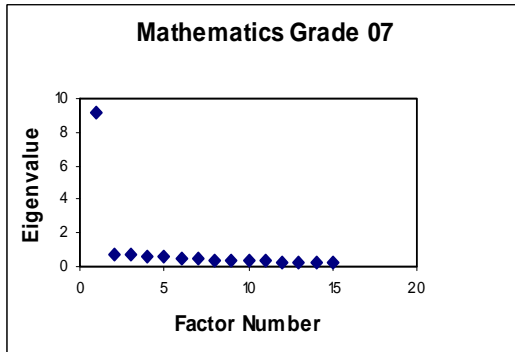
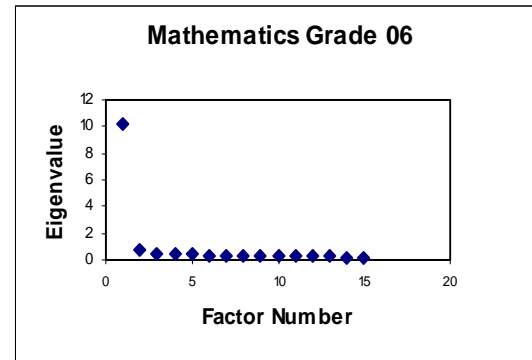
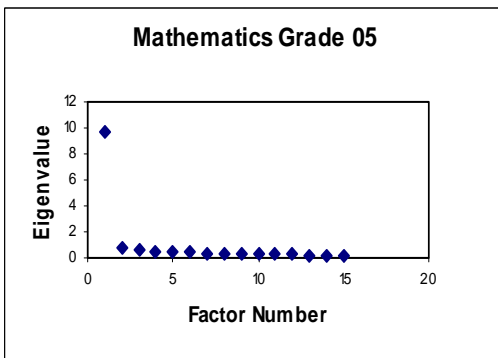
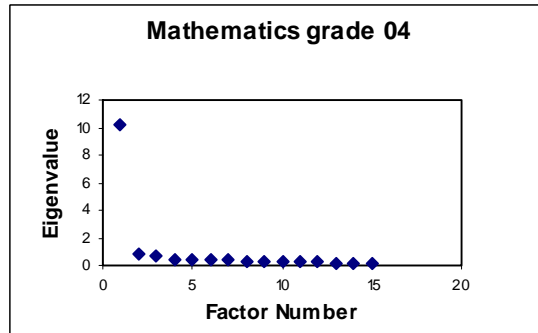
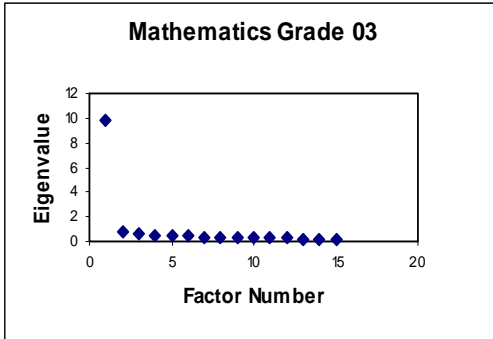
Grade	Number	Reading		Mathematics		Science		Writing		
		Eigenvalue	% Variance Explained	Eigenvalue	% Variance Explained	Eigenvalue	% Variance Explained	Eigenvalue	% Variance Explained	
	4	0.454	3.2	0.586	3.9	0.510	3.2			
	5	0.424	3.0	0.545	3.6	0.469	2.9			
	6	0.397	2.8	0.462	3.1	0.440	2.8			
	7	0.358	2.6	0.433	2.9	0.424	2.7			
	8	0.336	2.4	0.404	2.7	0.412	2.6			
	9	0.298	2.1	0.368	2.5	0.366	2.3			
	10	0.275	2.0	0.355	2.4	0.320	2.0			
	8	1	8.917	63.7	10.038	66.9				
		2	1.023	7.3	1.083	7.2				
		3	0.587	4.2	0.492	3.3				
4		0.459	3.3	0.460	3.1					
5		0.406	2.9	0.404	2.7					
6		0.402	2.9	0.369	2.5					
7		0.358	2.6	0.319	2.1					
8		0.348	2.5	0.313	2.1					
9		0.336	2.4	0.283	1.9					
10		0.298	2.1	0.262	1.7					
11	1	8.580	78.0	9.898	66.0	11.592	77.3	5.364	76.6	
	2	0.465	4.2	0.997	6.6	0.528	3.5	0.632	9.0	
	3	0.342	3.1	0.561	3.7	0.437	2.9	0.298	4.3	
	4	0.318	2.9	0.476	3.2	0.311	2.1	0.223	3.2	
	5	0.290	2.6	0.419	2.8	0.272	1.8	0.215	3.1	
	6	0.226	2.1	0.390	2.6	0.265	1.8	0.142	2.0	
	7	0.220	2.0	0.353	2.4	0.234	1.6	0.127	1.8	
	8	0.180	1.6	0.341	2.3	0.207	1.4			
	9	0.142	1.3	0.276	1.8	0.202	1.3			
	10	0.133	1.2	0.267	1.8	0.191	1.3			

APPENDIX E: Scree Plots for All Components

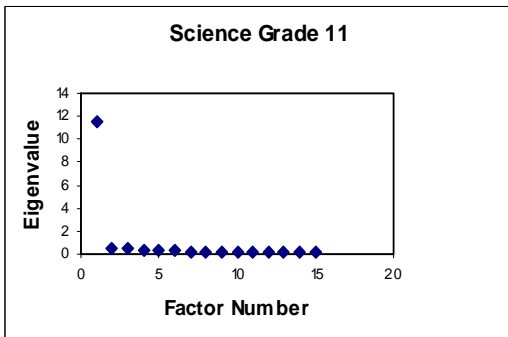
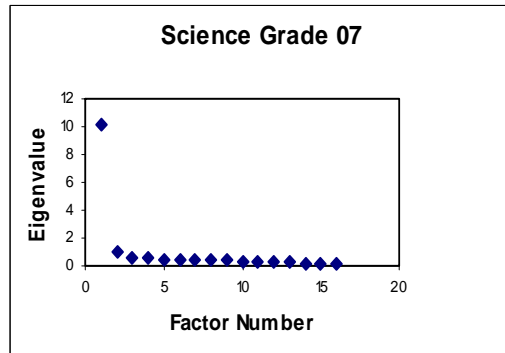
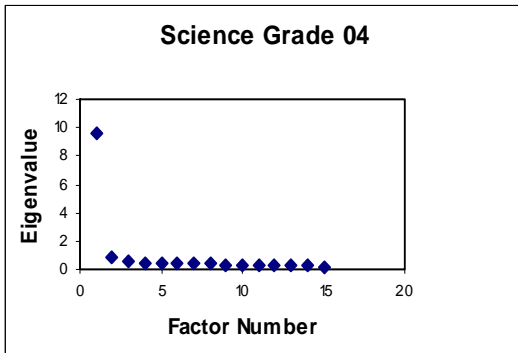
Reading



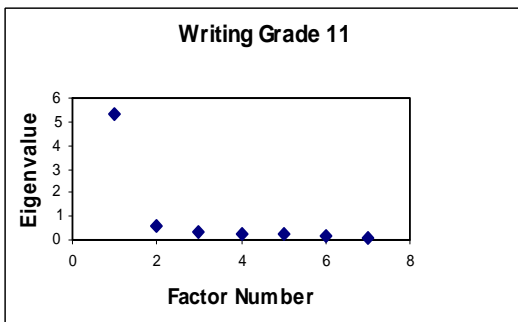
Mathematics



Science



Writing



APPENDIX F: Agreement between Teacher and Expert Scores by Item

Reading

Grade	Item	Number of Students*	% of Exact Agreement	% of Adjacent Agreement	% of Total Agreement
3	1	12	91.7	8.3	100.0
	2	12	91.7	8.3	100.0
	3	12	91.7	8.3	100.0
	4	13	100.0	0.0	100.0
	5	13	100.0	0.0	100.0
	6	13	100.0	0.0	100.0
	7	13	84.6	7.7	92.3
	8	12	91.7	8.3	100.0
	9	12	91.7	8.3	100.0
	10	12	91.7	8.3	100.0
	11	12	100.0	0.0	100.0
	12	12	100.0	0.0	100.0
	13	12	100.0	0.0	100.0
	14	11	91.7	0.0	91.7
4	1	11	90.9	0.0	90.9
	2	11	90.9	9.1	100.0
	3	11	100.0	0.0	100.0
	4	10	81.8	9.1	90.9
	5	10	90.9	0.0	90.9
	6	10	81.8	9.1	90.9
	7	10	100.0	0.0	100.0
	8	9	90.0	0.0	90.0
	9	9	90.0	0.0	90.0
	10	9	80.0	10.0	90.0
	11	9	90.0	0.0	90.0
	12	11	100.0	0.0	100.0
	13	11	100.0	0.0	100.0
	14	11	100.0	0.0	100.0
5	1	20	95.0	5.0	100.0
	2	19	100.0	0.0	100.0
	3	19	100.0	0.0	100.0
	4	21	95.2	4.8	100.0
	5	20	95.2	0.0	95.2
	6	20	95.0	5.0	100.0
	7	20	90.0	10.0	100.0
	8	18	94.4	5.6	100.0
	9	18	100.0	0.0	100.0
	10	18	100.0	0.0	100.0
	11	19	100.0	0.0	100.0
	12	19	100.0	0.0	100.0
	13	19	94.7	5.3	100.0
	14	18	100.0	0.0	100.0

Grade	Item	Number of Students*	% of Exact Agreement	% of Adjacent Agreement	% of Total Agreement
6	1	11	100.0	0.0	100.0
	2	11	90.9	0.0	90.9
	3	11	90.9	9.1	100.0
	4	11	90.9	0.0	90.9
	5	11	90.9	9.1	100.0
	6	11	90.9	9.1	100.0
	7	11	90.9	9.1	100.0
	8	11	100.0	0.0	100.0
	9	11	100.0	0.0	100.0
	10	11	100.0	0.0	100.0
	11	11	100.0	0.0	100.0
	12	11	100.0	0.0	100.0
	13	11	100.0	0.0	100.0
	14	11	100.0	0.0	100.0
7	1	16	100.0	0.0	100.0
	2	16	100.0	0.0	100.0
	3	16	81.3	18.8	100.0
	4	16	93.8	6.3	100.0
	5	16	100.0	0.0	100.0
	6	16	100.0	0.0	100.0
	7	15	100.0	0.0	100.0
	8	14	100.0	0.0	100.0
	9	14	92.9	7.1	100.0
	10	14	92.9	7.1	100.0
	11	14	92.9	7.1	100.0
	12	15	100.0	0.0	100.0
	13	15	100.0	0.0	100.0
	14	15	86.7	13.3	100.0
8	1	21	95.2	4.8	100.0
	2	21	95.2	4.8	100.0
	3	22	100.0	0.0	100.0
	4	22	90.9	9.1	100.0
	5	22	95.5	0.0	95.5
	6	21	90.5	9.5	100.0
	7	21	95.2	4.8	100.0
	8	20	100.0	0.0	100.0
	9	19	94.7	5.3	100.0
	10	19	94.7	5.3	100.0
	11	19	94.7	5.3	100.0
	12	18	100.0	0.0	100.0
	13	18	100.0	0.0	100.0
	14	18	94.4	5.6	100.0

Grade	Item	Number of Students*	% of Exact Agreement	% of Adjacent Agreement	% of Total Agreement
11	1	23	100.0	0.0	100.0
	2	23	100.0	0.0	100.0
	3	23	95.7	4.3	100.0
	4	23	95.7	4.3	100.0
	5	23	100.0	0.0	100.0
	6	23	100.0	0.0	100.0
	7	23	95.7	4.3	100.0
	8	23	95.7	4.3	100.0
	9	23	95.7	4.3	100.0
	10	23	95.7	4.3	100.0
	11	22	95.5	4.5	100.0

* Number of students includes those with complete pairs of ratings for the item.

Mathematics

Grade	Item	Number of Students*	% of Exact Agreement	% of Adjacent Agreement	% of Total Agreement
3	1	13	100.0	0.0	100.0
	2	13	100.0	0.0	100.0
	3	13	100.0	0.0	100.0
	4	13	100.0	0.0	100.0
	5	13	100.0	0.0	100.0
	6	13	100.0	0.0	100.0
	7	13	100.0	0.0	100.0
	8	13	100.0	0.0	100.0
	9	11	100.0	0.0	100.0
	10	12	100.0	0.0	100.0
	11	12	83.3	16.7	100.0
	12	12	83.3	16.7	100.0
	13	12	83.3	16.7	100.0
	14	12	100.0	0.0	100.0
	15	12	100.0	0.0	100.0
4	1	14	92.9	7.1	100.0
	2	14	100.0	0.0	100.0
	3	14	100.0	0.0	100.0
	4	14	92.9	7.1	100.0
	5	14	92.9	7.1	100.0
	6	14	100.0	0.0	100.0
	7	13	92.3	7.7	100.0
	8	14	92.9	7.1	100.0
	9	14	85.7	14.3	100.0
	10	14	100.0	0.0	100.0
	11	14	100.0	0.0	100.0
	12	14	100.0	0.0	100.0
	13	14	85.7	7.1	92.9
	14	14	92.9	7.1	100.0
	15	14	92.9	7.1	100.0
5	1	11	100.0	0.0	100.0
	2	11	100.0	0.0	100.0
	3	11	100.0	0.0	100.0
	4	11	100.0	0.0	100.0
	5	11	100.0	0.0	100.0
	6	11	100.0	0.0	100.0
	7	11	100.0	0.0	100.0
	8	11	100.0	0.0	100.0
	9	11	100.0	0.0	100.0
	10	11	100.0	0.0	100.0
	11	11	100.0	0.0	100.0
	12	11	100.0	0.0	100.0
	13	11	100.0	0.0	100.0
	14	11	100.0	0.0	100.0
	15	11	90.9	9.1	100.0

Grade	Item	Number of Students*	% of Exact Agreement	% of Adjacent Agreement	% of Total Agreement
6	1	14	100.0	0.0	100.0
	2	14	100.0	0.0	100.0
	3	14	100.0	0.0	100.0
	4	14	100.0	0.0	100.0
	5	13	100.0	0.0	100.0
	6	14	100.0	0.0	100.0
	7	14	100.0	0.0	100.0
	8	14	92.9	7.1	100.0
	9	14	92.9	7.1	100.0
	10	14	100.0	0.0	100.0
	11	14	100.0	0.0	100.0
	12	14	100.0	0.0	100.0
	13	14	92.9	7.1	100.0
	14	14	92.9	7.1	100.0
	15	14	100.0	0.0	100.0
7	1	21	100.0	0.0	100.0
	2	21	95.2	4.8	100.0
	3	21	100.0	0.0	100.0
	4	21	95.2	4.8	100.0
	5	21	95.2	4.8	100.0
	6	21	90.5	4.8	95.2
	7	21	81.0	19.0	100.0
	8	21	95.2	0.0	95.2
	9	20	100.0	0.0	100.0
	10	19	94.7	5.3	100.0
	11	19	100.0	0.0	100.0
	12	18	100.0	0.0	100.0
	13	18	100.0	0.0	100.0
	14	18	100.0	0.0	100.0
	15	18	100.0	0.0	100.0
8	1	11	100.0	0.0	100.0
	2	11	90.9	0.0	90.9
	3	11	100.0	0.0	100.0
	4	11	100.0	0.0	100.0
	5	11	90.9	0.0	90.9
	6	11	72.7	27.3	100.0
	7	11	90.9	9.1	100.0
	8	11	100.0	0.0	100.0
	9	11	90.9	9.1	100.0
	10	11	81.8	9.1	90.9
	11	11	90.9	9.1	100.0
	12	11	90.9	9.1	100.0
	13	12	91.7	0.0	91.7
	14	12	100.0	0.0	100.0
	15	11	91.7	0.0	91.7

Grade	Item	Number of Students*	% of Exact Agreement	% of Adjacent Agreement	% of Total Agreement
11	1	22	100.0	0.0	100.0
	2	21	95.5	0.0	95.5
	3	22	95.5	4.5	100.0
	4	22	95.5	4.5	100.0
	5	22	95.5	4.5	100.0
	6	22	95.5	4.5	100.0
	7	22	95.5	4.5	100.0
	8	22	100.0	0.0	100.0
	9	20	95.2	0.0	95.2
	10	22	100.0	0.0	100.0
	11	22	100.0	0.0	100.0
	12	22	95.5	0.0	95.5
	13	22	100.0	0.0	100.0
	14	22	100.0	0.0	100.0
	15	22	95.5	4.5	100.0

* Number of students includes those with complete pairs of ratings for the item.

Science

Grade	Item	Number of Students*	% of Exact Agreement	% of Adjacent Agreement	% of Total Agreement
4	1	16	100.0	0.0	100.0
	2	16	93.8	0.0	93.8
	3	16	93.8	6.3	100.0
	4	16	100.0	0.0	100.0
	5	15	100.0	0.0	100.0
	6	14	92.9	7.1	100.0
	7	15	100.0	0.0	100.0
	8	13	100.0	0.0	100.0
	9	14	100.0	0.0	100.0
	10	14	100.0	0.0	100.0
	11	14	100.0	0.0	100.0
	12	14	100.0	0.0	100.0
	13	14	100.0	0.0	100.0
	14	14	100.0	0.0	100.0
	15	14	100.0	0.0	100.0
7	1	18	100.0	0.0	100.0
	2	18	88.9	11.1	100.0
	3	18	94.4	5.6	100.0
	4	18	100.0	0.0	100.0
	5	18	100.0	0.0	100.0
	6	18	88.9	11.1	100.0
	7	18	94.4	5.6	100.0
	8	18	83.3	16.7	100.0
	9	18	88.9	11.1	100.0
	10	18	94.4	0.0	94.4
	11	18	100.0	0.0	100.0
	12	18	100.0	0.0	100.0
	13	18	100.0	0.0	100.0
	14	18	100.0	0.0	100.0
	15	18	94.4	5.6	100.0
	16	18	94.4	5.6	100.0
11	1	18	100.0	0.0	100.0
	2	18	100.0	0.0	100.0
	3	18	100.0	0.0	100.0
	4	18	100.0	0.0	100.0
	5	18	100.0	0.0	100.0
	6	18	94.4	5.6	100.0
	7	18	100.0	0.0	100.0
	8	18	100.0	0.0	100.0
	9	18	100.0	0.0	100.0
	10	17	94.4	0.0	94.4
	11	18	100.0	0.0	100.0
	12	18	100.0	0.0	100.0
	13	18	100.0	0.0	100.0

Grade	Item	Number of Students*	% of Exact Agreement	% of Adjacent Agreement	% of Total Agreement
	14	17	94.1	0.0	94.1
	15	18	94.4	0.0	94.4

* Number of students includes those with complete pairs of ratings for the item.

Writing

Grade	Item	Number of Students*	% of Exact Agreement	% of Adjacent Agreement	% of Total Agreement
11	1	23	100.0	0.0	100.0
	2	23	91.3	4.3	95.7
	3	22	95.5	4.5	100.0
	4	22	95.7	0.0	95.7
	5	22	100.0	0.0	100.0
	6	21	100.0	0.0	100.0
	7	21	95.2	4.8	100.0

* Number of students includes those with complete pairs of ratings for the item.

APPENDIX G: IAA Performance Theta Cuts and Transformation Constants

Subject	Grade	Theta Cuts			Slope	Intercept
		Foundational	Satisfactory	Mastery		
Reading	3	0.1300	1.0141	2.5554	29.10	470.49
	4	0.1606	0.9533	2.2373	29.58	471.80
	5	0.3002	0.9224	1.6128	52.96	451.15
	6	-0.0989	0.9395	2.2503	35.18	466.95
	7	0.2285	1.0623	2.9336	25.22	473.21
	8	0.3520	0.9456	2.3039	39.38	462.76
	11	0.0722	1.3064	3.5647	25.84	466.24
Mathematics	3	0.2641	0.8154	1.8005	51.96	457.64
	4	0.1101	0.8564	2.1916	40.10	465.66
	5	0.0605	0.9507	2.8620	21.52	479.54
	6	-0.0093	0.8134	2.0158	53.58	456.41
	7	0.1260	0.7382	1.9861	31.31	476.89
	8	0.0504	0.9781	2.3258	40.53	460.35
	11	0.1077	0.8287	2.4294	30.04	475.10
Science	4	-0.1476	0.7277	1.5537	73.12	446.79
	7	-0.0379	0.9222	1.8605	70.24	435.22
	11	-0.0917	0.7576	1.9360	36.84	472.09
Writing	11	-0.1002	0.7935	2.4546	46.30	463.26

Note: Previously, Writing was also administered in grades 3, 5, 6, & 8.

APPENDIX H: Item Statistics Summary

Reading

Grade		b-par	Infit	Outfit	Item Mean	Item-total Correlation
3	N of students	1,550	1,550	1,550	1,841	1,841
	N of items	14	14	14	14	14
	Mean	0.13	1.03	0.96	3.13	0.71
	STD	0.18	0.11	0.16	0.12	0.04
	Minimum	-0.30	0.85	0.70	2.93	0.64
	Maximum	0.38	1.19	1.24	3.42	0.78
4	N of students	1,746	1,746	1,746	1,902	1,902
	N of items	14	14	14	14	14
	Mean	0.02	1.02	0.94	3.20	0.73
	STD	0.31	0.07	0.14	0.20	0.03
	Minimum	-0.44	0.88	0.70	2.89	0.68
	Maximum	0.51	1.11	1.13	3.51	0.78
5	N of students	1,867	1,867	1,867	1,965	1,965
	N of items	14	14	14	14	14
	Mean	-0.05	1.04	0.95	3.27	0.72
	STD	0.22	0.10	0.17	0.15	0.04
	Minimum	-0.33	0.87	0.67	3.01	0.66
	Maximum	0.32	1.18	1.21	3.44	0.78
6	N of students	1,810	1,810	1,810	1,895	1,895
	N of items	14	14	14	14	14
	Mean	0.06	1.04	0.91	3.29	0.75
	STD	0.28	0.09	0.11	0.15	0.03
	Minimum	-0.51	0.89	0.74	3.08	0.70
	Maximum	0.42	1.19	1.11	3.59	0.79
7	N of students	1,856	1,856	1,856	1,928	1,928
	N of items	14	14	14	14	14
	Mean	0.13	1.03	0.93	3.36	0.74
	STD	0.25	0.14	0.17	0.13	0.05
	Minimum	-0.46	0.87	0.68	3.12	0.63
	Maximum	0.53	1.31	1.30	3.63	0.80
8	N of students	1,790	1,790	1,790	1,878	1,878
	N of items	14	14	14	14	14
	Mean	0.13	1.05	0.91	3.39	0.73
	STD	0.25	0.11	0.17	0.13	0.04
	Minimum	-0.27	0.90	0.66	3.17	0.64
	Maximum	0.56	1.28	1.21	3.60	0.79
11	N of students	1,665	1,665	1,665	2,061	2,061
	N of items	11	11	11	11	11
	Mean	0.19	1.02	0.87	3.42	0.81
	STD	0.28	0.15	0.28	0.11	0.04
	Minimum	-0.27	0.87	0.44	3.24	0.74
	Maximum	0.61	1.35	1.36	3.58	0.85

Mathematics

Grade		b-par	Infit	Outfit	Item Mean	Item-total Correlation
3	N of students	1,548	1,548	1,548	1,841	1,841
	N of items	15	15	15	15	15
	Mean	0.09	1.02	0.92	3.16	0.73
	STD	0.28	0.09	0.15	0.21	0.03
	Minimum	-0.30	0.91	0.70	2.75	0.65
	Maximum	0.56	1.27	1.24	3.45	0.77
4	N of students	1,744	1,744	1,744	1,901	1,901
	N of items	15	15	15	15	15
	Mean	0.04	1.04	0.92	3.32	0.74
	STD	0.33	0.10	0.18	0.20	0.05
	Minimum	-0.48	0.89	0.69	2.97	0.65
	Maximum	0.58	1.18	1.19	3.56	0.79
5	N of students	1,864	1,864	1,864	1,963	1,963
	N of items	15	15	15	15	15
	Mean	0.13	1.03	0.91	3.32	0.73
	STD	0.24	0.08	0.15	0.15	0.03
	Minimum	-0.29	0.85	0.64	3.10	0.65
	Maximum	0.48	1.21	1.15	3.59	0.78
6	N of students	1,807	1,807	1,807	1,892	1,892
	N of items	15	15	15	15	15
	Mean	0.20	1.04	0.91	3.32	0.75
	STD	0.31	0.15	0.19	0.16	0.05
	Minimum	-0.46	0.88	0.67	3.10	0.61
	Maximum	0.62	1.45	1.49	3.59	0.79
7	N of students	1,854	1,854	1,854	1,925	1,925
	N of items	15	15	15	15	15
	Mean	0.08	1.02	0.95	3.28	0.72
	STD	0.29	0.11	0.17	0.18	0.05
	Minimum	-0.59	0.88	0.72	2.96	0.62
	Maximum	0.56	1.24	1.25	3.63	0.79
8	N of students	1,789	1,789	1,789	1,877	1,877
	N of items	15	15	15	15	15
	Mean	0.23	1.04	0.92	3.37	0.75
	STD	0.24	0.18	0.23	0.12	0.06
	Minimum	-0.21	0.83	0.66	3.19	0.59
	Maximum	0.60	1.56	1.54	3.56	0.81
11	N of students	1,663	1,663	1,663	2,058	2,058
	N of items	15	15	15	15	15
	Mean	0.16	1.05	0.96	3.27	0.75
	STD	0.25	0.16	0.20	0.14	0.05
	Minimum	-0.27	0.83	0.67	3.05	0.64
	Maximum	0.60	1.39	1.32	3.52	0.82

Science

Grade		b-par	Infit	Outfit	Item Mean	Item-total Correlation
4	N of students	1,743	1,743	1,743	1,899	1,899
	N of items	15	15	15	15	15
	Mean	0.14	1.04	0.95	3.20	0.73
	STD	0.23	0.09	0.17	0.15	0.04
	Minimum	-0.21	0.89	0.65	2.83	0.66
	Maximum	0.73	1.22	1.24	3.42	0.79
7	N of students	1,853	1,853	1,853	1,926	1,926
	N of items	16	16	16	16	16
	Mean	0.22	1.04	0.92	3.37	0.73
	STD	0.35	0.09	0.19	0.18	0.04
	Minimum	-0.46	0.88	0.62	3.10	0.67
	Maximum	0.71	1.22	1.37	3.69	0.79
11	N of students	1,661	1,661	1,661	2,055	2,055
	N of items	15	15	15	15	15
	Mean	0.16	0.99	0.88	3.37	0.81
	STD	0.26	0.16	0.20	0.12	0.05
	Minimum	-0.07	0.79	0.55	2.98	0.68
	Maximum	0.95	1.35	1.35	3.49	0.86

Writing

Grade		b-par	Infit	Outfit	Item Mean	Item-total Correlation
	N of students	1,660	1,660	1,660	2,056	2,056
	N of items	7	7	7	7	7
11	Mean	0.19	1.05	0.91	3.35	0.82
	STD	0.20	0.29	0.29	0.11	0.06
	Minimum	0.00	0.81	0.59	3.15	0.71
	Maximum	0.52	1.55	1.42	3.45	0.86