Illinois Alternate Assessment
2010 Technical Manual

Illinois State Board of Education
Division of Assessment

Table of Contents

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 2010 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://education.umn.edu/NCEO/OnlinePubs/Policy16.htm).
- Alternate assessments aligned with grade-level achievement standards
- Alternate assessments based on alternate achievement standards.

The 2004 IDEA amendments reinforce the principle that children with disabilities may be appropriately assessed through one of these four alternatives. To qualify as an assessment under Title I, an alternate assessment must be aligned with the State's content standards, must yield results separately 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.

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	3, 5, 6, 8, 11	5	7	40 %

For 2010, the test length of the IAA stayed the same as 2009 for all subjects and grades. The 2010 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
03	1	10	71
03	2	4	29
04	1	10	71
04	2	4	29
05	1	9	64
05	2	5	36
06	1	9	64
06	2	5	36
07	1	10	71
07	2	4	29
08	1	10	71
08	2	4	29
11	1	11	100

Table 1.2b: Mathematics Blueprint

Grade	Goal	Number of Items	Percent of Items
03	6	7	47
03	7	2	13
03	8	3	20
03	9	2	13
03	10	1	7
04	6	7	47
04	7	2	13
04	8	2	13
04	9	2	13
04	10	2	13
05	6	6	40
05	7	2	13
05	8	3	20
05	9	2	13
05	10	2	13
06	6	4	27
06	7	2	13
06	8	3	20
06	9	4	27
06	10	2	13
07	6	4	27
07	7	3	20
07	8	3	20
07	9	3	20
07	10	2	13
08	6	4	27
08	7	3	20
08	8	3	20
08	9	2	13
08	10	3	20
11	6	5	33
11	7	2	13
11	8	3	20
11	9	4	27
11	10	1	7

Table 1.2c: Science Blueprint

Grade	Goal	Number of Items	Percent of Items
04	11	2	13
04	12	10	67
04	13	3	20
07	11	2	13
07	12	12	75
07	13	2	13
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
03	3	7	100
05	3	7	100
06	3	7	100
08	3	7	100
11	3	7	100

Item Development

Item Development Cycle

New items are acquired each year to establish an adequate item pool for test construction. The planning of new item development is based on content coverage and the number of test items needed for the test. 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 ensure 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. 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.

[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

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

Test Implementation Manual

The *IAA Test Implementation Manual* was developed by Pearson for 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

- Throughout January of 2010, Pearson, in partnership with ISBE, conducted multiple onsite trainings in locations statewide in preparation for the spring 2010 operational assessment.
- Each session was attended by approximately 100 Illinois IAA coordinators and educators.

- Additionally, in December 2009 and January 2010, Pearson and ISBE staff hosted (3) webinar training sessions, which were attended by nearly 500 Illinois IAA Coordinators and educators.

Training Facilitators

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

Training Materials

- All materials in support of the IAA Regional Trainings and spring 2010 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 2010 packaging and distribution requirements
- Regional 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 2010 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, …, y_T are the T scores that can be gained on the item. The values, n_{Ftk} and n_{Rtk}, 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_t. The "+" indicates total number over a particular index (Zwick, Donoghue, & Grima, 1993).

Table 1.4 2 × 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:

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

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

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

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

m_{Rk} = 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| ≤0 .17
B	Moderate DIF	Significant Mantel χ^2 and 0.17 < |SMD/SD| ≤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 2010 census items that present DIF. 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.

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

		Male/Female			White/Black			White/Hispanics		
Subject	**Grade**	**A**	**B**	**C**	**A**	**B**	**C**	**A**	**B**	**C**
Reading	3	14	0	0	14	0	0	14	0	0
	4	14	0	0	14	0	0	14	0	0
	5	13	1	0	14	0	0	14	0	0
	6	14	0	0	14	0	0	14	0	0
	7	14	0	0	13	1	0	14	0	0
	8	14	0	0	14	0	0	14	0	0
	11	10	1	0	11	0	0	11	0	0
Mathematics	3	14	1	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	13	2	0	15	0	0	15	0	0
Science	4	15	0	0	15	0	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	3	7	0	0	7	0	0	6	0	1
	5	7	0	0	6	0	1	7	0	0
	6	7	0	0	7	0	0	7	0	0
	8	7	0	0	7	0	0	7	0	0
	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:

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

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 2010. 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:

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

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	1999	44.88	11.71	0.94	2.92
	4	2120	45.93	10.83	0.93	2.83
	5	2025	46.32	10.88	0.94	2.68
	6	2128	47.26	10.49	0.94	2.58
	7	2075	47.65	10.80	0.95	2.45
	8	2023	48.39	9.85	0.94	2.48
	11	2104	38.60	8.63	0.95	1.95
Mathematics	3	1997	48.33	12.23	0.94	3.02
	4	2119	50.47	11.38	0.94	2.79
	5	2022	50.51	11.29	0.94	2.69
	6	2127	51.15	11.18	0.95	2.56
	7	2075	50.62	11.14	0.94	2.63
	8	2021	51.34	10.51	0.94	2.61
	11	2106	50.23	11.48	0.94	2.71
Science	4	2118	49.39	11.45	0.94	2.88
	7	2072	54.63	11.59	0.95	2.61
	11	2104	51.92	11.68	0.96	2.42
Writing	3	1995	22.29	6.19	0.89	2.03
	5	2020	23.12	5.61	0.89	1.87
	6	2126	23.03	5.77	0.90	1.82
	8	2021	24.12	5.18	0.89	1.74
	11	2103	23.89	5.84	0.92	1.63

Table 2.2a: Reliability Estimates by Ethnicity

Grade	Subgroup	Reading	Mathematics	Science	Writing
3	Asian	0.93	0.95	-	0.90
	Black	0.95	0.95	-	0.92
	Hispanic	0.95	0.95	-	0.90
	White	0.92	0.92	-	0.86
4	Asian	0.92	0.92	0.92	-
	Black	0.95	0.95	0.95	-
	Hispanic	0.94	0.95	0.95	-
	White	0.92	0.93	0.92	-
5	Asian	0.93	0.93	-	0.91
	Black	0.95	0.95	-	0.91
	Hispanic	0.95	0.95	-	0.91
	White	0.93	0.94	-	0.87
6	Asian	0.95	0.96	-	0.91
	Black	0.95	0.96	-	0.92
	Hispanic	0.93	0.94	-	0.90
	White	0.94	0.94	-	0.89
7	Asian	0.95	0.95	0.96	-
	Black	0.95	0.95	0.96	-
	Hispanic	0.96	0.95	0.95	-
	White	0.94	0.94	0.94	-
8	Asian	0.88	0.93	-	0.86
	Black	0.94	0.94	-	0.90
	Hispanic	0.94	0.95	-	0.90
	White	0.93	0.93	-	0.88
11	Asian	0.96	0.95	0.96	0.95
	Black	0.96	0.95	0.97	0.93
	Hispanic	0.95	0.95	0.96	0.91
	White	0.94	0.93	0.95	0.91

Table 2.2b: Reliability Estimates by LEP

Grade	Subgroup	Reading	Mathematics	Science	Writing
3	LEP	0.94	0.94	-	0.90
	Non-LEP	0.94	0.94	-	0.89
4	LEP	0.92	0.94	0.93	-
	Non-LEP	0.93	0.94	0.94	-
5	LEP	0.95	0.95	-	0.91
	Non-LEP	0.94	0.94	-	0.89
6	LEP	0.93	0.94	-	0.91
	Non-LEP	0.94	0.95	-	0.90
7	LEP	0.97	0.97	0.97	-
	Non-LEP	0.95	0.94	0.95	-
8	LEP	0.94	0.94	-	0.91
	Non-LEP	0.94	0.94	-	0.88
11	LEP	0.92	0.92	0.94	0.86
	Non-LEP	0.95	0.94	0.96	0.92

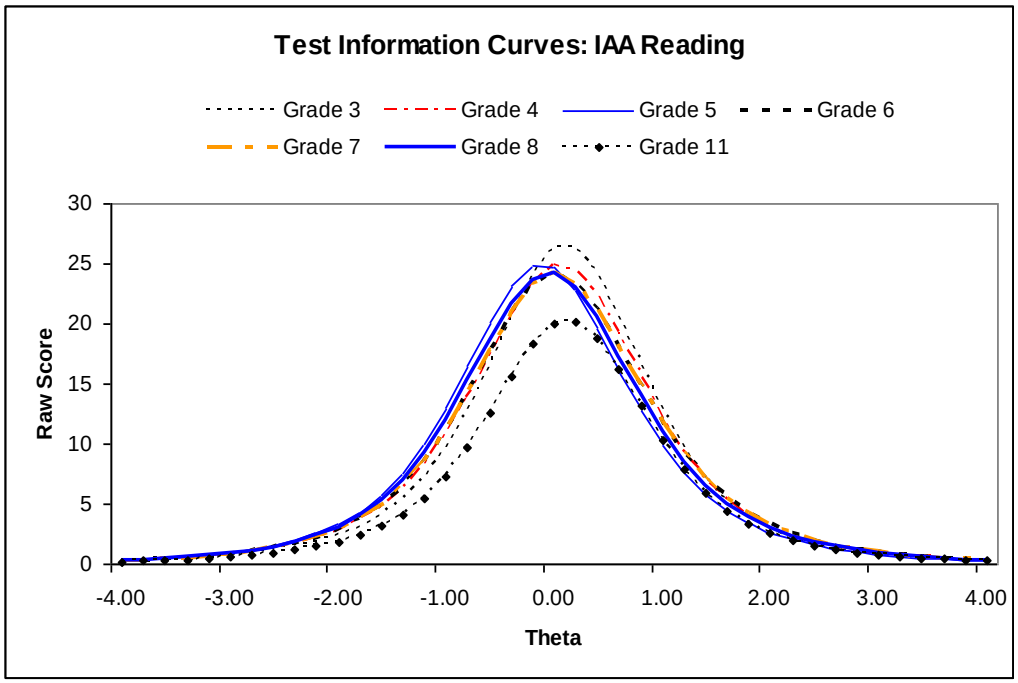
Table 2.2c: Reliability Estimates by Income

Grade	Subgroup	Reading	Mathematics	Science	Writing
3	Low-Income	0.94	0.94	-	0.89
	Non-Low-Income	0.93	0.94	-	0.89
4	Low-Income	0.94	0.94	0.94	-
	Non-Low-Income	0.93	0.94	0.93	-
5	Low-Income	0.94	0.95	-	0.89
	Non-Low-Income	0.94	0.94	-	0.89
6	Low-Income	0.94	0.94	-	0.90
	Non-Low-Income	0.94	0.95	-	0.90
7	Low-Income	0.95	0.95	0.95	-
	Non-Low-Income	0.95	0.94	0.95	-
8	Low-Income	0.94	0.94	-	0.90
	Non-Low-Income	0.93	0.93	-	0.88
11	Low-Income	0.94	0.94	0.95	0.91
	Non-Low-Income	0.95	0.95	0.96	0.93

IRT Test Information Function

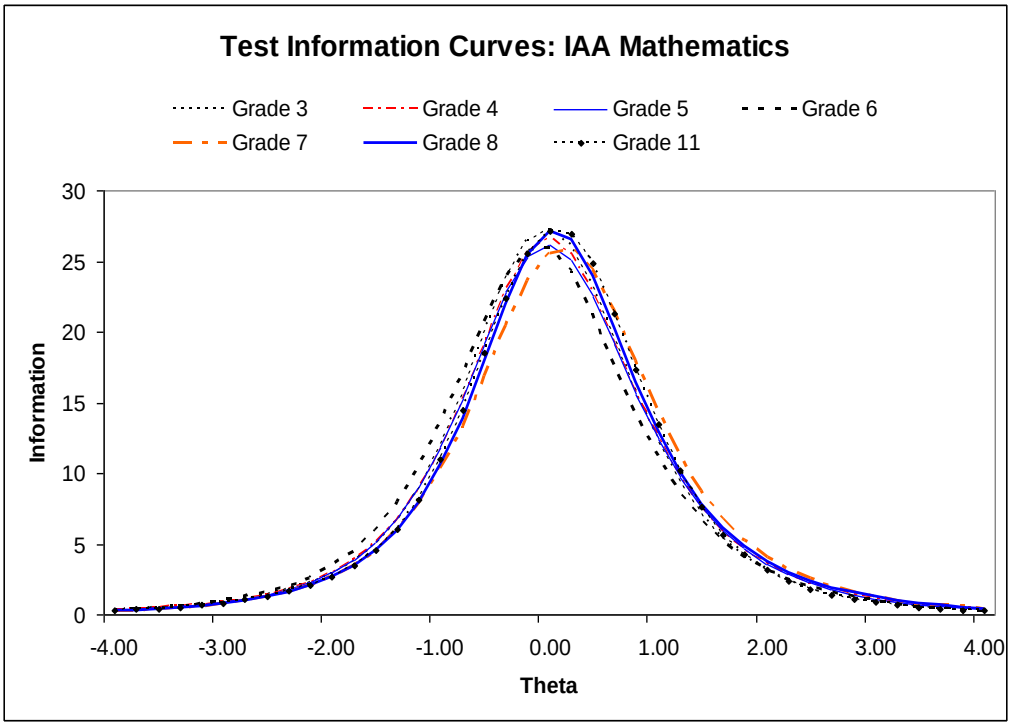
The reliability coefficients reported above were derived within the context of classical test theory and provide a single measure of precision for the entire test. With the Item Response Theory (IRT), it is possible to measure the relative precision of the test at different points on the scale. The amount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

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



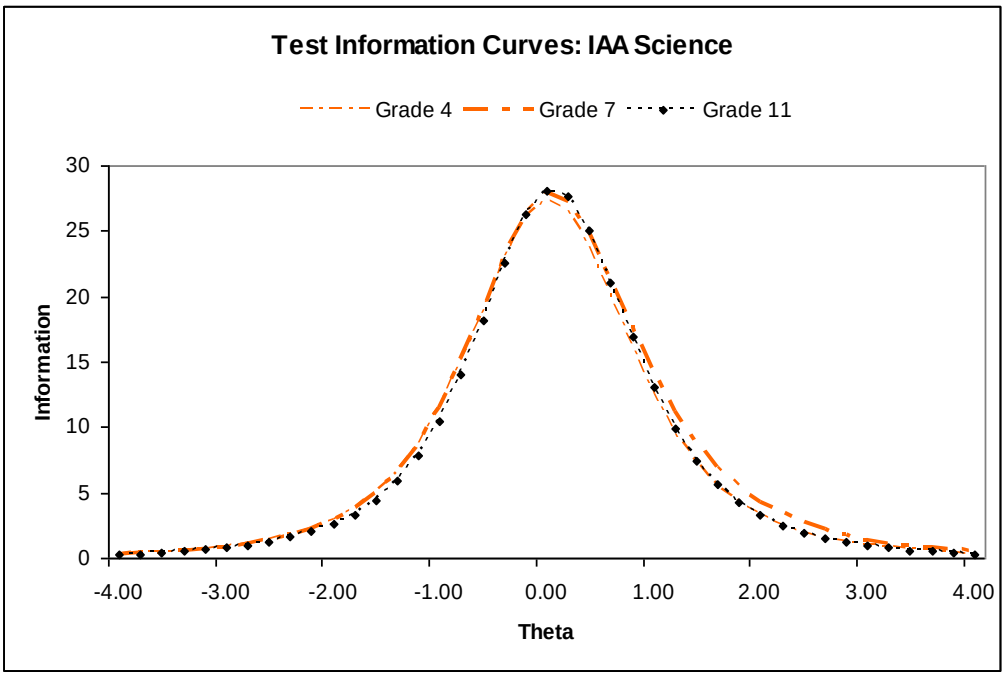
Note: Grades 3-8 have 14 items and grade 11 has 11 items.

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



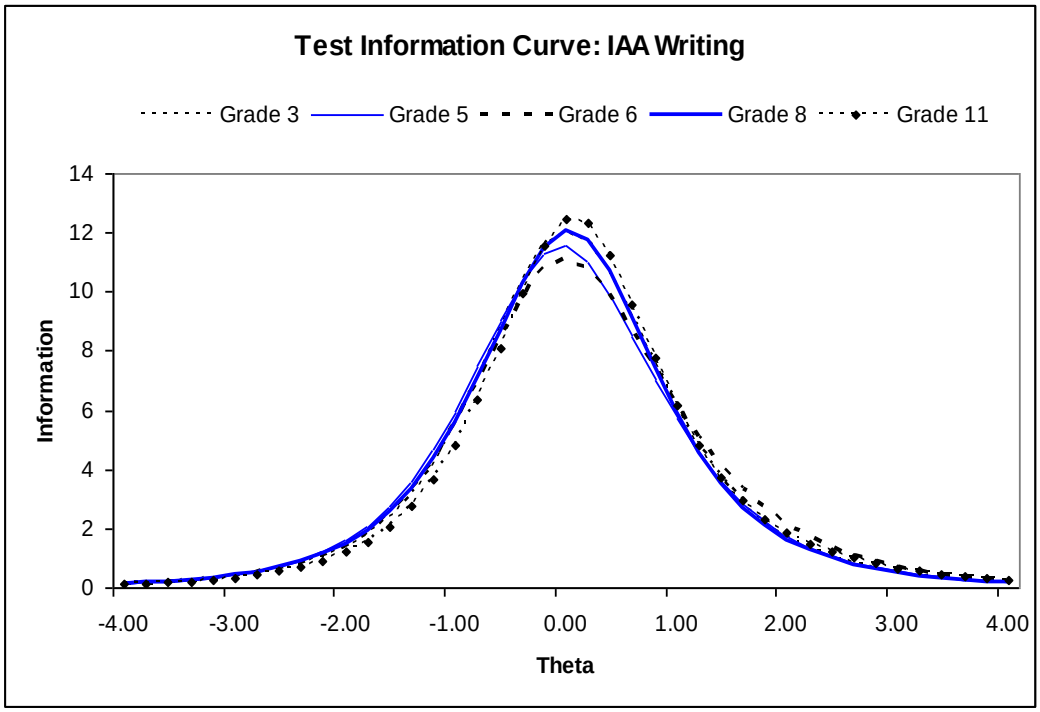
Note: Mathematics has 15 items for all grades.

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



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

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



Note: Writing has 7 items for all grades.

IRT Conditional SEM

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

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

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(θ) (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

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

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, 17.3 is the accuracy of Entry level and 21.5 is the accuracy of Foundational level. The sum of the diagonal values, 17.3, 21.5, 27.2, and 10.9, is the overall test classification accuracy (76.9). 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	17.3	3.1	0.0	0.2	20.7
R2	1.4	21.5	7.0	0.8	30.8
R3	0.0	3.0	27.2	5.6	35.8
R4	0.0	0.0	1.8	10.9	12.8
Observed	18.8	27.7	36.1	17.5	100.0

Table 2.4: IAA Classification Accuracy

Grade	Reading	Mathematics	Science	Writing
3	76.9	74.5		71.2
4	76.9	76.8	76.8	
5	74.0	78.4		75.7
6	76.8	77.7		70.9
7	75.9	77.7	78.7	
8	74.3	77.0		67.4
11	69.0	78.2	78.2	71.1

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.

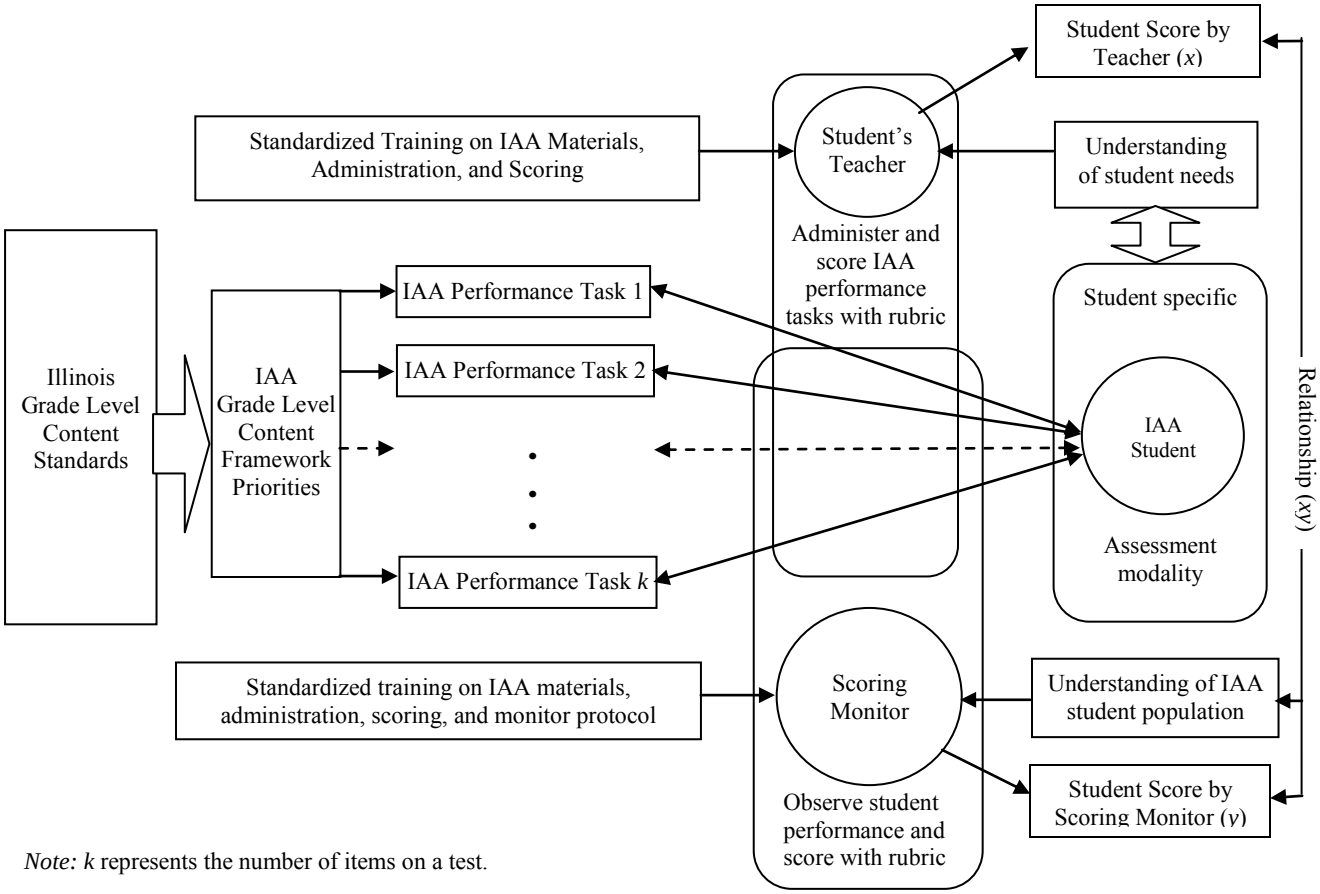


Figure 3.1 IAA performance model with validation component

Content-related Validity

The content validity of the IAA is established through content standard specification that defines the measurement of actual performance. It is fulfilled through item alignment study, test design, test/item review, and test/item analyses. As described in Chapter 1 of this report, the IAA measures actual student performance through trained teachers, a specified set of content-valid items, a test administration that is appropriate to the student's usual communication methods, and a standardized scoring rubric. Evidence of content validity detailed in Chapter 1 also includes descriptions of the test blueprint, the test construction process, and the decisions made for defining and developing the IAA test. In addition, an alignment study for each subject area was reported in April 2009 by the WIDA Consortium.

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; Raîche, 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 64.9% to 80.4% 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	47.80	70.14	-	22.07
4	64.32	24.79	16.89	-
5	58.98	66.39	-	51.69
6	43.95	38.70	-	58.81
7	158.39	28.10	28.19	-
8	31.86	22.62	-	19.96
11	83.90	38.89	55.05	24.89

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. In other words, 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	50.1	36.1	14.0	72	28	2.58
4	53.5	39.5	14.0	74	26	2.82
5	57.3	43.3	14.0	76	24	3.09
6	57.3	43.3	14.0	76	24	3.09
7	62.5	48.5	14.0	78	22	3.46
8	52.7	38.7	14.0	73	27	2.76
11	47.6	36.6	11.0	77	23	3.33
Mathematics						
3	67.9	52.9	15.0	78	22	3.53
4	66.0	51.0	15.0	77	23	3.40
5	70.9	55.9	15.0	79	21	3.73
6	73.6	58.6	15.0	80	20	3.91
7	65.1	50.1	15.0	77	23	3.34
8	57.0	42.0	15.0	74	26	2.80
11	61.3	46.3	15.0	76	24	3.09
Science						
4	56.9	41.9	15.0	74	26	2.79
7	68.9	52.9	16.0	77	23	3.31
11	63.1	48.1	15.0	76	24	3.21
Writing						
3	22.1	15.1	7.0	68	32	2.16
5	19.7	12.7	7.0	64	36	1.81
6	22.4	15.4	7.0	69	31	2.20
8	17.7	10.7	7.0	60	40	1.53
11	24.0	17.0	7.0	71	29	2.43

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	24.07	27.77	30.08
4	1.5	1.3	1.3	26.33	30.38	30.38
5	1.4	1.3	1.2	30.93	33.31	36.08
6	1.5	1.3	1.2	28.87	33.31	36.08
7	1.4	1.4	1.2	34.64	34.64	40.42
8	1.5	1.3	1.2	25.80	29.77	32.25
11	1.4	1.3	1.2	26.14	28.15	30.50
Mathematics						
3	1.5	1.4	1.1	35.27	37.79	48.09
4	1.7	1.4	1.2	30.00	36.43	42.50
5	1.4	1.4	1.2	39.93	39.93	46.58
6	1.5	1.3	1.3	39.07	45.08	45.08
7	1.6	1.4	1.2	31.31	35.79	41.75
8	1.7	1.4	1.3	24.71	30.00	32.31
11	1.6	1.5	1.3	28.94	30.87	35.62
Science						
4	1.8	1.2	-	23.28	34.92	-
7	1.6	1.3	1.2	33.06	40.69	44.08
11	1.6	1.3	1.2	30.06	37.00	40.08
Writing						
3	1.5	1.2	1.1	10.07	12.58	13.73
5	1.4	1.2	1.2	9.07	10.58	10.58
6	1.3	1.3	1.2	11.85	11.85	12.83
8	1.5	1.2	1.2	7.13	8.92	8.92
11	1.4	1.4	1.1	12.14	12.14	15.45

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 .90 to .92; the correlation between Reading and Science scores also ranges from .90 to .92; the correlation between Reading and Writing ranges from .88 to .90; the correlation between Mathematics and Writing ranges from .89 to .90; and the correlation between Mathematics and Science is from .92 to .93.

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 0.68 to 0.79 across subjects and grades.

Table 3.3: Correlation among IAA Assessments

Grade	Test	Reading	Mathematics	Science	Writing
3	Reading	1.00	0.91	-	0.89
	Mathematics	0.91	1.00	-	0.89
	Science	-	-	-	-
	Writing	0.89	0.89	-	1.00
4	Reading	1.00	0.90	0.90	-
	Mathematics	0.90	1.00	0.92	-
	Science	0.90	0.92	1.00	-
	Writing	-	-	-	-
5	Reading	1.00	0.91	-	0.89
	Mathematics	0.91	1.00	-	0.90
	Science	-	-	-	-
	Writing	0.89	0.90	-	1.00
6	Reading	1.00	0.91	-	0.89
	Mathematics	0.91	1.00	-	0.90
	Science	-	-	-	-
	Writing	0.89	0.90	-	1.00
7	Reading	1.00	0.91	0.91	-
	Mathematics	0.91	1.00	0.93	-
	Science	0.91	0.93	1.00	-
	Writing	-	-	-	-
8	Reading	1.00	0.92	-	0.88
	Mathematics	0.92	1.00	-	0.90
	Science	-	-	-	-
	Writing	0.88	0.90	-	1.00
11	Reading	1.00	0.90	0.92	0.90
	Mathematics	0.90	1.00	0.93	0.90
	Science	0.92	0.93	1.00	0.92
	Writing	0.90	0.90	0.92	1.00

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

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

Criterion-related Validity

In order to examine the criterion-related validity of the IAA, a study was conducted in 2010 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 2010 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 2010 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 2010 validation study has comparable percentages of male and female students with the spring 2010 IAA student population

Table 3.5: Spring 2010 IAA Student Population and Validation Sample

	N	Percentage	
		Male	**Female**
IAA Population	14,864	65.2%	34.8%
Validation Sample	166	66.4%	33.6%

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 94% for all subjects and grades, and the average percentage of total agreement exceeded 98% 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	10	97.8	1.5	99.3
	4	14	7	95.5	4.5	100.0
	5	14	15	99.6	0.4	100.0
	6	14	17	98.7	1.3	100.0
	7	14	11	100.0	0.0	100.0
	8	14	9	100.0	0.0	100.0
	11	11	36	98.2	0.5	98.7
Mathematics	3	15	8	94.2	5.0	99.2
	4	15	10	97.3	1.3	98.7
	5	15	16	96.0	3.2	99.2
	6	15	12	95.9	4.1	100.0
	7	15	8	98.3	1.7	100.0
	8	15	11	98.8	1.2	100.0
	11	15	37	97.7	1.8	99.5
Science	4	15	6	100.0	0.0	100.0
	7	16	12	98.9	1.1	100.0
	11	15	35	99.4	0.6	100.0
Writing	3	7	10	98.6	1.4	100.0
	5	7	12	97.8	2.2	100.0
	6	7	9	100.0	0.0	100.0
	8	7	10	95.7	4.3	100.0
	11	7	37	98.5	1.5	100.0

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 0.96 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		*r*
		Teacher	Expert	Teacher	Expert	Teacher	Expert	Teacher	Expert	
3	12	36.33	36.00	19.64	19.47	3	3	56	56	0.999
4	8	46.13	46.25	11.97	12.09	22	22	56	56	0.999
5	18	39.39	39.33	19.23	19.30	6	6	56	56	1.000
6	17	52.29	52.24	5.72	5.74	35	35	56	56	0.997
7	11	53.00	53.00	4.36	4.36	43	43	56	56	1.000
8	9	46.33	46.33	15.34	15.34	14	14	56	56	1.000
11	37	41.35	41.43	5.28	5.24	23	22	44	44	0.995

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

Grade	Number of Students	Mean		Std Deviation		Minimum		Maximum		*r*
		Teacher	Expert	Teacher	Expert	Teacher	Expert	Teacher	Expert	
3	8	51.38	52.00	7.56	7.09	38	41	60	60	0.986
4	10	54.80	55.20	6.30	6.81	38	37	60	60	0.971
5	18	50.22	50.56	14.79	14.84	1	1	60	60	0.996
6	12	54.33	54.08	8.26	8.12	29	29	60	60	0.997
7	8	54.38	54.38	3.25	3.20	49	49	59	59	0.986
8	12	49.33	49.50	13.63	13.65	16	16	60	60	1.000
11	38	54.08	54.11	7.74	7.70	25	25	60	60	0.997

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

Grade	Number of Students	Mean		Std Deviation		Minimum		Maximum		*r*
		Teacher	Expert	Teacher	Expert	Teacher	Expert	Teacher	Expert	
4	6	55.83	55.83	3.31	3.31	52	52	60	60	1.000
7	13	54.31	54.15	15.46	15.44	16	16	64	64	0.999
11	35	58.29	58.26	2.53	2.62	47	46	60	60	0.994

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

Grade	Number of Students	Mean		Std Deviation		Minimum		Maximum		*r*
		Teacher	Expert	Teacher	Expert	Teacher	Expert	Teacher	Expert	
3	10	26.70	26.80	1.89	1.81	22	22	28	28	0.986
5	13	24.15	24.31	5.30	5.33	12	12	28	28	0.998
6	9	24.22	24.22	5.91	5.91	9	9	28	28	1.000
8	10	23.30	23.20	5.74	6.18	12	12	28	28	0.995
11	37	26.70	26.70	1.58	1.49	22	23	28	28	0.966

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	Familiar	Somewhat Familiar	Not At All
☐	☐	☐	☐

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	☐	☐	☐	☐
Better than average	☐	☐	☐	☐
Average	☐	☐	☐	☐
Worse than average	☐	☐	☐	☐
Much worse than average	☐	☐	☐	☐
Prefer not to answer	☐	☐	☐	☐

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 77.0% to 86.0% across the grades and subjects. About 10.8% to 16.4% of the students were rated *familiar* by the teachers. *Somewhat familiar* and *not at all* made up only 3.2% to 7.2% 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	81.4	13.1	4.2	1.3	94.5
	4	83.4	13.0	2.5	1.1	96.4
	5	82.9	13.8	2.9	0.4	96.7
	6	78.8	16.4	3.8	1.0	95.2
	7	81.0	14.9	3.0	1.1	95.9
	8	86.0	10.8	2.5	0.8	96.8
	11	77.0	15.8	4.3	2.9	92.8
Mathematics	3	81.5	13.1	4.2	1.3	94.6
	4	83.5	12.9	2.5	1.1	96.4
	5	82.9	13.7	2.9	0.4	96.6
	6	78.8	16.4	3.8	1.0	95.2
	7	81.0	14.9	3.0	1.1	95.9
	8	85.9	10.8	2.4	0.8	96.7
	11	77.0	15.8	4.3	2.9	92.8
Science	4	83.5	12.9	2.5	1.1	96.4
	7	81.0	14.8	3.0	1.1	95.8
	11	77.0	15.7	4.3	2.9	92.7
Writing	3	81.5	13.1	4.2	1.3	94.6
	5	83.0	13.7	2.9	0.4	96.7
	6	78.8	16.4	3.8	1.0	95.2
	8	85.9	10.8	2.5	0.8	96.7
	11	77.0	15.7	4.3	2.9	92.7

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 92.7% 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 55.9% to 63.0% 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 21.0% to 27.7% of students were considered *better than average*, and 7.4% to 11.3% 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	9.0	27.5	55.9	3.7	1.0	3.0
	4	10.0	26.0	58.7	3.2	0.6	1.5
	5	9.0	24.5	60.7	3.1	0.8	1.9
	6	10.4	23.3	60.9	3.2	0.4	1.8
	7	8.4	25.4	61.3	2.5	0.7	1.7
	8	10.7	26.2	59.2	2.2	0.3	1.3
	11	8.1	27.3	58.5	1.9	0.4	3.7
Mathematics	3	9.1	24.4	59.0	3.8	0.8	2.9
	4	11.1	24.9	59.4	2.7	0.4	1.5
	5	9.9	24.2	60.6	2.7	0.6	2.0
	6	9.8	24.2	61.3	2.3	0.5	1.8
	7	7.8	24.4	63.0	2.5	0.8	1.5
	8	11.2	24.3	61.0	1.8	0.5	1.2
	11	7.4	22.9	61.9	3.5	0.8	3.5
Science	4	9.8	27.7	57.4	2.9	0.4	1.8
	7	8.6	24.6	62.4	1.9	0.7	1.9
	11	7.4	25.6	60.5	1.7	0.5	4.3
Writing	3	10.7	26.2	56.3	2.6	0.9	3.3
	5	10.2	23.9	60.6	2.6	0.7	2.1
	6	10.4	21.0	62.8	3.1	0.7	1.9
	8	11.3	24.6	60.7	1.7	0.4	1.3
	11	8.9	23.2	61.7	1.8	0.7	3.7

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:

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

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

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 2010 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 2010 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 2010 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:

SS = Theta*slope + intercept
SE = Theta*slope

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 2010 is presented in Tables 5.1a to 5.1d.

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

Performance	Grade 3		Grade 4		Grade 5		Grade 6		Grade 7		Grade 8		Grade 11	
Level	Low	High	Low	High	Low	High	Low	High	Low	High	Low	High	Low	High
Entry	300	473	300	476	300	466	300	462	300	478	300	476	300	467
Foundational	474	499	477	499	467	499	463	499	479	499	477	499	468	499
Satisfactory	500	544	500	537	500	536	500	545	500	546	500	552	500	557
Mastery	545	580	538	587	537	655	546	610	547	576	553	619	558	565

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

Performance	Grade 3		Grade 4		Grade 5		Grade 6		Grade 7		Grade 8		Grade 11	
Level	Low	High	Low	High	Low	High	Low	High	Low	High	Low	High	Low	High
Entry	300	470	300	469	300	480	300	455	300	480	300	461	300	477
Foundational	471	499	470	499	481	499	456	499	481	499	462	499	478	499
Satisfactory	500	550	500	553	500	540	500	563	500	538	500	554	500	547
Mastery	551	658	554	621	541	570	564	682	539	605	555	637	548	606

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

Performance	Grade 4		Grade 7		Grade 11	
Level	Low	High	Low	High	Low	High
Entry	300	435	300	432	300	468
Foundational	436	499	433	499	469	499
Satisfactory	500	559	500	565	500	542
Mastery	560	700	566	700	543	621

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

Performance Level	Grade 3		Grade 5		Grade 6		Grade 8		Grade 11	
	Low	High	Low	High	Low	High	Low	High	Low	High
Entry	300	436	300	465	300	442	300	449	300	458
Foundational	437	499	466	499	443	499	450	499	459	499
Satisfactory	500	558	500	558	500	578	500	565	500	576
Mastery	559	659	559	594	579	640	566	619	577	642

Based on the scale score cuts presented above, IAA students were classified into four performance levels: Entry, Foundational, Satisfactory, and Mastery. The results for 2010 are presented in Tables 5.2a to 5.2d along with the results for 2009. 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
4	2009	21	20	35	24
	2010	18	22	41	20
5	2009	24	18	23	36
	2010	23	18	20	40
6	2009	14	18	36	32
	2010	11	21	32	37
7	2009	15	20	42	24
	2010	15	15	36	33
8	2009	18	14	37	32
	2010	16	14	38	32
11	2009	13	18	31	38
	2010	12	17	34	36

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
4	2009	17	18	35	30
	2010	15	18	43	24
5	2009	16	20	41	23
	2010	12	22	43	23
6	2009	14	15	33	38
	2010	11	14	31	44
7	2009	16	15	41	29
	2010	14	13	40	32
8	2009	12	19	37	31
	2010	9	18	40	33
11	2009	16	14	43	26
	2010	14	13	45	28

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
7	2009	11	17	29	43
	2010	10	13	28	49
11	2009	12	13	28	47
	2010	11	12	30	47

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

Grade	Year	Entry	Foundational	Satisfactory	Mastery
3	2009	13	16	30	41
	2010	13	18	28	42
5	2009	11	17	43	29
	2010	9	13	57	21
6	2009	13	20	36	31
	2010	13	19	33	36
8	2009	13	18	30	40
	2010	12	16	43	29
11	2009	12	11	26	50
	2010	11	10	28	51

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).

Raîche, 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

Level 4:	*Level 3:*	*Level 2:*	*Level 1:*
The student correctly performs the task without assistance or with a single repetition of instructions or refocusing. • The student 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. ***The student then responds correctly.***	**The student correctly performs the task with a general prompt.** • 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: 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. ***The student then responds correctly.***	**The student correctly performs the task with specific prompts.** • If the student responds incorrectly to the task at Level 3 when given adequate wait time, the teacher provides specific prompts to direct the student's correct response such as: o Modeling exact response, "This is a picture of a dog, what is this?" (Show a picture of a dog). o After physically guiding the student to the correct response such as using hand over hand, the student then indicates the correct answer in his/her mode of communication. ***The student responds correctly after being given the correct answer.***	**The student does not perform the task at Level 2 or provides an incorrect response despite Level 2 support.** ***The student does not respond or does not respond correctly. Teacher demonstrates response and moves on to the next task.***

Illinois State Board of Education has adapted this rubric from the Colorado Student Assessment Program Alternate Level of Independence Performance Rubric. ISBE 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	369	47
12	300	53	300	54	300	97	300	64	300	46	300	72	400	26
13	300	53	300	54	300	97	300	64	300	46	300	72	418	18
14	356	53	349	54	300	97	317	64	366	46	300	72	428	14
15	391	29	385	30	300	53	359	35	397	25	342	39	435	12
16	411	20	405	21	322	37	384	25	414	18	370	28	440	11
17	422	16	417	17	343	30	398	20	424	15	385	22	444	10
18	429	14	425	14	358	26	409	18	432	13	396	19	448	9
19	435	12	431	13	370	23	416	16	437	11	405	17	451	9
20	439	11	436	12	379	21	423	14	442	10	412	16	453	8
21	443	10	441	11	387	20	428	13	445	9	417	14	456	8
22	447	10	444	10	394	18	433	12	449	9	422	14	458	8
23	450	9	448	10	400	17	437	12	452	8	427	13	461	8
24	452	9	451	9	405	17	440	11	454	8	431	12	463	7
25	455	8	453	9	410	16	444	11	457	8	434	12	465	7
26	457	8	456	8	415	15	447	10	459	7	438	11	467	7
27	459	8	458	8	419	15	450	10	461	7	441	11	469	7
28	461	8	460	8	423	14	453	10	463	7	444	11	471	7
29	463	7	462	8	427	14	455	10	465	7	447	11	473	7
30	465	7	464	8	431	14	458	9	466	7	450	10	475	7
31	467	7	466	8	434	14	460	9	468	7	453	10	477	7
32	468	7	468	8	438	14	462	9	470	7	455	10	479	8
33	470	7	470	7	441	13	465	9	472	7	458	10	481	8
34	472	7	472	7	445	13	467	9	473	6	460	10	484	8
35	473	7	474	7	448	13	470	9	475	6	463	10	486	8
36	475	7	476	7	451	13	472	9	477	6	466	10	489	9
37	477	7	478	7	455	13	474	9	478	6	468	10	492	9
38	479	7	480	7	458	13	477	9	480	7	471	10	495	10
39	480	7	481	8	462	14	479	9	482	7	473	10	499	11
40	482	7	483	8	465	14	482	9	483	7	476	10	504	12
41	484	7	485	8	469	14	484	9	485	7	479	11	511	14
42	486	7	487	8	472	14	487	10	487	7	482	11	519	17
43	488	8	489	8	476	15	489	10	489	7	485	11	535	25
44	490	8	492	8	480	15	492	10	491	7	488	11	565	46
45	492	8	494	8	485	15	495	10	493	7	491	12		
46	494	8	497	9	489	16	498	11	495	8	495	12		
47	497	9	499	9	494	17	502	11	498	8	498	13		
48	499	9	502	10	499	17	505	12	501	8	503	13		
49	503	10	505	10	506	19	510	12	504	9	507	14		
50	506	10	509	11	513	20	514	13	507	10	513	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	510	11	514	12	521	22	520	15	511	11	519	16		
52	515	13	519	13	531	24	527	16	516	12	526	18		
53	522	15	526	15	544	28	535	19	522	14	536	21		
54	531	19	536	19	562	35	548	23	531	17	550	26		
55	548	27	554	28	595	51	570	34	547	24	574	38		
56	580	52	587	53	655	95	610	63	576	45	619	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	302	73	389	39	300	98	339	57	300	74	350	55
16	312	51	350	40	415	21	300	54	377	31	339	41	386	30
17	347	36	377	28	430	15	330	38	399	22	368	29	407	21
18	366	28	392	22	438	12	351	31	411	18	384	23	419	17
19	379	24	402	19	444	10	366	26	420	15	396	20	427	15
20	389	21	410	17	449	9	378	23	427	14	404	18	434	13
21	397	19	416	15	452	8	387	21	433	12	412	16	439	12
22	404	18	422	14	455	8	395	20	437	12	418	15	443	11
23	410	17	426	13	458	7	401	18	441	11	423	14	447	10
24	415	16	430	12	460	7	407	17	445	10	427	13	450	10
25	420	15	434	12	462	7	413	16	448	10	432	13	453	9
26	424	15	437	11	464	6	417	16	451	9	435	12	456	9
27	428	14	440	11	466	6	422	15	454	9	439	12	458	8
28	431	14	443	11	468	6	426	15	456	9	442	11	460	8
29	435	13	446	10	469	6	430	14	459	9	445	11	463	8
30	438	13	448	10	471	6	434	14	461	8	448	11	465	8
31	442	13	451	10	472	6	438	14	463	8	451	11	467	8
32	445	13	453	10	474	5	441	14	465	8	454	10	469	8
33	448	13	456	10	475	5	445	13	467	8	456	10	470	7
34	451	12	458	10	476	5	448	13	469	8	459	10	472	7
35	454	12	460	10	477	5	451	13	471	8	461	10	474	7
36	457	12	463	10	479	5	454	13	473	8	464	10	476	7
37	460	12	465	10	480	5	458	13	475	8	467	10	477	7
38	463	12	467	10	481	5	461	13	477	8	469	10	479	7
39	466	12	469	10	483	5	464	13	479	8	471	10	481	7
40	469	12	472	10	484	5	468	13	481	8	474	10	483	7
41	472	12	474	10	485	5	471	13	483	8	476	10	485	7
42	475	13	477	10	487	5	474	14	485	8	479	10	487	8
43	478	13	479	10	488	6	478	14	487	8	482	10	489	8
44	481	13	482	10	489	6	481	14	489	8	484	11	491	8
45	484	13	484	10	491	6	485	14	491	8	487	11	493	8
46	487	13	487	11	492	6	489	15	493	8	490	11	495	8
47	491	14	490	11	494	6	493	15	496	9	493	11	497	8
48	495	14	493	11	496	6	497	15	498	9	496	12	499	9
49	499	15	496	11	498	6	502	16	501	9	499	12	502	9
50	503	15	499	12	499	7	507	17	504	10	503	12	505	9
51	507	16	503	12	502	7	512	17	507	10	507	13	508	10
52	512	17	507	13	504	7	518	18	510	11	512	14	511	10
53	518	18	511	14	507	8	525	19	514	11	517	15	515	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	524	19	517	15	510	8	532	21	518	12	522	16	519	12
55	532	21	523	16	513	9	541	23	523	13	529	17	525	13
56	541	23	530	18	518	10	552	25	529	15	537	19	531	15
57	553	27	539	21	523	12	566	29	537	17	548	22	539	17
58	570	33	553	26	531	15	585	36	549	21	563	28	550	21
59	601	49	577	38	545	21	620	52	569	30	590	40	570	30
60	658	92	621	71	570	39	682	97	605	57	637	73	606	54

Science

Raw Score	Grade 4		Grade 7		Grade 11	
	Scale Score	SE	Scale Score	SE	Scale Score	SE
1	300	134	300	128	300	67
2	300	134	300	128	300	67
3	300	134	300	128	300	67
4	300	134	300	128	300	67
5	300	134	300	128	300	67
6	300	134	300	128	300	67
7	300	134	300	128	300	67
8	300	134	300	128	300	67
9	300	134	300	128	300	67
10	300	134	300	128	300	67
11	300	134	300	128	300	67
12	300	134	300	128	300	67
13	300	134	300	128	300	67
14	300	134	300	128	300	67
15	300	134	300	128	325	67
16	300	73	300	128	369	36
17	300	52	300	70	393	25
18	306	42	300	49	407	20
19	327	36	300	40	417	17
20	342	32	317	34	424	15
21	355	29	332	31	429	14
22	365	27	344	28	434	13
23	374	25	354	26	439	12
24	382	24	363	24	442	11
25	390	22	371	23	446	11
26	396	21	378	22	449	10
27	402	21	384	21	451	10
28	408	20	390	20	454	10
29	413	19	396	19	457	9
30	418	19	401	19	459	9
31	423	19	406	18	461	9
32	428	18	411	18	463	9
33	432	18	415	18	466	9
34	437	18	419	17	468	9
35	441	18	424	17	470	9
36	445	18	428	17	472	9
37	449	18	432	17	474	9
38	454	18	436	17	476	9
39	458	18	440	17	478	9
40	462	18	444	17	480	9
41	466	18	448	17	482	9
42	471	18	452	17	484	9
43	475	18	457	17	487	9
44	480	18	461	17	489	9
45	485	19	465	17	491	9
46	490	19	470	18	494	10
47	495	20	474	18	496	10
48	500	20	479	18	499	10
49	506	21	484	19	502	10
50	512	22	489	19	505	11
51	519	23	494	20	508	11
52	526	24	500	20	512	12
53	535	26	506	21	516	13

Raw Score	Grade 4		Grade 7		Grade 11	
	Scale Score	SE	Scale Score	SE	Scale Score	SE
54	545	28	513	22	521	14
55	556	30	520	23	527	15
56	570	34	528	24	534	17
57	588	40	537	26	542	20
58	615	49	548	28	556	25
59	661	71	560	31	579	35
60	700	132	575	34	621	66
61			594	40		
62			621	49		
63			668	69		
64			700	128		

Writing

Raw Score	Grade 3		Grade 5		Grade 6		Grade 8		Grade 11	
	Scale Score	CSE	Scale Score	SE	Scale Score	SE	Scale Score	SE	Scale Score	SE
1	300	103	300	59	300	91	300	93	300	85
2	300	103	300	59	300	91	300	93	300	85
3	300	103	300	59	300	91	300	93	300	85
4	300	103	300	59	300	91	300	93	300	85
5	300	103	300	59	300	91	300	93	300	85
6	300	103	300	59	300	91	300	93	300	85
7	300	103	361	59	300	91	300	93	303	85
8	336	56	400	33	331	50	323	51	359	46
9	374	39	424	23	367	36	359	36	391	32
10	395	32	437	19	388	29	380	30	409	26
11	411	27	447	17	403	26	395	26	422	23
12	423	25	455	15	415	23	407	23	432	21
13	433	23	461	14	425	22	417	21	441	19
14	442	22	467	13	434	21	425	20	448	18
15	450	21	472	13	442	20	433	19	455	17
16	457	21	477	12	450	19	440	19	461	17
17	465	21	482	12	458	19	447	18	468	17
18	472	21	486	12	465	19	453	18	474	17
19	480	21	491	12	473	19	460	18	480	17
20	488	21	495	12	480	20	467	19	486	17
21	496	22	500	13	488	20	474	19	493	18
22	505	23	505	13	497	21	482	20	501	19
23	515	25	511	14	507	23	490	22	509	21
24	527	27	518	15	518	25	500	24	520	23
25	541	31	526	18	532	28	513	27	533	27
26	561	37	538	21	551	34	531	33	552	33
27	596	54	558	30	583	48	562	48	585	47
28	659	101	594	57	640	89	619	90	642	85

APPENDIX C: Classification Consistency

Reading

R3

Level	R1	R2	R3	R4	True
R1	17.3	3.1	0.0	0.2	20.7
R2	1.4	21.5	7.0	0.8	30.8
R3	0.0	3.0	27.2	5.6	35.8
R4	0.0	0.0	1.8	10.9	12.8
Observed	18.8	27.7	36.1	17.5	100.0

R7

Level	R1	R2	R3	R4	True
R1	14.3	2.4	0.1	0.4	17.1
R2	1.1	11.2	5.8	0.9	19.0
R3	0.0	1.7	28.9	9.9	40.5
R4	0.0	0.0	1.8	21.6	23.4
Observed	15.4	15.3	36.5	32.8	100.0

R4

Level	R1	R2	R3	R4	True
R1	16.0	2.8	0.1	0.2	19.1
R2	1.4	16.6	7.0	0.6	25.7
R3	0.0	2.6	32.0	6.4	41.0
R4	0.0	0.0	1.9	12.3	14.2
Observed	17.5	21.9	41.0	19.6	100.0

R8

Level	R1	R2	R3	R4	True
R1	14.7	2.7	0.3	0.5	18.3
R2	1.3	8.6	5.5	0.8	16.1
R3	0.0	2.3	30.6	10.3	43.2
R4	0.0	0.0	2.1	20.4	22.4
Observed	16.0	13.6	38.5	31.9	100.0

R5

Level	R1	R2	R3	R4	True
R1	20.6	3.5	0.2	0.4	24.8
R2	1.9	12.4	5.6	1.8	21.7
R3	0.0	2.4	10.9	7.2	20.6
R4	0.0	0.0	2.8	30.1	32.9
Observed	22.5	18.4	19.5	39.6	100.0

R11

Level	R1	R2	R3	R4	True
R1	10.7	1.4	0.1	0.7	12.9
R2	0.9	14.6	7.0	2.8	25.3
R3	0.0	1.5	25.5	14.7	41.7
R4	0.0	0.0	1.9	18.2	20.1
Observed	11.6	17.5	34.4	36.4	100.0

R6

Level	R1	R2	R3	R4	True
R1	9.8	1.9	0.0	0.2	11.9
R2	0.9	16.5	5.1	1.0	23.5
R3	0.0	2.2	24.2	9.4	35.8
R4	0.0	0.0	2.6	26.2	28.8
Observed	10.7	20.6	32.0	36.7	100.0

Mathematics

M3

Level	M1	M2	M3	M4	True
M1	17.7	3.4	0.2	0.2	21.4
M2	1.3	13.3	5.8	0.7	21.1
M3	0.0	3.2	21.4	7.4	32.1
M4	0.0	0.0	3.2	22.2	25.4
Observed	18.9	19.9	30.6	30.5	100.0

M4

Level	M1	M2	M3	M4	True
M1	13.5	2.4	0.1	0.3	16.3
M2	1.1	13.3	6.4	0.7	21.4
M3	0.0	2.3	34.7	8.0	45.1
M4	0.0	0.0	2.0	15.3	17.3
Observed	14.6	17.9	43.2	24.3	100.0

M5

Level	M1	M2	M3	M4	True
M1	11.0	2.0	0.0	0.2	13.3
M2	0.8	17.1	5.8	0.5	24.3
M3	0.0	2.5	35.6	7.9	46.0
M4	0.0	0.0	1.8	14.7	16.5
Observed	11.9	21.7	43.2	23.3	100.0

M6

Level	M1	M2	M3	M4	True
M1	10.2	1.7	0.0	0.1	12.1
M2	0.8	10.1	4.5	0.6	16.1
M3	0.0	1.7	23.6	9.8	35.1
M4	0.0	0.0	3.0	33.7	36.8
Observed	11.0	13.5	31.1	44.3	100.0

M7

Level	M1	M2	M3	M4	True
M1	13.4	2.1	0.1	0.2	15.8
M2	1.0	8.8	4.7	0.4	14.9
M3	0.0	2.3	32.6	8.6	43.5
M4	0.0	0.0	2.8	22.9	25.7
Observed	14.5	13.2	40.2	32.2	100.0

M8

Level	M1	M2	M3	M4	True
M1	8.6	1.6	0.0	0.1	10.4
M2	0.8	14.0	5.2	0.6	20.5
M3	0.0	2.0	31.1	9.5	42.7
M4	0.0	0.0	3.2	23.2	26.5
Observed	9.4	17.6	39.5	33.4	100.0

M11

Level	M1	M2	M3	M4	True
M1	12.6	1.6	0.0	0.1	14.3
M2	1.0	9.6	5.1	0.3	16.1
M3	0.0	1.6	36.9	8.6	47.1
M4	0.0	0.0	3.4	19.0	22.4
Observed	13.6	12.9	45.5	28.0	100.0

Science

S4

Level	S1	S2	S3	S4	True
S1	10.5	1.4	0.0	0.1	12.0
S2	0.9	13.8	5.2	0.6	20.6
S3	0.0	2.4	21.5	9.2	33.1
S4	0.0	0.0	3.4	31.0	34.4
Observed	11.4	17.6	30.1	40.9	100.0

S7

Level	S1	S2	S3	S4	True
S1	9.2	1.4	0.0	0.1	10.6
S2	0.8	10.2	4.3	0.5	15.8
S3	0.0	1.5	21.0	9.7	32.2
S4	0.0	0.0	3.1	38.3	41.4
Observed	9.9	13.1	28.3	48.6	100.0

S11

Level	S1	S2	S3	S4	True
S1	9.9	1.3	0.0	0.3	11.4
S2	0.9	9.2	3.4	0.9	14.5
S3	0.0	1.4	24.1	11.0	36.5
S4	0.0	0.0	2.6	35.0	37.6
Observed	10.8	11.8	30.1	47.2	100.0

Writing

W3

Level	W1	W2	W3	W4	True
W1	11.4	2.2	0.0	0.3	14.0
W2	1.1	13.1	7.0	2.1	23.3
W3	0.1	2.3	17.4	10.0	29.8
W4	0.0	0.0	3.6	29.3	32.9
Observed	12.6	17.6	28.1	41.7	100.0

W8

Level	W1	W2	W3	W4	True
W1	10.3	3.1	0.6	1.0	14.9
W2	1.2	10.8	9.0	1.8	22.8
W3	0.0	2.1	31.8	11.8	45.6
W4	0.0	0.0	2.1	14.5	16.6
Observed	11.5	15.9	43.4	29.1	100.0

W5

Level	W1	W2	W3	W4	True
W1	8.4	1.7	0.1	0.3	10.4
W2	0.8	8.9	7.5	0.8	18.0
W3	0.0	2.0	47.9	9.4	59.4
W4	0.0	0.0	1.7	10.5	12.2
Observed	9.3	12.6	57.2	21.0	100.0

W11

Level	W1	W2	W3	W4	True
W1	10.2	2.1	0.2	0.6	13.2
W2	1.0	6.4	4.8	1.6	13.8
W3	0.0	1.3	21.2	15.4	38.0
W4	0.0	0.0	1.8	33.2	35.0
Observed	11.2	9.9	28.0	50.9	100.0

W6

Level	W1	W2	W3	W4	True
W1	11.4	2.7	0.1	0.4	14.6
W2	1.2	13.7	7.2	1.9	24.0
W3	0.0	2.5	22.9	10.8	36.1
W4	0.0	0.0	2.5	22.8	25.3
Observed	12.7	18.9	32.7	35.8	100.0

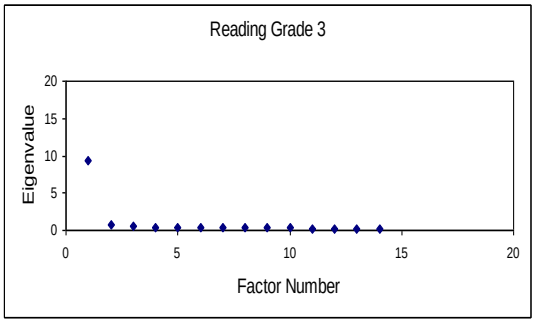
APPENDIX D: First Ten Eigenvalues from the Principal Component Analysis

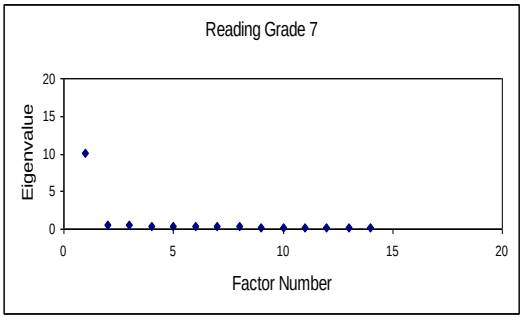
Grade	Number	Reading		Mathematics		Science		Writing	
		Eigenvalue	% Variance Explained	Eigenvalue	% Variance Explained	Eigenvalue	% Variance Explained	Eigenvalue	% Variance Explained
3	1	9.321	66.6	10.112	67.4			4.964	70.9
	2	0.685	4.9	0.704	4.7			0.571	8.2
	3	0.504	3.6	0.570	3.8			0.372	5.3
	4	0.461	3.3	0.508	3.4			0.356	5.1
	5	0.425	3.0	0.412	2.7			0.290	4.1
	6	0.390	2.8	0.392	2.6			0.282	4.0
	7	0.365	2.6	0.370	2.5			0.165	2.4
	8	0.346	2.5	0.357	2.4			-	-
	9	0.322	2.3	0.322	2.1			-	-
	10	0.290	2.1	0.300	2.0			-	-
4	1	9.089	64.9	10.312	68.7	9.668	64.5		
	2	0.723	5.2	0.910	6.1	1.024	6.8		
	3	0.593	4.2	0.531	3.5	0.512	3.4		
	4	0.533	3.8	0.483	3.2	0.464	3.1		
	5	0.464	3.3	0.369	2.5	0.449	3.0		
	6	0.380	2.7	0.339	2.3	0.402	2.7		
	7	0.353	2.5	0.302	2.0	0.372	2.5		
	8	0.345	2.5	0.296	2.0	0.342	2.3		
	9	0.316	2.3	0.278	1.9	0.324	2.2		
	10	0.304	2.2	0.257	1.7	0.298	2.0		
5	1	9.395	67.1	10.315	68.8			4.836	69.1
	2	0.659	4.7	0.666	4.4			0.493	7.0
	3	0.511	3.6	0.521	3.5			0.409	5.8
	4	0.473	3.4	0.467	3.1			0.366	5.2
	5	0.423	3.0	0.397	2.6			0.344	4.9
	6	0.388	2.8	0.368	2.5			0.290	4.1
	7	0.326	2.3	0.338	2.3			0.262	3.7
	8	0.324	2.3	0.319	2.1			-	-
	9	0.309	2.2	0.307	2.0			-	-
	10	0.297	2.1	0.283	1.9			-	-
6	1	9.575	68.4	10.664	71.1			5.051	72.2
	2	0.717	5.1	0.741	4.9			0.460	6.6
	3	0.515	3.7	0.484	3.2			0.382	5.5
	4	0.426	3.0	0.439	2.9			0.372	5.3
	5	0.405	2.9	0.361	2.4			0.287	4.1
	6	0.360	2.6	0.334	2.2			0.251	3.6
	7	0.323	2.3	0.302	2.0			0.197	2.8
	8	0.290	2.1	0.292	1.9			-	-
	9	0.279	2.0	0.265	1.8			-	-
	10	0.256	1.8	0.239	1.6			-	-
7	1	10.176	72.7	10.217	68.1	10.906	68.2		
	2	0.517	3.7	0.875	5.8	0.881	5.5		
	3	0.456	3.3	0.543	3.6	0.525	3.3		

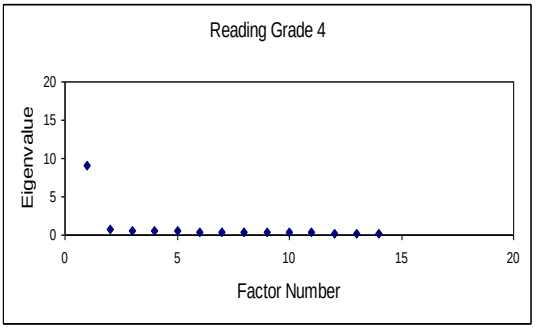
Grade	Number	Reading		Mathematics		Science		Writing	
		Eigenvalue	% Variance Explained	Eigenvalue	% Variance Explained	Eigenvalue	% Variance Explained	Eigenvalue	% Variance Explained
	4	0.371	2.7	0.445	3.0	0.469	2.9		
	5	0.359	2.6	0.437	2.9	0.435	2.7		
	6	0.323	2.3	0.350	2.3	0.396	2.5		
	7	0.307	2.2	0.306	2.0	0.363	2.3		
	8	0.277	2.0	0.301	2.0	0.322	2.0		
	9	0.259	1.9	0.275	1.8	0.296	1.9		
	10	0.231	1.6	0.265	1.8	0.257	1.6		
8	1	9.524	68.0	9.899	66.0			4.911	70.2
	2	0.769	5.5	1.014	6.8			0.615	8.8
	3	0.495	3.5	0.621	4.1			0.399	5.7
	4	0.451	3.2	0.515	3.4			0.332	4.7
	5	0.413	3.0	0.436	2.9			0.318	4.5
	6	0.370	2.6	0.389	2.6			0.296	4.2
	7	0.321	2.3	0.357	2.4			0.130	1.9
	8	0.311	2.2	0.300	2.0			-	-
	9	0.275	2.0	0.262	1.7			-	-
	10	0.253	1.8	0.241	1.6			-	-
11	1	8.841	80.4	10.184	67.9	11.428	76.2	5.531	79.0
	2	0.431	3.9	0.837	5.6	0.626	4.2	0.522	7.5
	3	0.331	3.0	0.597	4.0	0.430	2.9	0.321	4.6
	4	0.275	2.5	0.558	3.7	0.355	2.4	0.201	2.9
	5	0.229	2.1	0.442	2.9	0.327	2.2	0.174	2.5
	6	0.219	2.0	0.393	2.6	0.282	1.9	0.136	1.9
	7	0.190	1.7	0.315	2.1	0.238	1.6	0.115	1.6
	8	0.152	1.4	0.285	1.9	0.215	1.4	-	-
	9	0.136	1.2	0.277	1.8	0.201	1.3	-	-
	10	0.102	0.9	0.248	1.7	0.199	1.3	-	-

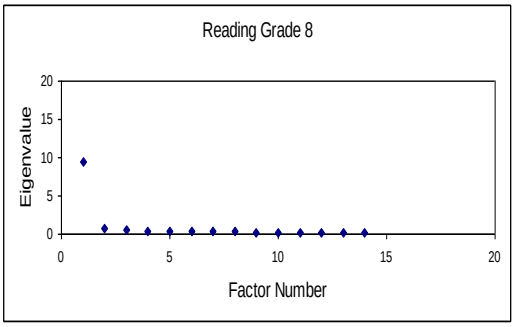
APPENDIX E: Scree Plots for All Components

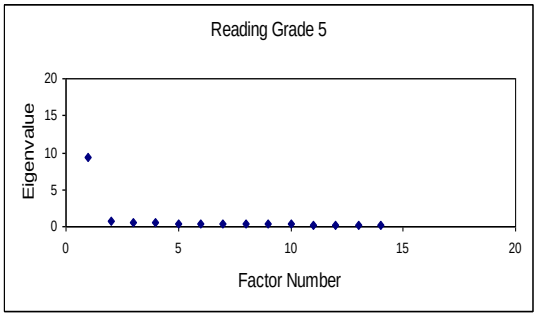
Reading

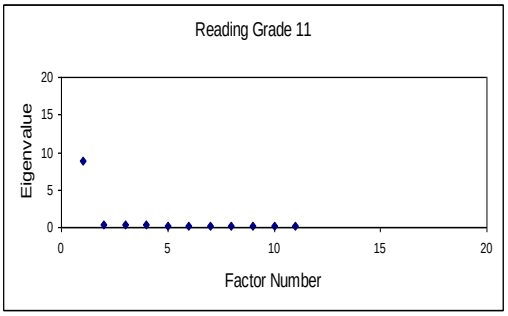


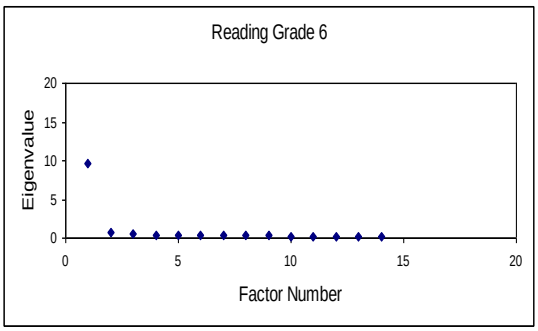




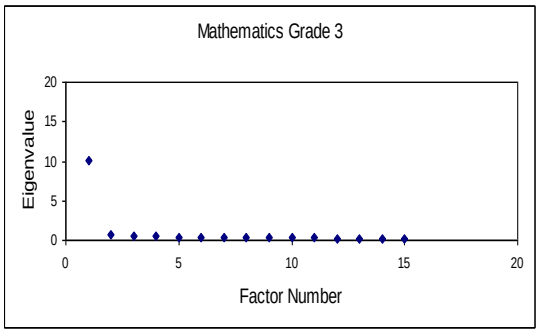


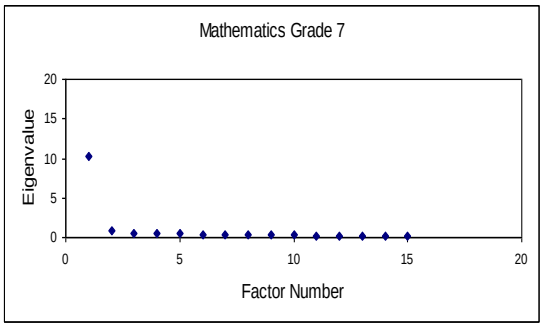


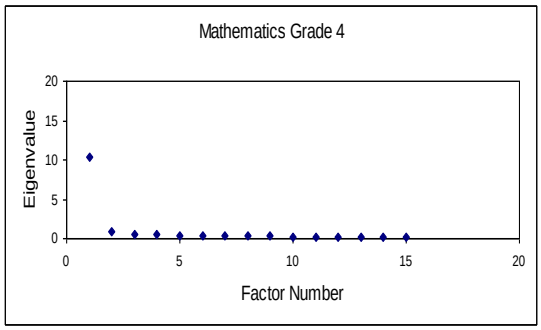


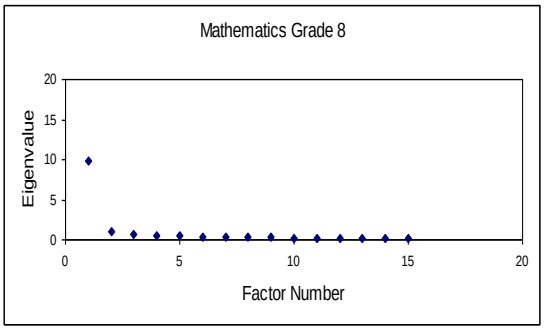


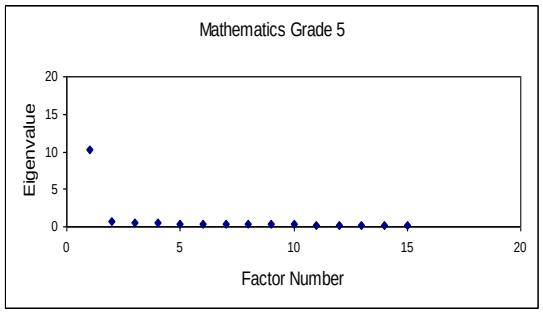
Mathematics

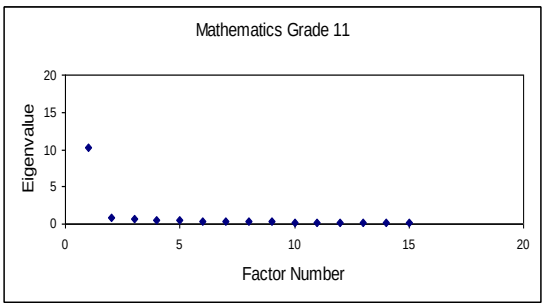


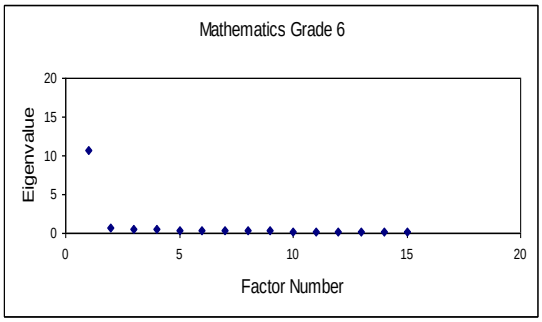




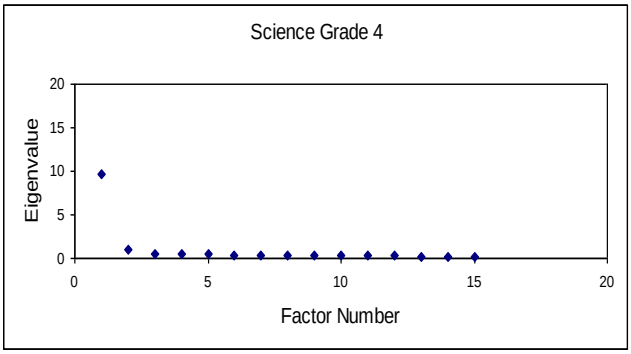


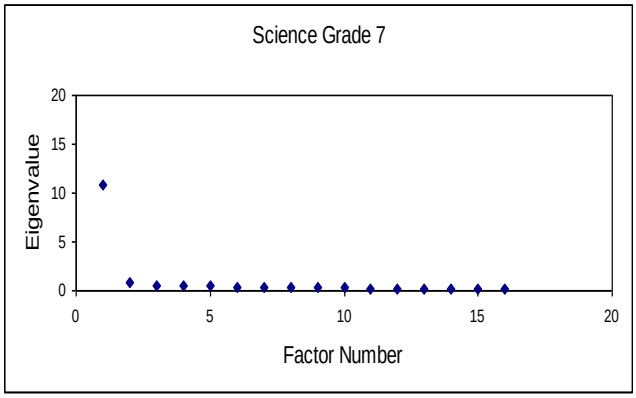


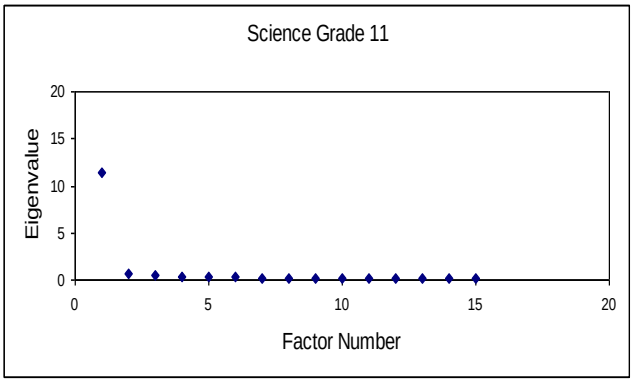




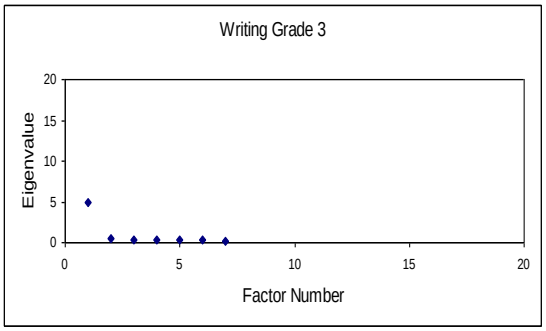
Science

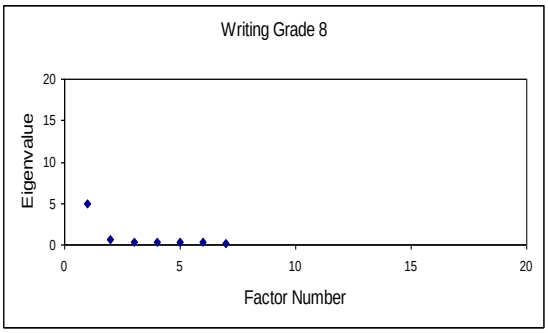

Science Grade 4
Eigenvalue
20
15
10
5
0
0
5
10
15
20
Factor Number

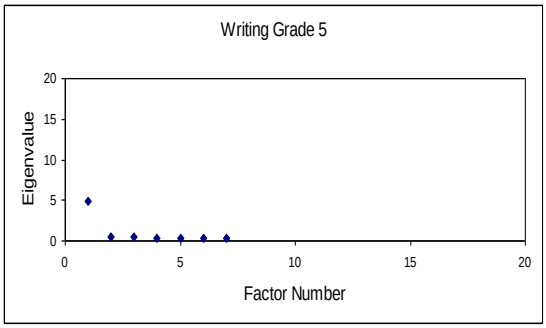

Science Grade 7
Eigenvalue
20
15
10
5
0
0
5
10
15
20
Factor Number

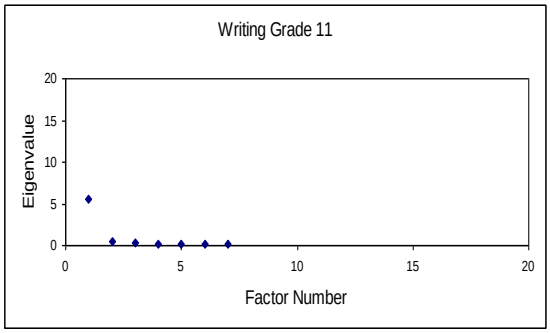

Science Grade 11
Eigenvalue
20
15
10
5
0
0
5
10
15
20
Factor Number

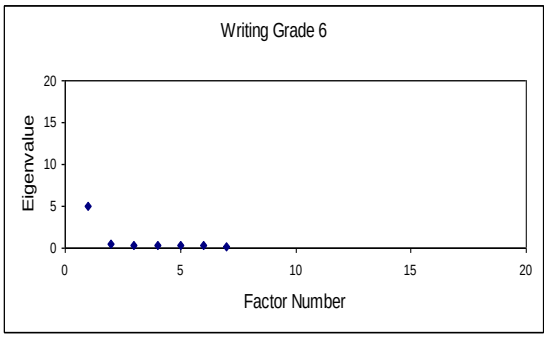
Writing


Writing Grade 3
Eigenvalue
Factor Number
0
5
10
15
20


Writing Grade 8
Eigenvalue
Factor Number
0
5
10
15
20


Writing Grade 5
Eigenvalue
Factor Number
0
5
10
15
20


Writing Grade 11
Eigenvalue
Factor Number
0
5
10
15
20


Writing Grade 6
Eigenvalue
Factor Number
0
5
10
15
20

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	10	100.0	0.0	100.0
	2	10	90.0	0.0	90.0
	3	10	100.0	0.0	100.0
	4	10	100.0	0.0	100.0
	5	10	90.0	10.0	100.0
	6	10	100.0	0.0	100.0
	7	10	100.0	0.0	100.0
	8	10	100.0	0.0	100.0
	9	9	100.0	0.0	100.0
	10	10	100.0	0.0	100.0
	11	9	100.0	0.0	100.0
	12	9	100.0	0.0	100.0
	13	9	100.0	0.0	100.0
	14	9	88.9	11.1	100.0
4	1	7	100.0	0.0	100.0
	2	7	100.0	0.0	100.0
	3	7	100.0	0.0	100.0
	4	7	100.0	0.0	100.0
	5	7	100.0	0.0	100.0
	6	7	100.0	0.0	100.0
	7	7	100.0	0.0	100.0
	8	7	100.0	0.0	100.0
	9	8	87.5	12.5	100.0
	10	8	75.0	25.0	100.0
	11	8	100.0	0.0	100.0
	12	8	87.5	12.5	100.0
	13	8	87.5	12.5	100.0
	14	8	100.0	0.0	100.0
5	1	13	100.0	0.0	100.0
	2	13	100.0	0.0	100.0
	3	14	100.0	0.0	100.0
	4	15	100.0	0.0	100.0
	5	15	100.0	0.0	100.0
	6	14	100.0	0.0	100.0
	7	14	100.0	0.0	100.0
	8	16	100.0	0.0	100.0
	9	16	100.0	0.0	100.0
	10	16	100.0	0.0	100.0
	11	16	100.0	0.0	100.0
	12	16	93.8	6.3	100.0

Grade	Item	Number of Students*	% of Exact Agreement	% of Adjacent Agreement	% of Total Agreement
	13	16	100.0	0.0	100.0
	14	16	100.0	0.0	100.0
6	1	16	100.0	0.0	100.0
	2	16	100.0	0.0	100.0
	3	16	93.8	6.3	100.0
	4	16	100.0	0.0	100.0
	5	16	100.0	0.0	100.0
	6	17	100.0	0.0	100.0
	7	17	100.0	0.0	100.0
	8	17	100.0	0.0	100.0
	9	17	100.0	0.0	100.0
	10	17	100.0	0.0	100.0
	11	17	100.0	0.0	100.0
	12	17	94.1	5.9	100.0
	13	17	94.1	5.9	100.0
	14	17	100.0	0.0	100.0
7	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
8	1	9	100.0	0.0	100.0
	2	9	100.0	0.0	100.0
	3	9	100.0	0.0	100.0
	4	8	100.0	0.0	100.0
	5	9	100.0	0.0	100.0
	6	9	100.0	0.0	100.0
	7	9	100.0	0.0	100.0
	8	9	100.0	0.0	100.0
	9	9	100.0	0.0	100.0
	10	9	100.0	0.0	100.0
	11	9	100.0	0.0	100.0
	12	9	100.0	0.0	100.0
	13	9	100.0	0.0	100.0
	14	9	100.0	0.0	100.0
11	1	37	100.0	0.0	100.0
	2	37	97.3	2.7	100.0
	3	37	100.0	0.0	100.0

Grade	Item	Number of Students*	% of Exact Agreement	% of Adjacent Agreement	% of Total Agreement
	4	37	100.0	0.0	100.0
	5	37	100.0	0.0	100.0
	6	37	100.0	0.0	100.0
	7	36	97.2	0.0	97.2
	8	36	97.2	0.0	97.2
	9	36	97.2	0.0	97.2
	10	35	97.1	0.0	97.1
	11	35	94.3	2.9	97.1

* 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	8	100.0	0.0	100.0
	2	8	100.0	0.0	100.0
	3	8	87.5	12.5	100.0
	4	8	87.5	0.0	87.5
	5	8	75.0	25.0	100.0
	6	8	100.0	0.0	100.0
	7	8	100.0	0.0	100.0
	8	8	100.0	0.0	100.0
	9	8	100.0	0.0	100.0
	10	8	87.5	12.5	100.0
	11	8	100.0	0.0	100.0
	12	8	100.0	0.0	100.0
	13	8	87.5	12.5	100.0
	14	8	87.5	12.5	100.0
	15	8	100.0	0.0	100.0
4	1	10	100.0	0.0	100.0
	2	10	100.0	0.0	100.0
	3	10	100.0	0.0	100.0
	4	10	100.0	0.0	100.0
	5	10	100.0	0.0	100.0
	6	10	100.0	0.0	100.0
	7	10	100.0	0.0	100.0
	8	10	90.0	10.0	100.0
	9	10	100.0	0.0	100.0
	10	10	100.0	0.0	100.0
	11	10	100.0	0.0	100.0
	12	10	90.0	0.0	90.0
	13	10	100.0	0.0	100.0
	14	10	100.0	0.0	100.0
	15	10	80.0	10.0	90.0
5	1	17	100.0	0.0	100.0
	2	16	93.8	6.3	100.0
	3	16	100.0	0.0	100.0
	4	16	100.0	0.0	100.0
	5	16	100.0	0.0	100.0
	6	16	100.0	0.0	100.0
	7	16	87.5	6.3	93.8
	8	16	100.0	0.0	100.0
	9	16	100.0	0.0	100.0
	10	17	94.1	0.0	94.1
	11	17	100.0	0.0	100.0
	12	17	94.1	5.9	100.0
	13	17	88.2	11.8	100.0
	14	17	94.1	5.9	100.0
	15	17	88.2	11.8	100.0

Grade	Item	Number of Students*	% of Exact Agreement	% of Adjacent Agreement	% of Total Agreement
	1	12	100.0	0.0	100.0
	2	12	100.0	0.0	100.0
	3	12	83.3	16.7	100.0
	4	12	100.0	0.0	100.0
	5	12	100.0	0.0	100.0
	6	12	100.0	0.0	100.0
	7	12	91.7	8.3	100.0
6	8	12	100.0	0.0	100.0
	9	11	100.0	0.0	100.0
	10	11	90.9	9.1	100.0
	11	11	100.0	0.0	100.0
	12	11	81.8	18.2	100.0
	13	11	100.0	0.0	100.0
	14	11	90.9	9.1	100.0
	15	11	100.0	0.0	100.0
	1	8	100.0	0.0	100.0
	2	8	100.0	0.0	100.0
	3	8	100.0	0.0	100.0
	4	8	100.0	0.0	100.0
	5	8	100.0	0.0	100.0
	6	8	100.0	0.0	100.0
	7	8	100.0	0.0	100.0
7	8	8	100.0	0.0	100.0
	9	8	100.0	0.0	100.0
	10	8	100.0	0.0	100.0
	11	8	100.0	0.0	100.0
	12	8	100.0	0.0	100.0
	13	8	100.0	0.0	100.0
	14	8	87.5	12.5	100.0
	15	8	87.5	12.5	100.0
	1	11	100.0	0.0	100.0
	2	11	100.0	0.0	100.0
	3	11	90.9	9.1	100.0
	4	10	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	8	11	100.0	0.0	100.0
	9	12	100.0	0.0	100.0
	10	12	100.0	0.0	100.0
	11	12	100.0	0.0	100.0
	12	12	100.0	0.0	100.0
	13	12	91.7	8.3	100.0
	14	12	100.0	0.0	100.0
	15	12	100.0	0.0	100.0
11	1	36	100.0	0.0	100.0
	2	37	97.3	2.7	100.0

Grade	Item	Number of Students*	% of Exact Agreement	% of Adjacent Agreement	% of Total Agreement
	3	37	100.0	0.0	100.0
	4	37	100.0	0.0	100.0
	5	36	97.2	2.8	100.0
	6	37	91.9	5.4	97.3
	7	37	97.3	2.7	100.0
	8	38	94.7	5.3	100.0
	9	38	97.4	2.6	100.0
	10	37	100.0	0.0	100.0
	11	37	100.0	0.0	100.0
	12	37	97.3	2.7	100.0
	13	37	97.3	0.0	97.3
	14	37	100.0	0.0	100.0
	15	37	94.6	2.7	97.3

* 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	6	100.0	0.0	100.0
	2	6	100.0	0.0	100.0
	3	6	100.0	0.0	100.0
	4	6	100.0	0.0	100.0
	5	6	100.0	0.0	100.0
	6	6	100.0	0.0	100.0
	7	6	100.0	0.0	100.0
	8	6	100.0	0.0	100.0
	9	6	100.0	0.0	100.0
	10	6	100.0	0.0	100.0
	11	6	100.0	0.0	100.0
	12	6	100.0	0.0	100.0
	13	6	100.0	0.0	100.0
	14	6	100.0	0.0	100.0
	15	6	100.0	0.0	100.0
7	1	12	100.0	0.0	100.0
	2	12	100.0	0.0	100.0
	3	11	100.0	0.0	100.0
	4	12	100.0	0.0	100.0
	5	11	90.9	9.1	100.0
	6	11	100.0	0.0	100.0
	7	11	90.9	9.1	100.0
	8	11	100.0	0.0	100.0
	9	12	100.0	0.0	100.0
	10	12	100.0	0.0	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	12	100.0	0.0	100.0
	15	12	100.0	0.0	100.0
	16	12	100.0	0.0	100.0
11	1	35	100.0	0.0	100.0
	2	35	100.0	0.0	100.0
	3	35	100.0	0.0	100.0
	4	35	100.0	0.0	100.0
	5	35	100.0	0.0	100.0
	6	35	100.0	0.0	100.0
	7	35	100.0	0.0	100.0
	8	35	100.0	0.0	100.0
	9	35	100.0	0.0	100.0
	10	35	97.1	2.9	100.0
	11	35	100.0	0.0	100.0
	12	35	100.0	0.0	100.0
	13	35	100.0	0.0	100.0

Grade	Item	Number of Students*	% of Exact Agreement	% of Adjacent Agreement	% of Total Agreement
	14	35	97.1	2.9	100.0
	15	34	97.1	2.9	100.0

* 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
3	1	10	100.0	0.0	100.0
	2	10	100.0	0.0	100.0
	3	10	90.0	10.0	100.0
	4	10	100.0	0.0	100.0
	5	10	100.0	0.0	100.0
	6	10	100.0	0.0	100.0
	7	10	100.0	0.0	100.0
5	1	12	100.0	0.0	100.0
	2	12	100.0	0.0	100.0
	3	12	100.0	0.0	100.0
	4	12	100.0	0.0	100.0
	5	13	92.3	7.7	100.0
	6	13	92.3	7.7	100.0
	7	13	100.0	0.0	100.0
6	1	9	100.0	0.0	100.0
	2	9	100.0	0.0	100.0
	3	9	100.0	0.0	100.0
	4	9	100.0	0.0	100.0
	5	9	100.0	0.0	100.0
	6	9	100.0	0.0	100.0
	7	9	100.0	0.0	100.0
8	1	10	100.0	0.0	100.0
	2	10	100.0	0.0	100.0
	3	10	90.0	10.0	100.0
	4	10	80.0	20.0	100.0
	5	10	100.0	0.0	100.0
	6	10	100.0	0.0	100.0
	7	10	100.0	0.0	100.0
11	1	37	94.6	5.4	100.0
	2	37	97.3	2.7	100.0
	3	37	100.0	0.0	100.0
	4	37	100.0	0.0	100.0
	5	37	97.3	2.7	100.0
	6	37	100.0	0.0	100.0
	7	37	100.0	0.0	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	3	-0.4771	0.6332	1.6716	56.55	464.19
	5	-0.4924	0.5795	2.4193	31.92	481.51
	6	-0.2150	0.9431	2.5322	49.51	453.31
	8	0.0980	1.0778	2.3781	50.54	445.53
	11	-0.1002	0.7935	2.4546	46.30	463.26

APPENDIX H: Item Statistics Summary

Reading

Grade		b-par	Infit	Outfit	Item Mean	Item-total Correlation
3	N of students	1442	1442	1442	1664	1664
	N of items	14	14	14	14	14
	Mean	0.08	1.03	0.97	3.22	0.70
	STD	0.23	0.08	0.15	0.14	0.03
	Minimum	-0.27	0.85	0.65	2.96	0.66
	Maximum	0.50	1.13	1.17	3.44	0.76
4	N of students	1688	1688	1688	1934	1934
	N of items	14	14	14	14	14
	Mean	0.04	1.05	0.97	3.27	0.68
	STD	0.31	0.15	0.28	0.16	0.06
	Minimum	-0.63	0.88	0.65	2.99	0.54
	Maximum	0.49	1.37	1.65	3.55	0.75
5	N of students	1601	1601	1601	1881	1881
	N of items	14	14	14	14	14
	Mean	-0.09	1.00	0.92	3.30	0.71
	STD	0.20	0.12	0.20	0.11	0.04
	Minimum	-0.53	0.82	0.61	3.11	0.63
	Maximum	0.25	1.25	1.26	3.54	0.77
6	N of students	1646	1646	1646	1978	1978
	N of items	14	14	14	14	14
	Mean	0.05	1.04	0.92	3.37	0.71
	STD	0.26	0.14	0.24	0.13	0.05
	Minimum	-0.42	0.88	0.53	3.18	0.56
	Maximum	0.42	1.42	1.50	3.61	0.75
7	N of students	1451	1451	1451	1915	1915
	N of items	14	14	14	14	14
	Mean	0.04	1.02	0.91	3.40	0.74
	STD	0.28	0.09	0.18	0.12	0.03
	Minimum	-0.44	0.91	0.58	3.23	0.66
	Maximum	0.47	1.23	1.22	3.61	0.77
8	N of students	1488	1488	1488	1867	1867
	N of items	14	14	14	14	14
	Mean	-0.02	1.05	0.92	3.45	0.69
	STD	0.31	0.13	0.21	0.13	0.05
	Minimum	-0.68	0.91	0.70	3.26	0.57
	Maximum	0.38	1.37	1.43	3.70	0.74
11	N of students	1086	1086	1086	1826	1826
	N of items	11	11	11	11	11
	Mean	0.12	1.03	0.88	3.53	0.77
	STD	0.28	0.16	0.27	0.09	0.04
	Minimum	-0.32	0.84	0.58	3.37	0.67
	Maximum	0.61	1.39	1.36	3.67	0.82

Mathematics

Grade		b-par	Infit	Outfit	Item Mean	Item-total Correlation
3	N of students	1473	1473	1473	1662	1662
	N of items	15	15	15	15	15
	Mean	0.06	1.02	0.91	3.23	0.70
	STD	0.36	0.08	0.17	0.23	0.05
	Minimum	-0.60	0.94	0.64	2.81	0.60
	Maximum	0.68	1.20	1.26	3.59	0.75
4	N of students	1628	1628	1628	1932	1932
	N of items	15	15	15	15	15
	Mean	0.01	1.04	0.89	3.36	0.70
	STD	0.36	0.09	0.19	0.20	0.05
	Minimum	-0.53	0.90	0.54	2.98	0.57
	Maximum	0.65	1.26	1.28	3.66	0.75
5	N of students	1592	1592	1592	1878	1878
	N of items	15	15	15	15	15
	Mean	0.06	1.03	0.92	3.35	0.71
	STD	0.35	0.10	0.17	0.19	0.04
	Minimum	-0.45	0.89	0.59	3.02	0.63
	Maximum	0.62	1.19	1.17	3.58	0.77
6	N of students	1687	1687	1687	1977	1977
	N of items	15	15	15	15	15
	Mean	0.05	1.02	0.89	3.40	0.73
	STD	0.35	0.14	0.22	0.18	0.05
	Minimum	-0.46	0.77	0.48	3.08	0.59
	Maximum	0.63	1.43	1.47	3.64	0.81
7	N of students	1647	1647	1647	1914	1914
	N of items	15	15	15	15	15
	Mean	-0.05	1.01	0.90	3.37	0.71
	STD	0.28	0.17	0.22	0.15	0.07
	Minimum	-0.86	0.83	0.68	3.13	0.52
	Maximum	0.32	1.38	1.38	3.72	0.78
8	N of students	1650	1650	1650	1864	1864
	N of items	15	15	15	15	15
	Mean	0.17	1.04	0.89	3.42	0.69
	STD	0.33	0.18	0.24	0.17	0.07
	Minimum	-0.45	0.81	0.64	2.99	0.51
	Maximum	0.89	1.45	1.41	3.67	0.77
11	N of students	1621	1621	1621	1828	1828
	N of items	15	15	15	15	15
	Mean	0.12	1.06	0.94	3.37	0.71
	STD	0.31	0.13	0.20	0.17	0.06
	Minimum	-0.37	0.87	0.61	3.06	0.57
	Maximum	0.60	1.34	1.31	3.58	0.78

Science

Grade		b-par	Infit	Outfit	Item Mean	Item-total Correlation
4	N of students	1767	1767	1767	1930	1930
	N of items	15	15	15	15	15
	Mean	0.04	1.04	0.91	3.29	0.68
	STD	0.27	0.14	0.21	0.17	0.07
	Minimum	-0.46	0.89	0.55	2.83	0.47
	Maximum	0.69	1.47	1.47	3.51	0.75
7	N of students	1699	1699	1699	1912	1912
	N of items	16	16	16	16	16
	Mean	0.14	1.03	0.90	3.41	0.72
	STD	0.33	0.11	0.20	0.16	0.05
	Minimum	-0.32	0.91	0.64	3.12	0.59
	Maximum	0.72	1.38	1.44	3.62	0.77
11	N of students	1330	1330	1330	1826	1826
	N of items	15	15	15	15	15
	Mean	0.08	1.00	0.86	3.49	0.76
	STD	0.28	0.13	0.20	0.12	0.04
	Minimum	-0.35	0.85	0.58	3.21	0.68
	Maximum	0.67	1.22	1.16	3.65	0.80

Writing

Grade		b-par	Infit	Outfit	Item Mean	Item-total Correlation
3	N of students	1309	1309	1309	1661	1661
	N of items	7	7	7	7	7
	Mean	0.07	1.10	1.03	3.20	0.69
	STD	0.37	0.09	0.11	0.16	0.03
	Minimum	-0.40	1.01	0.86	2.97	0.64
	Maximum	0.62	1.21	1.17	3.40	0.73
5	N of students	1413	1413	1413	1876	1876
	N of items	7	7	7	7	7
	Mean	0.02	1.02	0.98	3.29	0.69
	STD	0.19	0.10	0.14	0.10	0.03
	Minimum	-0.28	0.91	0.81	3.14	0.65
	Maximum	0.24	1.15	1.18	3.40	0.72
6	N of students	1484	1484	1484	1976	1976
	N of items	7	7	7	7	7
	Mean	0.14	1.03	0.96	3.28	0.71
	STD	0.31	0.11	0.12	0.13	0.03
	Minimum	-0.24	0.91	0.83	3.10	0.65
	Maximum	0.53	1.22	1.17	3.43	0.75
8	N of students	1270	1270	1270	1865	1865
	N of items	7	7	7	7	7
	Mean	0.04	1.00	0.92	3.44	0.68
	STD	0.20	0.19	0.27	0.08	0.06
	Minimum	-0.26	0.79	0.58	3.29	0.58
	Maximum	0.36	1.35	1.38	3.52	0.75
11	N of students	1157	1157	1157	1826	1826
	N of items	7	7	7	7	7
	Mean	0.17	1.01	0.87	3.44	0.76
	STD	0.17	0.25	0.29	0.11	0.08
	Minimum	0.00	0.77	0.59	3.20	0.60
	Maximum	0.52	1.54	1.45	3.52	0.82