TR 2008-01

Validity study:
A collection of evidence about the Ohio Graduation Tests.

Terrence Moore

Ohio Department of Education

May 9, 2008

Revision	Description	Revised by	Approval	Date
Int. Rel.	Edited by T. Moore	TWM		5.9.08
TR 2008 – 08				

Validity study:
A collection of evidence about the Ohio Graduation Tests.

Terrence Moore, Ohio Department of Education

Edited, 20 November 2007
Edited, 18 December 2007
- Revised Table I
- Revised Table 14
- Changed section headers to soften the wording
- Added strand designations to table E
- Corrected equation 1
- Revised part 4 (conclusion)
- Added SEM for dissimilar pairs (equations 3 & 4 plus discussion)
- Added Table 3 and Table 4 data for the ACT and the PLAN tests

Table of Content

0. Background. 4
0.1 Test Purpose. 4
0.2 Validation and validity. 5
0.3 Validity and AERA Standards. 6
0.4 Organization of this Document: Two views of validity: 7
PART 1 – Process related Validity 8
1.0 Processes Used to Create the Tests. 8
1.1 Development of Standards. 8
1.2 Item Development. 9
1.2.1 Item Review. 9
1.3 Test Blueprint. 11
1.4 Standard setting. 11
1.5 Scoring. 12
1.6 Equating and scaling. 13
1.7 Field Testing. 14
1.8 Capstone committee. 14
1.9 Appeals. 15
PART 2 – Empirical Studies Pertaining to the Validity of Inferences. 16
2. Overview. 16
2.1 Alignment. 16
2.2 Reliability, correlations and exploratory factor analysis. 17
2.3 Structural equation modeling (SEM). 20
2.3.1 Preparing the data file for SEM. 20
2.3.2 Congenaric analysis of the Ohio Graduation Tests. 20
2.3.3 2nd order confirmatory factor analysis models. 22
2.3.4 Regression models. 24
2.4 G Studies and Classification Consistency. 29
2.5 Test Success. 37
PART 3 – Discussion. 41
PART 4 - Conclusion. 42
References 43
Appendix A: Preliminary results of the Science Alignment Study 47
Appendix B: Constructing a data file 49
Appendix C: Congeneric analyses 51
Appendix D: Analysis of content areas by strand 57
Appendix E: 2nd Order Factor Analysis Models by Difficulty 67
Appendix F: Split of content standards 77
Appendix G: Regression models 82
Appendix I: Test Success Data 113

Two chief problems of the theory of knowledge are the question of meaning and the question of verification. The first question asks under what condition a sentence has meaning, in the sense of cognitive, factual meaning. The second one asks how we get to know something, how we can find out if a given sentence is true or false. (Carnap, 1936)

0. Background.

The Ohio Graduation Tests (OGT) are administered to Ohio high school students in the Tenth Grade as a prerequisite for the conferring of an Ohio High School Graduation Diploma. Students are required to pass all five portions of the OGT: (Reading, Mathematics, Writing, Science and Social Studies) and are provided with multiple opportunities to demonstrate proficiency in each content area. Also, according to state law, students that do not perform at the proficient level on the test will be provided with intervention to improve the student's ability in the content area where the student was judged to be performing below proficient.

Schools are also judged on the basis of the performance of students on the OGT. A school (or district) is judged to be proficient when 75% of the school's 10th graders pass the OGT. An additional indicator classifies an educational entity as having adequate performance when 85% of the entity's 11th graders have passed the test (ODE, 2007).

Because of the impact that OGT test results can have on schools and students, the inferences made from the results of the OGT need to be valid. In addition to students being denied diplomas, school attendance and school funding can be affected by the performance of examinees on the OGT.

0.1 Test Purpose.

The OGT testing program is used to determine which students will qualify for an Ohio diploma as a consequence of successful testing. Those who do not have test related successes are encouraged to obtain the necessary knowledge, skills and abilities (KSA's) to be successful on the OGT in a future test administration. One of the intentions of the testing programs is to inform students, parents and teachers of deficiencies in the KSA's of individual students. Because the student may continue to retake the OGT, remediation is one of the possible consequences of an unsuccessful attempt at one and all portions of the OGT. Many schools offer remediation in the immediate period following identification of students that have not demonstrated proficiency at the academic content of the OGT. As an inducement to providing remediation, students must attend 10 hours of remediation to qualify for the summer test administration.

While the use of the test as a measure of teacher performance is discouraged, the test results may be used to locate weaknesses related to instructional content and other factors that might influence how much students learn such as local instructional content and materials.

The results of the Ohio testing program are made public so that local officials and citizenry can take action to improve the performance of the schools or redirect students into alternative instructional programs either within a school or at an alternative school.

0.2 Validation and validity.

A good working definition of validity, for the purposes of this study, is provided by Messick (1993):

> "Validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretic rationales support the *adequacy* and *appropriateness of inferences* and *actions* based on test scores or other modes of assessment."

The purpose of this study is to present a collection of evidence (some of it empirical) as well as an explanation or argument (theoretic rationale) to assert that the Ohio Graduation Test is suitable for making valid inferences about the level of knowledge, skills and abilities that should be required of Ohio High School graduates and therefore the results of testing are also valid for making judgments about the efficacy of educational treatments at the schools Ohio students attend.

For the purposes of this paper, validation is the process of collecting evidence, analyzing evidence and making a judgment about the suitability of the tests for making inferences about student learning. Validity is asserted from the extent to which the inferences are thought to be valid based on the accumulation of evidence that weighs more heavily in favor of the inferences being valid than weighs in on the disconfirming collection of evidence. To some extent, the condition of validity will be asserted by the lack of evidence suggesting that the tests lack validity, a situation that is common in the testing of hypotheses in the social sciences.

0.3 Validity and AERA Standards.

The Standards for Educational and Psychological Testing (AERA, 1999) lists sources for validity evidence as shown in Table 0. The table briefly describes the applicability of the AREA document to the information provided in this document.

Table 0 – AERA listed sources for validity evidence.

AERA Source of Validity Evidence	**Correspondence in this document**
Evidence Based on Test content (AERA, p 11)	1.1 Development of Standards 1.2 Item Development 1.2.1 Item Review 2.1 Alignment 2.2 Reliability, Correlations, and Exploratory Factor Analysis
Evidence Based on Response Processes (AERA, p 12)	1.5 Scoring 2.2 Reliability, Correlations, and Exploratory Factor Analysis 2.4 G Studies and Classification Consistency
Evidence Based on Internal Structure (AERA, p 13)	2.3 Structural Equation Modeling (and subsections 2.3.2, 2.3.3, 2.3.4) 2.4 G Studies and Classification Consistency
Evidence Based on Relations to Other Variables (AERA, p 13)	
Convergent and Discrminant Evidence (AERA, p14)	2.3 Structural Equation Modeling (and subsections 2.3.2, 2.3.3, 2.3.4)
Test Criterion Relationships (AERA p 14)	1.1 Development of Standards 1.2.1 Item Review 1.4 Standard Setting 2.1 Alignment
Validity Generalization (AERA, p 15)	2.4 G Studies and Classification Consistency
Evidence Based on Consequences of Testing (AERA, p 16)	2.5 Test Success.

0.4 Organization of this Document: Two views of validity:

In this study, validity evidence has been collected and examined from two perspectives: "Part 1 – Process related Validity" and "Part 2 – Empirical Studies Pertaining to the Validity of Inferences." Part 1 lists many of the processes including processes used to define the domain of the tests, to set the performance standards, to qualify test items for the test, and the construction of the tests themselves. Part 2 provides descriptions of studies, computations and other data that represent reasoned inquiry into the quality of tests produced using the processes described in Part 1.

Parts 3 and 4 are for discussion and conclusions

PART 1 – Process related Validity

1.0 Processes Used to Create the Tests.

The processes used to develop the test are rich in the participation of stakeholders drawn from the community of grade appropriate educators thereby fulfilling Cronbach's (1989, p. 163, 1988, p. 6) notion of "Validation as a community process"[1]. Committees have been used to determine the curriculum to be assessed, to review items for content, to review items for fairness and sensitivity, to set standards, and to test the alignment of the tests with the curriculum. Community participation is not only integral in the processes used to develop and interpret the test, community participation increases the sense of fairness by giving voice to many faces as well as tapping into common understanding so that interpretation is central to the way that educators, parents and community members think about the construct of proficiency. Table 1 shows a list of some committees used in OGT processes.

Table 1 ~ Committees participating in activities integral to the Ohio Graduation Tests

Standards writing committees	Content advisory committees
Fairness and Sensitivity committees	Rangefinding committees
Standard Setting Committees	Technical Advisory Committees
Alignment Study Committees	

"Committees" is used in plural because the committees may be specific to a content area and because the committees may be formed for activities associated with a particular test administration. For example, a Rangefinding Committee will be convened for each subject and for each Spring administration; in servicing the OGT there have been more than a dozen Rangefinding Committees formed and quite possibly several dozen Rangfinding Committees have been formed. The work of these committees will be described in subsequent paragraphs.

1.1 Development of Standards.

Ohio embarked on the development of content based assessments by determining the content standards to describe the expectations that Ohio educators, parents, and community members have for what Ohio students should know and be able to do in the spring of their 10th grade year (also known as the "Common Expectations"). The process is described in the document, ACSDevelopmentProcess[2].pdf, (ODE, undated) and began in 1997 with the formation of six writing teams, one for each content area (the arts, English language arts, foreign languages, mathematics, science, and social studies). The initial emphasis was on describing the expectations for students exiting 12th grade, and

[1] According to Cronbach, "In science, technology, and policy alike, a theory succeeds by commanding widespread support in the relevant community, and the process is social as much as rational. Acceptance of constructions is inherently a community process…." This perspective seems especially relevant to content referenced or criterion referenced testing that lacks a variable to predict because both content and suitable performance are synthesized from the community of educators and those with a stake in educational outcomes (no matter how varied and unquantifiable those outcomes might be); the community is expert.

the expectations for successful transition from high school to either work or post secondary education.

The development of standards initially focused on English language arts and on mathematics using a model draft of content standards provided by the Ohio Department of Education (ODE). A joint council of persons representing the ODE, the Ohio Board of Regents (Ohio's post secondary education administration body) and composed of teachers, university professors, business advisors, and parents, revised the expectation into content standards complete with examples and benchmarked to grades 3, 5, 8 and 12. Public input was solicited through electronic posting of the draft standards and revisions were made. The standards for English Language Arts and for Mathematics were adopted by the Ohio Board of Education (a board comprised of elected officials and gubernatorial appointees) in December 2001.

Content standards were developed for Science and Social Studies using an advisory committee and a team for writing the content standards was formed based on the nominations from school districts, professional organizations, colleges and universities, community leaders , Educational Service Centers, Regional Professional Development Centers, teacher's unions and the Business Roundtable. The writing team reviewed standards available from national sources, international sources and other states. The team wrote standards, benchmarks and indicators for grades k though 12. Local and state professional organizations were invited to provide input for revision. Then the draft content standards were submitted to a public engagement campaign where electronic drafts were made available for comment and revisions were made as deemed appropriate by the writing committee. The Academic Content Standards for Science and Social Studies were adopted by the State Board of Education in December 2002.

The Academic Content Standards are available as both electronic and hardcopy versions. Because the Content Standards serve as the basis for the content of the Ohio Graduation Tests, teachers, parents and other interested parties can inform themselves of the content found on Ohio assessment instruments by studying the Academic Content Standards.

1.2 Item Development.

Test item development has been contracted out to test venders. ODE has written specifications (ODE, 2003(3); ODE, 2003(4); ODE, 2004(2); ODE, 2004(3)) that provide instructions to item developers on a benchmark by benchmark basis and include instruction on item type, stimulus attributes and response attributes. Item development specifications are a reflection of the Academic Content Standards. Item writers write with an eye toward assessing a particular benchmark in the content standards and identify the benchmark that is the target of the item.

1.2.1 Item Review.

Item development review involves collaboration between item writers, curriculum experts at ODE, outside content experts (typically college professors); a content advisory committee that includes classroom teachers, school administrators, district content

experts, and parents; and a fairness and sensitivity committee also composed of classroom teachers, school administrators, district content experts, and parents.

As an example of a committee, Table 2 summarizes the roster of participants in the Mathematics Content Advisory Committee for 6 July 2005.

Table 2 – Math Content Advisory Committee for 6 July 2005

Initials	Affiliation
M. G.	Coshocton High School
R. H.	South Euclid/Lyndhurst (Brush High School)
K. H.	Science/Math Network
D. H.	Morgan Local Schools
V. J-F.	Wright Patterson AFB
S. K.	Xenia High School
C. K.	Beachwood City Schools (Beachwood HS)
C. K.	Columbus Public Schools
S. K.	Champion Middle School
J. K.	Cleveland State University
M. L.	Columbus State Community College
S. M.	Toledo Public Schools
D. M.	Mansfield City Schools
S. P.	Western Local (Western HS)
K. P.	Tri-Village Local Schools (Tri-Village High School)
B. R.	Chippewa Local (Chippewa High School)
M. R.	University of Rio Grande (Higher Education)
T. R.	Ohio Northern University (Dept. of Mathematics)
J. R.	Cincinnati Public (Robert A Taft HS)
B. S.	Not Specified
B. S.	Preble County ESC
D. S-K.	Auburn Joint Vocational District (Auburn Career Center)
K. T.	Warren City Schools (Harding HS)
M. W.	Reading Community Schools (Reading JR./Sr. High)
R. W.	Willard City School District (Willard High School)

The committee shown in Table 2 is comprised primarily of high school teachers in the content area of Mathematics. There are several reasons for this committee to be weighted heavily toward classroom teachers. First, classroom teachers will be among those most familiar with the academic content standards and the meanings of the terms found in content standards. Teachers should be able to categorize test items into the educational rhetoric for the content area or identify those items that do not fit the content area. Second, because of their classroom experience, teachers can make good judgments and offer insight into the ways that students will interpret questions. Third, the use of the nomination process for forming committees is best known to school personnel making teachers among the most likely to be nominated for the committee. The nominating

process should have the effect of filtering out eager participants that might come to the committee with a particular ideological agenda that could make the committee less representative of Ohio communities than it might be otherwise.

1.3 Test Blueprint.

Tests are collections of test items. The selection of the items that appear on a given test is governed by the "Test Blueprint" (ODE, 2003; ODE, 2003(3); ODE, 2004; ODE,2006; ODE2006 (2)). Test blueprints cover the selection of items by standard and by benchmark including the number of total points, and the type of item (MC or multiple choice, CR or constructed response worth 2 points or 4 points, etc.) as well as the overall point total and some other aspects of item properties that need to be controlled. The purpose of the Test Blueprint is to ensure that every operational form is comparable to every prior operational form including the form used for the setting of performance standards. The test blueprint also assists in practices intended to rotate the selection of test items within the domain of all the content standards despite use of a limited length test.

Blueprint development was an informal committee activity intended to define structural features of tests so that the tests could be produced, so that the tests included point distributions that seems to be a fair expectation, so that tests produce point totals that could be used to report subscale performance and so that the tests provided a sufficient sample of the domain of the content area.

1.4 Standard setting.

Standard setting is the process where ability levels are defined for required performance deemed "proficient" for purposes of conferring an Ohio High School Diploma. If one thinks of assessment in the context of a) *what* is to be tested and b) *how much* performance indicates proficiency, the Academic Content Standards deal with the "what" of testing while standard setting deals with the "how much" aspect of the inferences to be made using the OGT. The standard setting process establishes the convergence between test items and performance expectations.

Standards were set in 2004 for Reading and Mathematics (Bunch, 2004) and in 2005 in Writing, Science and Social Studies (Bunch, Inman and Miles, 2005). For both standard setting activities, the method used was based on the bookmark method of Mitzel, Lewis, Patz and Green (2001).

The bookmark method of standard setting was performed using 24 panelists for Reading, 25 for Mathematics, 25 for Writing, 25 for Science and 21 for Social Studies. Panelists included teachers, school administrators, school board members, higher education representatives, as well as business and community leaders and parents. Effort was made to make the panels a cross-section of races and gender.

The standard setting process is a collaborative effort intended to give voice to participants and allow for mutual influence among the panelists. A three step process was used to allow the panelists first to make independent judgments, then to react to each other and

finally to react to the impact of the proposed ability levels on a large number of examinees.

Recommendations from the standard setting committees were adopted by the Ohio Board of Education for use in making inferences from operational test administration data.

1.5 Scoring.

The OGT uses both Selected Response (multiple choice or MC) and Constructed Response (or CR) test items. At present, all responses are scanned with the scanning machines making records of the responses of examinees to the selected response questions. Practices are in place to ensure that MC items are correctly keyed and that the scanning machines are accurately calibrated.

Constructed response items are read and scored by college educated and trained readers employed by test vendors or their subcontractors. CR scoring performance is monitored by the firm doing the scoring using a variety of techniques including the use of 100% second reads. When scores reported by the two readers are not adjacent (e.g. one assigns a score of 0 and the other assigns a score of 2) additional reading by another party brings the score back into adjacency. A test that has a total of six CR items will likely have been read, in part, by 12 or more readers. Arguably, the score from a single reader could make the difference between a student being judged proficient or not proficient. However, for that single reader to make such an impact, the scores from the other readers would need to place the other scores in the range of the proficient cut (along with scores on the multiple choice questions). Every student's score is a result from a community of readers.

There are a whole host of "business rules" for scoring practices including retraining and discharge of readers exhibiting poor consistency; to cover all those rules is beyond the scope of this paper.

Rangefinding is a technique where a committee reviews a set of example papers from a field test to define how points are to be assigned to reward the quality of examinee responses. Vendor scoring directors sift through as many as several thousand examinee responses to find papers that, in the experience of the scoring director, would constitute both discrete and transcendent scoring points to set the scoring scale. When a test question is written loosely, the scoring director will also select papers that represent the scope of the responses in addition to range of responses.

The rangefinding activity is a community process and the rangefinding committee may consist of 10 or more participants most of whom are classroom teachers but might also include administrators or content experts from higher education. While there is an interest is making the rangefinding committee diverse from the perspective of gender and race, rangefinding is not a political process but a technical process. The rangefinding process draws on the collective knowledge of the community of teachers.

Rangefinding committees are convened following collection of field test data and the results are used to score field test items and to provide a basis for operational scoring of the same items submitted to rangefinding.

1.6 Equating and scaling.

Ohio is using the Rasch single parameter equating model, a subset of the more general partial credit model (Wright and Masters, 1982; Masters, 1982, Wright and Stone, 1980). The implementation of the equating model is through the application of commercially available and extensively documented software (Linacre, 2005) under the name of WINSTEPS®. The use of commercially available software enhances the transparency of the equating process and ensures the software is available for public inspection. The use of software that is widely distributed also helps to ferret out problems that might be in the software code or problems of interpretation.

The primary advantage of using the Item Response Theory approach to equating of forms is the ease with which the ability levels defined as proficient during standard setting can be applied to test forms comprised of items that differ from the test form used for the standard setting process. Ohio practice is to set the mean difficulty of the standard setting form to zero and then calibrate all items to that same scale. When items are used from the bank, they can all serve as equating links because they have been calibrated to the same metric used to define proficiency.

The Spring test administration is the first test administered to the target group – Ohio 10th graders. The Spring administration test equating uses post equating and stability checks to ensure that items (or perhaps even passages) that perform in an uncharacteristic fashion are no longer considered to be calibrated to the same scale as the other items (Wright and Stone, 1979; Wright and Douglas, 1975). Summer and Fall test administrations utilize pre-equated test scoring tables without stability checks; the practice is considered appropriate because the Summer and Fall administrations are for students that did not demonstrate proficiency during the Spring administration and, therefore, the population of Summer and Fall examines is not representative of the population that was used to perform the original calibration.

Conversion from raw scores to ability is accomplished using the scoring table produced by WINSTEPS® or by arithmetic means. Conversion from the scoring table in the ability units expressed in logits to scaled score points is by means of an equation approved by the Ohio Board of Education subsequent to the standard setting process (Bunch, 2004; Bunch, Inman and Miles, 2005). Rounding of scaled scores to integer values for each possible raw score is accomplished using an Ohio Rounding Rule. Ohio has defined 400 scaled score points as proficient for all tests and uses an appropriate raw score multiplier to spread a test of 46 to 48 raw score points across a range of scaled score points from about 200 points to 600 points, depending on the subject and test administration. Scaled score points are used because they allow a fixed scale of 400 for the proficient cut, allow the other cuts to be expressed in a test invariant scale, are more consistent than raw score points and are easier for the public to understand than student abilities expressed in logits.

These processes have been approved by an independent board called the Technical Advisory Committee comprised of seasoned experts either from academia or because of past practitioner experience in large scale testing. The execution of steps associated with Equating and Scaling are performed by the vendor contracted for the OGT and are independently verified by an independent quality assurance firm and often replicated ODE personnel.

1.7 Field Testing.

Field testing is used to examine the suitability of new items for inclusion on an operational Ohio Graduation Test. Items are field tested two different ways: independent field tests and embedded field tests. The preferred method is the embedded field test where some new items are placed on operational forms without the knowledge of the examinee which items are operational and which are field test. Stand alone field test were administered to seed an item bank prior to the first operational administration and may be used in the future to add items to the item bank at a rate that cannot be provided through embedded field testing or to keep the operational test duration within some reasonable time limit.

Field test performance of each new item is inspected for properties that indicate the item is performing as intended. A field test item will only be placed into the item bank after it has been calibrated, found to have predictive validity (by exhibiting a point biserial correlation of .25 or greater) and is thought to be a fair and accurate measure of ability across race and gender. Field test items can be rejected, emended for retest, admitted to the item bank or passed before the fairness and sensitivity committee a second time.

All operational items on the OGT have been field tested.

1.8 Capstone committee.

Ohio engages a Technical Advisory Committee to provide advice on policies and practices. The Technical Advisory Committee is comprised of academics and practitioners. The academics are well published in the field of educational assessment while the practitioners have been associated with operational testing at local, state, and/or national levels and are well seasoned in the practices and problems of large scale assessment.

The duties of the TAC include the review of technical reports and the providing of advice on the adoption of administrative policy and technical aspects of administration, computation, and reporting associated with the OGT.

The TAC typically meets for face to face meetings of a day or two in duration on one or two occasions per year. Additionally the TAC will telephone conference as proper. For example, in the immediate aftermath of an administration, draft technical reports will be transmitted to the TAC, the TAC members review the reports and then telephone conference about the content of the reports. TAC members can also be used for extemporaneous consulting on issues relevant to the testing program.

The use of a TAC has several advantages for Ohio. Many of the TAC members serve or have served in an advisory capacity for other states and other large scale testing programs. Therefore they have an awareness of issues that have been either advantageous or detrimental to other large scale assessment programs. As a committee, the TAC is long in the tooth and brings wisdom that can only be obtained through years of cumulative experience. The TAC also functions as a board of peers where each of the members has a respect for other members often based on expertise demonstrated through the process of peer review publication. It should be noted that the responsibility of the final adoption of recommendations of the TAC is at the discretion of the Ohio Department of Education. This places a firewall between the TAC and the duties of actually administering the tests thus freeing up TAC to express thoughts without the prejudicial adherence to past operational practices. Finally, it should be noted that the TAC functions independently of vendors that the state employs to write test items, construct tests, score tests and generate performance reports for parents, school and districts and for the State. Test vendors often employ personnel that have considerable technical skills and the TAC provides guidance and feedback on the proposals the vendors make.

1.9 Appeals.

Ohio has a process whereby a score reported to a school or district may be appealed. The basis of the appeal could be over some procedural issue or over a belief (or disbelief) related to a reported score. Most of the appeals are related to procedural matters such as submittal of answer documents with errors of student identification. For example, there will be instances where an examinee has a score reported in Reading but the student is believed to have taken the Mathematics portion of the test. An appeal will be processed where the examinee's answer document is reviewed and scored as though the responses were aimed at the Mathematics portion of the test. This problem might occur when an examinee recorded his or her responses in the wrong portion of the answer document …intending to respond to Mathematics but recording the responses in another section. The appeals process is an avenue to correct this and other types of errors.

There are occasions where an examinee will appeal a score on the belief that the reported score is close to a proficient level and a rescoring might produce a different result. There are a few instances where this type of appeal might be successful. The reason for an occasional success with the strategy is rooted in the ambiguity of scoring constructed response items. There is some amount of chance that an examinee score in the very immediate vicinity of the proficient score will be within a standard measurement error of a proficient performance; sometimes being classed as proficient as a matter of chance or classified as not proficient, also as a matter of chance.

The appeals process serves primarily to correct procedural errors such as use of an incorrect answer document so that valid inferences may be made about students.

PART 2 – Empirical Studies Pertaining to the Validity of Inferences.

2. Overview.

In Part 1, this paper outlined the practices that have been used, and in most cases continue to be used, for building and administering the Ohio Graduation Tests. Part 1 showed the participation of committees in several aspects of the development of the OGT. The participation of a diverse community of persons interested in education, including teachers, school administrators, academics, parents, and business leaders, helps to build tests from which valid inferences can be made. Part 2 is for confirmatory evidence.

According to Messick (1993, p14), "Inferences are hypotheses, and the validation of inferences is hypothesis testing." A similar theme appears in the Standards for Educational and Psychological Testing (1999, p9): "Validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by the proposed uses of tests." Part 2 will focus on the use of evidence for testing the hypothesis:

> H_0: The processes used in developing the OGT program and scoring tests result in an instrument suited for making valid inferences about the abilities of 10th grade examinees against Ohio's Academic Content Standards.

The evidence to be presented include the following:

1. Alignment studies: tests of whether the OGT's are an accurate reflection of Ohio's Academic Content Standards
2. A congeneric analysis of strands and a 2nd order factor analysis to show the internal consistency of strands. Both the congeneric analysis and the 2nd order factor analyses test whether the items found in a typical test covary in a manner that one might conclude that they are all related to an underlying construct (such as the mathematics ability a 10th grade student should have)
3. Structural equation regression models to test the convergent and divergent validity of the OGT content areas

2.1 Alignment.

Ohio has performed alignment studies on three of the five content areas of the OGT. Alignment studies were used to examine the extent to which the Ohio Graduation Tests in Reading, Mathematics, and Science are consistent with the Ohio's Academic Content Standards.

The method used for the alignment studies was developed as the Web Alignment Tool (WAT) provided by Dr. Norman Webb and the Wisconsin Educational Research Center at the University of Wisconsin (Webb, 2005). The WAT uses panels of active classroom teachers to classify test items in comparison with the Benchmarks found in the Academic Content Standards and, once again, invokes the concept of validity as a community process (Cronbach, 1989) because of the use of panelists.

Results of the alignment studies in Reading and Mathematics (Keene, 2006) have been submitted under NCLB and are associated with Ohio's unconditional approval (Johnson, 2006) through the US Department of Education's (or Ed's) peer review process (still another example of a community process). Results of the Science alignment study are being analyzed and, at this time, look comparable to the results of the alignment study in Mathematics. The Ed has not called for the submission of data for Science but that directive is anticipated in the next several months.

Preliminary results of the Science Alignment Study are summarized in Appendix A. The results (Table A5) indicate that alignment for the Science portion of the test is comparable to the results of the alignment study for the Reading and Mathematics portions of several Ohio Tests and approved by the Ed.

2.2 Reliability, correlations and exploratory factor analysis.

Data from the Spring 2006 administration was chosen for this study because the spring administration is more representative in showing what students in the spring of their sophomore year know and can do. Administrations occurring in the Summer and Fall are for examinees that did not have success during a prior Spring administration and are less representative of Ohio 10th graders.

The test questions comprising each content area exhibit internal consistency as shown by the relatively high Cronbach alpha statistics as shown in Table 3.

Table 3 – Cronbach alpha statistics for the five content areas on the Ohio Graduation Tests[2]

	Test Reliabilities	
Content Area	**OGT**	**ACT[3]**
Reading	.829	.85
Mathematics	.886	.91
Science	.850	.80
Social Studies	.861	n.a.
Writing	.711	.64[4]

N_{OGT}=112,171

The ratio of the estimated true variance to the observed variance, Cronbach's alpha, (Nunnally and Bernstein, 1994; pp. 212) within content areas seems high when using alpha to measure test reliability. Only Writing falls below 0.8 and, presumably, Writing does so because there are fewer items on a Writing test (Writing has 14 items as compared to around 38 for tests in other content areas). Data for a well accepted national test, the ACT, are shown for comparison. The ACT test has 40 to 75 MC items per content area except for ACT Writing, a 30 minute single essay test. OGT reliabilities seem reasonable when compared to the ACT.

[2] Kane (2006, p. 21) includes reliability as a component of a validity study.
[3] ACT, 2007. Page 42. & ACT, 2005. Page 2. ACT does not have a Social Studies component.
[4] This data are for the ACT Writing test, ACT 2005, page 2. ACT assesses a writing sample using the Writing test and assesses mechanics using the multiple choice English portion of the ACT test.

Table 4 shows the correlations between scores for the five content areas of the OGT.

Table 4a – Correlations between scores for a large sample of examinees on the Spring 2006 Administration of the OGT.

	Reading	Mathematics	Science	Social Studies	Writing
Reading	1	.679	.727	.763	.684
Math	.679	1	.781	.729	.612
Science	.727	.781	1	.807	.604
Social Studies	.763	.729	.807	1	.624
Writing	.684	.612	.604	.624	1

Note: All Correlations are significant at the 0.01 level (2-tailed).
N=112,171

Table 4b shows the correlations between content areas for the five OGT tests (above the diagonal) and four ACT tests below the diagonal.

Table 4b – Comparison of the OGT correlations of Table 4a to the ACT. ACT correlations are in bold and below the diagonal.

	Reading	Mathematics	Science	Social Studies	Writing
Reading	1	.679	.727	.763	.684
Math	**.64**	1	.781	.729	.612
Science	**.69**	**.75**	1	.807	.604
Social Studies	**n.a.**	**n.a.**	**n.a.**	1	.624
Writing / **English**	**.78**	**.70**	**.70**	**n.a.**	1

In general, the correlations within the OGT are comparable to the correlations within the ACT. For example, Science and Reading are more highly correlated than Mathematics and Reading. Writing (English) and Reading are more highly correlated than Writing and either Mathematics and Science.

The differences in the correlations for Writing/English with other subjects between the ACT and the OGT may reflect structural differences in the tests. The OGT Writing test includes both a written component and a multiple choice component while the ACT English test includes only multiple choice items.

A similar comparison can be made using the PLAN test. The PLAN test, although once commonly uses as a practice test for the ACT, has been reframed and validated as a test for the growth in skills of tenth grade examinees after controlling for the skills of an eighth grader. Although the content standards for the three tests, the OGT, the ACT and the PLAN, may differ, the tests are aimed at almost the same grade level and should assess similar sets of skills.

Table 4c shows the correlations between content areas for the five OGT tests (above the diagonal) and four PLAN tests below the diagonal. The patterns of correlations within the PLAN test, the ACT test, and the OGT tests are similar. The similarities persist despite important differences in the tests including the presence or absence of constructed response items, different or slightly different numbers of items on the tests (for example, the ACT English test is 75 items in 45 minutes; ACT Reading is 40 items in 30minutes, ACT, 2007. p. 5-7), and the ACT and PLAN are for a more self selected group that could affect skills and motivation as well as countless other aspects of testing.

Table 4c – Comparison of the OGT correlations of Table 4a to the PLAN Test. PLAN correlations[5] are in bold and below the diagonal.

	Reading	Mathematics	Science	Social Studies	Writing
Reading	1	.679	.727	.763	.684
Math	**.62**	1	.781	.729	.612
Science	**.68**	**.69**	1	.807	.604
Social Studies	**n.a.**	**n.a.**	**n.a.**	1	.624
Writing / **English**	**.74**	**.69**	**.69**	**n.a.**	1

While the data of Table 3 indicate that, in general, the test items seem to covary with a common latent variable, the data does not indicate if the common variation among items in one test is any different than items found in some other content area (see Table 4). For example, is Mathematics a test of Reading skills more than a test of Mathematics skills? The data show that item scores in each content area covary with item scores in other content areas. Taken together (Tables 3 and 4) the saliency of academic content areas is lost in data that shows relatively high correlations across content areas.

In an attempt to find saliency in the content areas, an exploratory factor analysis (EFA) was performed. Out of the 171 components in the analysis, one component had an eigenvalue of 27.6 followed by a 3.5, a 2.4, and a 2.1. Altogether, there were 19 components with eigenvalues exceeding 1 supporting the notion that the tests have a moderate one-dimensional character (some of the higher factor loadings for the first component are associated with constructed response items). However the EFA did little to support the notion that the tests are effectively measuring salient content areas otherwise there would have been five substantial components, one for each content area. Another technique was needed to show that content areas are measuring different abilities.[6]

[5] Roberts & Noble, 2004. Page 28

[6] Dr. Tom Hirsch has suggested that factor analysis using strand totals and rotated factor loadings has produced some evidence supporting internal structure. This may be attempted at some later time.

2.3 Structural equation modeling (SEM).

LISREL was chosen as the software tool for structural equation models because it has the ability to process the dichotomous scored (multiple choice) and polytomous scored (constructed response) items using polychoric correlations. While other software packages can also compute polychoric correlations, LISREL seems to have an unlimited capacity to compute large matrices and was found to compute over 13,000 correlations for 167 variables in a single analysis for one trial. LISREL also has the advantages of being senior among SEM software competitors, well documented from a user perspective, and ubiquitous in the literatures where SEM is commonly used. It also is a relative bargain to purchase and comparatively friendly to use.

2.3.1 Preparing the data file for SEM.

Data for the study were from the Spring 2006 OGT administration. The data were provided in files for each content area. The five files were concatenated by matching of lithocodes. Lithocodes are unique identifiers for each answer document; matching across content areas collects a single student's performance results into a single cross content record. After concatenation, incomplete records (i.e. fewer than all five content areas) and records for examinees not classed as 10th grade were jettisoned.

Data preparation included the reduction of the number of observations from about 125,000 to a more manageable 10,000 records. The data were sampled down using a 14 cell typology that matched the sample to the district enrollment by school typology using the ODE classification system (7 classes) and to match the sample for proportions of white and non-white examinees (2 classes). Additional details are provided in Appendix B.

2.3.2 Congeneric analysis of the Ohio Graduation Tests.

The congeneric model tests the fit of the data to examine if all of the measures (or test items) in a single content area are of the same genus and assess the same construct (Byrne, 1998, p94). The congeneric model is a single factor model and provides less structure to fit the data than some alternative models (e.g. a 2nd order factor analysis model). Therefore, it may be more difficult to get a good fit with the congeneric model. A discussion of the congeneric model is available in Jöreskog and Sörbom, 1988.

Ohio uses a scoring system that resembles a single factor model, at the content level, because the performance standards are set for an entire content area instead of deconstructing performance into strands. For example, in Mathematics, performance standards are set for all of Mathematics instead of having separate performance standards for each strand (Number Sense, Patterns Functions and Algebra, Data analysis, etc.). Ohio scoring practice treats correct response points in one strand as equal to points that may have been earned on another strand; points is points, or so it seems, as points are interchangeable within a test but not across tests. So points earned in Number Sense may be "substituted" for points in Algebra because the they are portions of a singular mathematics ability but not for points in history (a strand in the Social Studies test).

The data are shown for the five congeneric models (one each for Reading, Mathematics, Writing, Science, and Social Studies) in Appendix C and summarized as shown in Table 5.

Table 5 does not include measures based on the χ^2 statistic despite the historical prevalence of the statistic for assessing model fit and both the reporting of the statistic in SEM (e.g. LISREL) and the mention of the statistic in venerated literature such as Bagozzi and Yi (1988). As pointed out by Gerbing and Anderson (1993), "…the χ^2 value tended to reject the null hypothesis of acceptable fit more than it should, even for models that were perfectly specified, a finding that only exacerbated the oversensitivity of this test for models with some, although perhaps minor, misspecification." The use of χ^2 as a test for these models will be exacerbated by the large number of degrees of freedom. Therefore, χ^2 is not presented as a criterion in Table 5.

Table 5 – Evaluation of the Congeneric (Single Factor) models for OGT Subject Tests.

	Measures of Model Quality (Byrne, 1989, p54-55; Byrne, 1998, p112-113; Bagozzi & Yi, 1988)				
Content Area	**Are all items significant?**	**AGFI ≥ 0.90 ?**	**RMR ≤ 0.050 ?**	**RMSEA ≤ 0.050 ?**	**# Negative error variances**
Reading	Yes	0.96	0.014	0.032	0
Mathematics	Yes	0.96	0.024	0.031	0
Writing	Yes	0.90	0.043	0.081	0
Science	Yes	0.98	0.020	0.023	0
Social Studies.	Yes	0.98	0.004	0.020	0

The data show that for the measures of Reading, Mathematics, Science and Social Studies, the criteria are met for acceptable modeling. Writing met four of the five criteria shown in Table 3. In examining the data, the error variances were typically much larger for constructed response items than for multiple choice items. Presumably, this difference is due to the discrepancy on the number of points assigned to some of the items. Most of the items are dichotomous but constructed response items could be worth 2, 4, 6, or even 12 points. The data show, not unexpectedly, that error variances tend to rise as the number of points on an item rise. For example, the two Writing Applications items worth 12 points each have error variances of 1.77 points and 1.85 points; the two Writing Conventions items (6 points each) have error variances of .53 and .82 points. The concentration of points in four items (36 out of 48 total points) on the Writing test may also explain why the RMSEA is higher than the other tests.

Considering the unique distribution of points on the Writing test, the data support the notion that combining all the points on a single test into a single indicator of ability is consistent with the data for the congeneric model. The use of the data to indicate a single measure of knowledge, skill and ability for each of the five OGT tests as well as a single performance standard for defining proficient performance in each of the five standards is supported by the data.

2.3.3 2nd order confirmatory factor analysis models.

The tests are presented to the public both as a single score and as scores in strands. For example, the is an overall Mathematics score as well as scores in Algebra, Data Analysis, Geometry, Measurement, and Number Sense.

Detailed data for the 2nd order confirmatory factor analysis is shown in Appendix D. Once the metric is set for a particular eta (η) variable, the paths for the remaining lambda-y (λ_y) variables are both positive and significant in all of the models.[7].

[7] In the models, the metric for each strand is set by the "first" item in the analysis and, therefore, that item's path coefficient is not tested for statistical significance.

A comparison of the congeneric model with the 2nd order factor analysis using content strands is shown in Table 6.

Table 6 – Comparison of the 2nd order factor model based on strand identity of the content to the congeneric model

	χ^2 Fit Statistics			
Test	**Congeneric**	**2nd order**	**Change in χ^2**	**Change in d.f.**
Reading	5995	5859	136	4
Mathematics	5543	5388	155	5
Writing	5586	4486	100	16
Science	3685	3515	170	4
Social Studies	3219	3181	38	4

While all of the changes in the χ^2 statistic in Table 6 are highly significant in a statistical sense, the overall fit of the data to the model is almost unaffected by using the 2nd order model instead of the congeneric model. This data suggests that while there is some dimensionality to the tests by strand, overall examinee performance on individual tests is not influenced much by skills that are *unique to particular strands*.

As an alternative to the strand based confirmatory factor analysis model, a 2nd order models were constructed by assigning test items into four levels of difficulty called "Hard," "Medium Hard," "Medium Easy," and "Easy" (see appendix E). Four levels were chosen because each content area has 4 (or so) strands. The difficulty based model can be compared to determine whether difficulty or strands provide a superior fit the data to the model.

Table 7 – Comparison of the difficulty based model to the 2nd order factor model

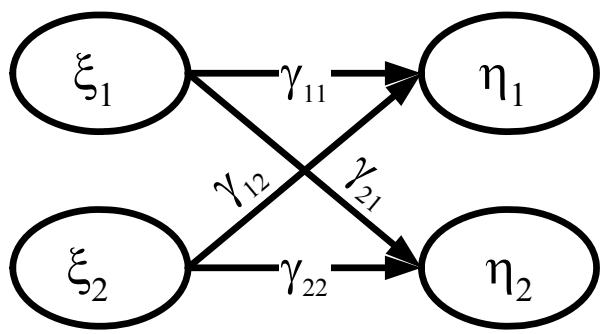
	χ^2 Fit Statistics			
Test	**2nd order based on strands**	**2nd order based on difficulty**	**Change in χ^2**	**Change**
Reading	5859	4328	1531	- 26 %
Mathematics	5388	4145	1243	- 23 %
Writing	4486	5186	- 700	+ 15 %
Science	3515	3287	228	- 6 %
Social Studies	3181	2786	395	- 12 %

The results of Table 7 show that difficulty grouping of items tends to produce a better covariance model than grouping by strand; with the exception of Writing where the strand based grouping offers a better means of organizing the variance in the model. This data suggest that despite the saliency of academic strands, examinee skill at a subject transcends the variation in skill associated with strands. Simply put, examinees that perform well on a test tend to do well with little regard to the strand.

2.3.4 Regression models.

Structural equation modeling was used to simultaneously examine the convergent and discriminant validity of the test measures (Reading, Mathematics, Writing, Science and Social Studies) used for the OGT. The perspective chosen for showing convergent and discriminant validity is that subsets of items from one test should be a better predictor of subset performance on that same test than a subset of test items from of a different content area. For example, Reading performance should be a better predictor of Reading performance than it is of Mathematics performance and the converse should also hold. The implementation of the perspective is accomplished by using a single administration where examinees have taken all five tests, then splitting each test into nominally equivalent forms. For example, a Mathematics Test of 48 points becomes two Mathematics Tests of 24 points each, etc. (see appendix F). If there is a distinct latent variable of Mathematics ability and a distinct latent variable of Reading ability, tests of Mathematics ability should predict performance on other tests of Mathematics ability better than tests of Reading ability predict tests of Mathematics ability. Figure 1 shows a diagram for this study.

Figure 1 – Diagram of a model for testing the discrimminant and convergent validity of two tests.



Note: the ξ (ksi) variables and the η (eta) variables that are subscripted the same are from the same full length test form.

The diagram of Figure 1 can be expressed as two simultaneous equations:

$$\xi_1 = \gamma_{1\text{-}2}\eta_2+\gamma_{1\text{-}1}\eta_1+b_1+e_1 \qquad \text{eq. 1}$$

$$\xi_2 = \gamma_{2\text{-}1}\eta_1+\gamma_{2\text{-}2}\eta_2+b_2+e_2 \qquad \text{eq. 2}$$

Convergent validity is supported when the coefficients $\gamma_{1\text{-}1}$ (gamma$_{1\text{-}1}$) and $\gamma_{2\text{-}2}$ are significant and large indicating that two test instruments (the hypothetical test forms split from the same subject test form) are measuring the same or similar latent variables (Angoff, 1988). Discriminant validity is supported when the coefficients $\gamma_{1\text{-}2}$ and $\gamma_{2\text{-}1}$ are small and insignificant showing that, in the presence of a predictor variable from the same focal test, a predictor variable from the test of a different subject is a poor predictor of the focal test and helps demonstrate discriminant validity evidence (Kane, 2006 p. 39).

Structural equation modeling was chosen to analyze the data for all five tests with a separate analysis for each dyadic pair of tests resulting in 10 models as shown in Table 8.

Table 8 – Dyadic pairs modeled to test for convergent and discriminant validity.

Model	Variable 1	Variable 2
1	Mathematics	Reading
2	Mathematics	Writing
3	Mathematics	Science
4	Mathematics	Social Studies
5	Reading	Writing
6	Reading	Science
7	Reading	Social Studies
8	Writing	Science
9	Writing	Social Studies
10	Science	Social Studies

The data for the study was prepared by splitting each form into two parametrically equivalent forms, each hypothetical form is half the length of the administered test form. As noted by Kane (2006, p. 36), "Convergent evidence may be based on correlations between different tests developed for this purpose or *between existing tests of the same label* (italic added)." Splitting a single test in halves is way of using existing tests of the same label.

The data of Table 9 show that for each split, the number of points are equal and the number of items are equal for all content areas except for Writing where the items (15 items in Writing) cannot be split equally. The items were also split by assigning item types equally to the two groups. For example, in Science, the two four point constructed response items were assigned to different groups. Then the two point constructed response items were assigned, in pairs of matched difficulty such that the sum of the nominal difficulties of the sets of items are driven toward equality. The practice of assignment continued into the multiple choice items always assigning pairs of comparable difficulty and always assigning between the two items to drive the two groups of items toward the same level of difficulty. It is hoped that this method of assignment produces groups of test items that have not only similar difficulty and similar structure but also similar variances of difficulties. The results of splitting the items in each content area are summarized in Table 9 and detailed in Appendix G.

Table 9 – Comparison of the points, items and difficulties of the split items from the Spring 2006 administration

	Group 1			Group 2		
Area	**Number of points**	**Number of items**	**Mean Rasch**	**Number of points**	**Number of items**	**Mean Rasch**
Reading	24	19	-.0233	24	19	-.02502
Mathematics	23	19	.042091	23	19	.09
Writing	24	7	-.03789	24	8	-.04
Science	24	19	.085614	24	19	.084682
Social Studies	24	19	.075523	24	19	.082361

The coefficients for each dyad in table 8 and equations 1 and 2 are shown in Table 10.

Table 10a – Results of convergent and discriminanat validity comparisons of dyads of tests

Model	Variable 1	Variable 2	$\gamma_{1\text{-}1}$	$\gamma_{1\text{-}2}$	$\gamma_{2\text{-}2}$	$\gamma_{2\text{-}1}$
1	Mathematics	Reading	0.99 (28.79)	0.018 (1.4)	1.06 (34.64)	-0.051 (-2.93)
2	Mathematics	Writing	0.99 (28.79)	0.022 (2.17)	1.51 (49.03)	-0.33 (-12.07)
3	Mathematics	Science	1.05 (24.1)	-0.016 (-.62)	1.17 (15.62)	-0.12 (-3.22)
4	Mathematics	Social Studies	1.27 (20.65)	-0.27 (-5.70)	1.33 (27.14)	-0.35 (-9.86)
5	Reading	Writing	1.04 (32.95)	-0.024 (-1.42)	1.92 (28.95)	-0.77 (-11.85)
6	Reading	Science	1.15 (28.83)	-0.15 (-5.13)	1.16 (16.53)	-0.16 (-5.28)
7	Reading	Social Studies	1.09 (25.18)	-0.079 (-2.23)	1.05 (26.15)	-0.059 (-2.09)
8	Writing	Science	1.58 (44.22)	-0.40 (-12.28)	1.10 (17.46)	-.10 (-6.71)
9	Writing	Social Studies	1.59 (43.61)	-0.42 (-12.48)	1.02 (33.11)	-0.026 (-2.15)
10	Science	Social Studies	1.26 (13.75)	-0.25 (-4.13)	0.98 (21.41)	0.020 (.56)

Note: t statistic shown, in parentheses, below path coefficient.

The results shown in Table 10a provide estimates for 20 equations (two equations per dyadic model in Table 8) and forty parameter estimates. The hypothesis found at the beginning of Part 2[8] is supported when one split half of a test is a predictor variable for

[8] H_0: The processes used in developing the OGT program and scoring tests result in an instrument suited for making valid inferences about the abilities of 10th grade examinees against Ohio's Academic Content Standards.

the other half because the coefficients (i.e. $\gamma_{1\text{-}1}$ or $\gamma_{2\text{-}2}$) are positive, significant, and large. In looking at the data of Table 10, this condition is met for 20 of the 20 possible cases. In the 20 equations where coefficients are computed for predicting a subject from another subject (i.e. $\gamma_{1\text{-}2}$ or $\gamma_{2\text{-}1}$), 17 of those cases produce negative coefficients that whether statistically significant or not, support the hypothesis. Two of the remaining cases (Mathematics predicted by Reading and Social Studies predicted by Science) the coefficients although positive are not statistically significant and support the hypothesis. The final coefficient (Mathematic predicted by Writing) is statistically significant but very small; the cross subject coefficient is .022 compared to the same subject coefficient of .99 or nearly fifty times larger; and also supports the hypothesis.

The 10 models of Table 10a often produced models with negative error variances (see Appendix G). Like any variance, a negative error variance (sometimes called a Heywood case) is a conceptual impossibility. Some analysts think that a solution with a negative error variance is not a legitimate solution but other analysts disagree (SAS, 1999). Reasons cited for negative error variances include:

- bad prior communality estimates
- too many common factors
- too few common factors
- not enough data to provide stable estimates
- the common factor model is not an appropriate model for the data

There were no prior communality estimates so that is not a reason for negative error variances in this study. The models, showing a total of four factors across a nominal 76 items (or 19 items per factor) would preclude too many common factors and suggest the opposite - too few common factors, except the congeneric models did not produce negative error variances. With 10,000 observations, there should be sufficient data. It is also possible that the common factor model is not appropriate to the data, yet there is ample evidence from the congeneric models, that a single common factor does not produce negative error variances. Therefore, one can choose between two possibilities: 1. The crossed subject factor models (equations 1 and 2) are not appropriate for the data which supports discriminant validity or 2. The negative error variances are small and can be ignored in which case the data of Table 10 supports discriminant validity. Whichever perspective one chooses, the negative error variances are very small.

As a matter of logic, the structural equation models are being used to find evidence that disconfirms the contention that the OGT tests are suited for making valid inferences. Acceptance of the SEM's shows no evidence that the OGT's lack validity for making inferences. Rejection of the SEM's results in the same logical position except there is less evidence that failed to refute the hypothesis.

Two additional models were analyzed to serve as sources for comparison to the models of Table 10a. Both of those models were made using dissimilar content areas for the dependent and independent variables and are shown as equations 3a, 3b, 4a and 4b.

$$\xi_{Read} = \gamma_{Read\text{-}Math}\eta_{Math}+\gamma_{Read\text{-}Social}\eta_{social}+b_1+e_1 \quad \text{eq. 3a}$$

$$\xi_{Science} = \gamma_{Science\text{-}Math}\eta_{Math}+\gamma_{Science\text{-}Social}\eta_{Social}+b_2+e_2 \quad \text{eq. 3b}$$

$$\xi_{Science} = \gamma_{Science\text{-}Read}\eta_{Read}+\gamma_{Science\text{-}Math}\eta_{Math}+b_3+e_3 \quad \text{eq. 4a}$$

$$\xi_{Social} = \gamma_{Social\text{-}Read}\eta_{Read}+\gamma_{Social\text{-}Math}\eta_{Math}+b_4+e_4 \quad \text{eq. 4b}$$

The results of analyzing these models are shown as Table 10b where each variable was modeled using the same data as for Table 10a but only the items in the -1 subscripted test half (see Appendix F for the items in the split tests). This practice maintained the same data sets for the analyses of both Table 10a and Table 10b.

Table 10b – Results of convergent and discriminanat validity comparisons of dissimilar test models

Model	Dependent Variable (dv)	Independent Variable 1 (iv1)	Independent Variable 2 (iv2)	$\gamma_{dv\text{-}iv1}$	$\gamma_{dv\text{-}iv2}$
eq 3a	Reading	Math	Social Studies	0.15 (7.07)	0.77 (28.46)
eq 3b	Science	Math	Social Studies	0.44 (14.50)	0.56 (15.60)
eq 4a	Science	Reading	Math	0.40 (14.50)	0.61 (16.16)
eq 4a	Social Studies	Reading	Math	0.66 (27.27)	0.33 (17.84)

Notes: 1. Details are provided in appendix G in Tables G – 11a, b, c, & d and in Figures G – 2 & 3
2. t statistic shown, in parentheses, below path coefficient

The results show that in the absence of a same content area covariate (equations 3a, 3b, 4a, & 4b; Table 10b) the regression coefficients are all statistically significant and moderately positive or larger. However, in the presence of a same content area covariate (see table 10a), the regression coefficients for the dissimilar content area coefficients are negative (in 17 out of 20 cases) or positive but statistically insignificant (in 2 out of 20 cases) or significant but small (in 1 case out of 20 cases). The largest dissimilar content area regression coefficient in Table 10a is 0.077 while the smallest dissimilar regression coefficient in Table 10b is 0.15. The largest dissimilar content area coefficient in Table 10b (0.77) is smaller than the smallest similar content regression coefficient in Table 10a (0.98). This data support the notion that each of the content areas (Reading, Mathematics, Science, Social Studies, and Writing) is measuring something distinctly different from the other content areas. For example, structural equation modeling shows that Mathematics is measuring something distinctly different than Reading. However, the interpretation that Mathematics is actually measuring Mathematics is affirmed, not by the structural equation modeling, but by the alignment studies.

2.4 G Studies and Classification Consistency.

The spring 2006 OGT was analyzed for generalizability and classification consistency. Tables 11a, 11b, 11c, 11d, and 11e show kappa coefficients by performance level and group identity.

Table 11a – Estimated Kappa Coefficients for the Spring 2006 OGT Administration of Reading

Group	Basic	Proficient	Accelerated	Advanced	All
All	0.64335	0.68330	0.72123	0.67304	0.52076
Male	0.65263	0.68745	0.72176	0.66212	0.51433
Female	0.63036	0.67126	0.72062	0.68356	0.53246
American Indian	0.06628	0.70152	0.72115	0.63088	0.50264
Asian	0.61987	0.67045	0.72034	0.69726	0.53184
Black	0.67854	0.71057	0.70412	0.58664	0.51127
Hispanic	0.66586	0.70642	0.71440	0.63027	0.50666
White	0.63329	0.66731	0.72107	0.68373	0.53843
Multi	0.64298	0.68182	0.72075	0.65920	0.51623
Other	0.66959	0.70092	0.72395	0.67065	0.50736
Public	0.64634	0.68647	0.72140	0.66738	0.51825
Non-Public	0.54108	0.60597	0.69634	0.70839	0.60818

Table 11b. – Estimated Kappa Coefficients for the Spring 2006 OGT Administration of Mathematics

Group	Basic	Proficient	Accelerated	Advanced	All
All	0.64201	0.68454	0.72220	0.70901	0.50774
Male	0.65027	0.68562	0.72343	0.71088	0.50927
Female	0.63634	0.68451	0.72558	0.70740	0.51031
American Indian	0.65160	0.69922	0.72836	0.69826	0.48503
Asian	0.55177	0.63194	0.71165	0.72774	0.55214
Black	0.69124	0.71712	0.70413	0.64874	0.50536
Hispanic	0.65870	0.70432	0.72030	0.68512	0.49124
White	0.62015	0.66794	0.71882	0.71384	0.52259
Multicultural	0.68982	0.72258	0.72201	0.68854	0.50090
Other	0.65134	0.69451	0.72509	0.71891	0.50784
Public	0.64562	0.68754	0.68754	0.70681	0.50797
Non-Public	0.54338	0.61200	0.70371	0.71847	0.56456

Table 11c – Estimated Kappa Coefficients for the Spring 2006 OGT Administration of Writing

Group	Basic	Proficient	Accelerated	Advanced	All
All	0.62177	0.66850	0.71122	0.49674	0.55776
Male	0.63256	0.67736	0.71522	0.44585	0.55124
Female	0.59699	0.65391	0.70082	0.53650	0.57205
American Indian	0.62883	0.70275	0.71649	0.28496	0.55349
Asian	0.59518	0.65723	0.70056	0.61099	0.55710
Black	0.64500	0.69796	0.70925	0.31524	0.56101
Hispanic	0.64879	0.69522	0.71700	0.43398	0.54463
White	0.60863	0.65864	0.70368	0.51065	0.56392
Multi-Ethnic	0.62722	0.67396	0.71514	0.44955	0.56261
Other	0.62311	0.67731	0.70756	0.46777	0.54074
Public	0.62255	0.66971	0.71179	0.47706	0.55718
Non-Public	0.27408	0.62027	0.66861	0.61212	0.59350

Table 11d – Estimated Kappa Coefficients for the Spring 2006 OGT Administration of Science

Group	Basic	Proficient	Accelerated	Advanced	All
All	0.67311	0.72219	0.73168	0.68599	0.51578
Male	0.62813	0.67867	0.69102	0.64505	0.45323
Female	0.60478	0.66897	0.67378	0.62041	0.44554
American Indian	0.64339	0.68792	0.67136	0.58466	0.44938
Asian	0.59655	0.66701	0.69190	0.67274	0.47008
Black	0.63348	0.65438	0.58406	0.47217	0.46292
Hispanic	0.64200	0.67993	0.64962	0.58148	0.45292
White	0.55259	0.62486	0.65176	0.60433	0.41690
Multi-Ethnic	0.60205	0.66022	0.65421	0.59078	0.42147
Other	0.67681	0.70874	0.72207	0.68920	0.48434
Public	0.62682	0.68093	0.68905	0.63470	0.45512
Non-Public	0.37524	0.48531	0.56607	0.53365	0.35990

Table 11e – Estimated Kappa Coefficients for the Spring 2006 OGT Administration of Social Studies

Group	Basic	Proficient	Accelerated	Advanced	All
All	0.65504	0.69839	0.72707	0.71603	0.51331
Male	0.67874	0.71449	0.74027	0.73236	0.53145
Female	0.60579	0.66242	0.69355	0.68006	0.47245
American Indian	0.64652	0.67668	0.69168	0.66299	0.46861
Asian	0.63334	0.68241	0.72072	0.72417	0.52747
Black	0.66741	0.69525	0.67227	0.62943	0.46977
Hispanic	0.67006	0.70449	0.69828	0.67136	0.47868
White	0.62038	0.67011	0.71180	0.70482	0.50568
Multi-Ethnic	0.65134	0.69276	0.71013	0.69195	0.48805
Other	0.68422	0.70983	0.73765	0.73358	0.52772
Public	0.65577	0.70164	0.72436	0.71059	0.50765
Non-Public	0.44122	0.51774	0.62313	0.63293	0.46399

The data of Tables 11a through 11e show good overall decision consistency across race and gender. The Cohen's Kappa statistic (Cohen, 1960) is based on a chance adjusted model of rater agreement. Landis and Koch (1977) are often cited for the qualitative interpretation of the kappa statistic: Less than 0 = poor agreement, 0 to .2 = slight agreement, .21 to .4 = fair agreement, .41 to .6 = moderate agreement, .61 to .8 = substantial agreement, and .81 to 1.0 = almost perfect agreement. In cells where decision consistency drops, those cells are often cells with few (or fewer) examinees. There are no instances where the decision consistency is negative or close to zero.

Tables 12a, 12b, 12c, 12d, and 12e show estimated reliabilities by performance level and group identity.

Table 12a – Estimated Reliability Coefficients for the Spring 2006 OGT Administration of Reading

Estimates Of OGT Reading Test Reliability (Cronbach's α) For Each Sub-content Area By Subgroup (Grade 10 students only) Spring 2006					
Group/Subgroup	Total Reading	Informational Text	Reading Processes	Literary Text	Acquisition of vocabulary
Public students (N=136,964)	.894127	.749555	.683785	.681924	.544306
Non-public (N=13,672)	.792816	.584793	.480880	.478158	.368341
Gender					
Female (N=74,207)	.884277	.737068	.661174	.649495	.524115
Male (N=76,179)	.900034	.756985	.698615	.695685	.566819
Ethnicity					
American Indian (N=235)	.906663	.762989	.726955	.715232	.576607
Asian-Pacific Is. (N=1,865)	.892179	.761599	.683839	.661637	.523201
Black-African (N=20,147)	.894147	.738756	.690886	.703808	.545829
Hispanic Amer. (N=2,753)	.898683	.770628	.697802	.683737	.543238
Multi-Cultural (N=2,628)	.890257	.730300	.671126	.683579	.559083
Unknown (N=3,773)	.908996	.773846	.716042	.729333	.590334
Other (N=808)	.911883	.785113	.741929	.715792	.619246
White (N=118,427)	.879711	.717943	.652847	.645576	.514129
Total Group (N=150,381)	.893679	.747896	.683060	.679185	.547187

Notes: Some examinees failed to report group classification(s).

Table 12b – Estimated Reliability Coefficients for the Spring 2006 OGT Administration of Mathematics

Estimates of OGT Mathematics Test Reliability (Cronbach's α) For Each Sub-content Area By Subgroup (grade 10 students only) Spring 2006						
Group/Subgroup	Total Mathematics	Number Sense	Measurement	Geometry	Data Analysis	Patterns, Functions & Algebra
Public students (N=137,673)	.906928	.689522	.692595	.558427	.657661	.713819
Non-public (N=13,675)	.862183	.584304	.639058	.478806	.492497	.622517
Gender						
Female (N=74,632)	.900182	.673550	.688385	.536060	.631113	.698876
Male (N=76,272)	.912237	.703401	.696607	.577654	.675750	.727782
Ethnicity						
American Indian (N=241)	.907005	.687077	.693769	.585508	.660521	.683873
Asian-Pacific Isl. (N=1,878)	.902129	.657550	.739886	.522445	.592679	.728756
BL-AA (N=20,478)	.880001	.651478	.599994	.465176	.634181	.627886
Hispanic (N=2,792)	.893330	.652919	.658421	.501952	.640073	.664638
Multicultural (N=2,495)	.899104	.684827	.671440	.526871	.641589	.695143
Other (N=811)	.920369	.740155	.735004	.577575	.699710	.727465
White (N=118,697)	.897175	.663558	.670431	.536910	.619817	.701633
Total Group (N=151,348)	.906574	.689185	.693123	.557655	.654511	.713634

Table 12b – Estimated Reliability Coefficients for the Spring 2006 OGT Administration of Writing

Estimates Of OGT Writing Test Reliability (Cronbach's α) For Each Sub-content Area By Subgroup (Grade 10 students only) Spring 2006				
Group/Subgroup	Total Writing	Process	Application	Content
Public students (N=136,480)	.833254	.597647	.708588	.685636
Non-public (N=13,677)	.698433	.463633	.561318	.489336
Gender				
Female (N=74,082)	.810283	.584140	.653878	.625363
Male (N=76,185)	.840334	.606523	.724283	.703888
Unknown (N=250)	.825885	.636629	.672466	.665953
Ethnicity				
American Indian (N=237)	.840156	.600906	.709522	.617073
Asian-Pacific Isl. (N=1,873)	.829518	.620253	.688684	.715233
Black-African American (N=20,023)	.814615	.578281	.689763	.651842
Hispanic American (N=2,760)	.862971	.627592	.727529	.737712
Multicultural (N=2,482)	.829369	.616984	.699996	.691782
Other (N=801)	.855658	.628302	.761337	.732826
Unknown (N=144)	.859649	.691223	.757160	.769331
White (N=118,425)	.833219	.575326	.606909	.681880
Total Group (N=150,517)	.832921	.602184	.707231	.684955

Table 12d – Estimated Reliability Coefficients for the Spring 2006 OGT Administration of Science

Estimates Of OGT Science Test Reliability (Cronbach's α) For Each Sub-content Area By Subgroup (Grade 10 students only) Spring 2006					
Group/Subgroup	Total Science	Earth Science	Life Science	Physical Science	Technology
Public students (N=133,861)	.874545	.635372	.671356	.659479	.607204
Non-public (N=13,393)	.821762	.489015	.651464	.517932	.490730
Gender					
Female (N=72,598)	.859913	.696252	.650224	.624506	.581867
Male (N=74,397)	.884617	.657461	.692327	.682995	.632757
Ethnicity					
American Indian (N=234)	.864770	.620915	.637685	.680585	.558649
Asian Pacific Is. (N=1,837)	.880738	.651311	.714706	.619816	.623787
Black-African (N=19,868)	.822907	.549900	.484081	.609635	.528671
Hispanic (N=2,697)	.854540	.599003	.592570	.644047	.589596
Multicultural (N=2,423)	.869357	.608897	.666309	.643897	.574405
Other (N=776)	.887821	.641590	.714655	.688499	.636952
White (N=115,598)	.860955	.599921	.668533	.616104	.579136
Total Group (N=147,254)	.873803	.629691	.673662	.657068	.607363

Notes: Some examinees failed to report group classification(s).

Table 12e – Estimated Reliability Coefficients for the Spring 2006 OGT Administration of Science

Estimates Of OGT Social Studies Test Reliability (Cronbach's α) For Each Sub-content Area By Subgroup (Grade 10 students only) Spring 2006					
Group/Subgroup	Total Social Studies	History	People/ Geography	Economics	Social Studies Skills
Public students (N=137,009)	.896036	.704830	.740472	.617482	.635614
Non-public (N=13,646)	.827808	.595845	.606659	.457186	.487376
Gender					
Female (N=74,300)	.884276	.675301	.723101	.574025	.615964
Male (N=76,116)	.904736	.728435	.752409	.648347	.654830
Ethnicity					
American Indian (N=238)	.884972	.656179	.712471	.665956	.605788
Asian-Pacific Is. (N=1,870)	.898048	.693453	.716135	.590710	.620218
Black-African (N=20,125)	.879989	.663358	.712291	.592131	.573295
Hispanic (N=2,769)	.891999	.689788	.743661	.612657	.587134
Multicultural (N=2,479)	.894910	.703205	.729480	.607120	.662343
Unknown (N=141)	.913613	.735153	.734653	.573259	.637452
Other (N=796)	.909049	.746278	.765637	.628152	.672010
White (N=118,480)	.886874	.691572	.719244	.588100	.617028
Total Group (N=150,655)	.895348	.704046	.738173	.613702	.636500

Notes: Some examinees failed to report group classification(s).

Tables 12a through 12e show high levels of internal consistency as indicated by the reliability estimates. The reliability estimates are comparable to the G coefficients computed and reported for Reading, Mathematics, Science, Social studies and Writing as reported in Appendix H despite the data in Appendix H using a different number of examinees from a Spring administration for a different year for the G studies. Results produced using different methods and samples, yet providing very similar estimates, tend to increase the confidence in the methods used for building and scoring the OGT.

2.5 Test Success.

Students may take the OGT as many as seven times with their graduating class cohort. For the cohort graduating in 2007 the administrations include three spring administrations (2005, 2006, and 2007), two summer administrations (2005 and 2006) and two fall administrations (also 2005 and 2006). While Ohio law precludes the Office of Assessment from following examinees through those seven administrations, some estimates have been made of the success rates for examinees by gender and by ethnicity. The data are confounded by an inability to know, with certainty, how many examinees attempted all seven administrations, how many examinees are taking the test because they move into the State of Ohio after the spring administration of their sophomore year, how many examinees have been incorrectly classified or erratically classified by demographic and how many examinees withdrew and when they withdrew (or why they withdrew). Tables 14a through 14e show counts for number of examinees sitting for the first test administration as well as an estimate of the number of examinees that were either successful on any one of the tests or attempted the tests in the last cohort test administration (the spring administration of their Senior year). The tables also show success rates on the OGT by ethnicity and gender.

Table 14a – Estimated Success Rates for Examinees through Seven Test Administrations in OGT Writing

Populations	Grade 10 participation	Participation in the single success tournament[9]	Success rate[10]	Success rate relative to…
Ethnicity based[11]				**White**
*	2382			
Blank	212			
American Indian	247	235	98.7%	99.7%
Asian	1870	2055	98.6%	99.6%
Black	18946	17823	97.0%	97.9%
Hispanic	2594	2511	97.0%	97.9%
White	118902	118051	99.1%	100.0%
Multi-racial	2083	2125	98.4%	99.4%
Other	511	646	95.8%	96.7%
Gender based				**Male**
Blank	917	814	97.5%	99.3%
Female	72348	73166	99.3%	101.1%
Male	74482	71474	98.2%	100.0%

Note: Footnotes to Table 14a are applicable to Tables 14a, 14b, 14c, 14d, and 14e.

[9] A single success tournament rate is the sum of all those who were successful on the test up to the last test administration plus those attempting the last administration.

[10] The success rate is the sum of all those who passed divided by the Single success tournament participation count.

[11] The code "*" and "blank" only appear in the Spring 2005 administration for all five tests.

Table 14b – Estimated Success Rates for Examinees through Seven Test Administrations in OGT Reading

Populations	Grade 10 participation	Participation in the single success tournament	Success rate	Success rate relative to…
Ethnicity based				White
*	2423			
Blank	220			
American Indian	253	250	98.8%	99.7%
Asian	1873	2059	98.3%	99.2%
Black	19278	18115	96.8%	97.6%
Hispanic	2626	2563	97.2%	98.0%
White	119218	119188	99.2%	100.0%
Multi-racial	2109	2183	98.8%	99.7%
Other	514	581	94.6%	95.4%
Gender based				Male
Blank	941	861	96.9%	98.3%
Female	72646	73910	99.1%	100.6%
Male	74927	74195	98.5%	100.0%

Table 14c – Estimated Success Rates for Examinees through Seven Test Administrations in OGT Mathematics

Populations	Grade 10 participation	Participation in the single success tournament	Success rate	Success rate relative to …
Ethnicity based				White
*	2444			
Blank	215			
American Indian	253	224	97.3%	99.0%
Asian	1876	2079	99.1%	100.9%
Black	19421	17037	90.5%	92.1%
Hispanic	2638	2466	95.1%	96.7%
White	119149	116880	98.3%	100.0%
Multi-racial	2097	2072	96.9%	98.6%
Other	517	675	90.5%	92.1%
Gender based				Male
Blank	951	766	93.3%	95.8%
Female	72736	70752	97.2%	99.7%
Male	74923	71695	97.4%	100.0%

Table 14d – Estimated Success Rates for Examinees through Seven Test Administrations in OGT Social Studies

Populations	Grade 10 participation	Participation in the single success tournament	Success rate	Success rate relative to …
Ethnicity based				**White**
*	2362			
Blank	206			
American Indian	247	231	97.4%	99.4%
Asian	1862	2031	97.6%	99.6%
Black	18775	16363	90.6%	92.4%
Hispanic	2590	2402	94.7%	96.6%
White	118674	115753	98.0%	100.0%
Multi-racial	2077	2024	96.6%	98.6%
Other	509	659	91.0%	92.9%
Gender based				**Male**
Blank	921	755	94.0%	96.7%
Female	72138	69726	96.8%	99.6%
Male	74243	70698	97.3%	100.0%

Table 14e – Estimated Success Rates for Examinees through Seven Test Administrations in OGT Science

Populations	Grade 10 participation	Participation in the single success tournament	Success rate	Success rate relative to …
Ethnicity based				**White**
*	2399			
Blank	208			
American Indian	251	216	95.4%	98.2%
Asian	1875	2014	95.6%	98.4%
Black	19005	15628	81.6%	84.0%
Hispanic	2614	2328	90.3%	92.9%
White	118830	114780	97.2%	100.0%
Multi-racial	2086	1974	94.3%	97.1%
Other	518	720	84.9%	87.3%
Gender based				**Male**
Blank	933	731	89.5%	93.1%
Female	74504	71591	94.7%	98.5%
Male	72349	72494	96.1%	100.0%

The data of Tables 14a through 14e fail to show large differences in success rates between groups with the possible exception of Science where those self reporting as Black are succeeding at 84% of the rate for whites, after adjusting for nominal rates[12] of educational persistence.

[12] The concept of "nominal rates" is invoked to reflect uncertainty in demographic shifts. Exactly why a demographic group grows (or shrinks) compared to the baseline (Spring 2005) administration when the cohort under study was tested as 10th graders is not known. Factors that could impact these rates include dropping out of school, in-migration and out-migration and changes in self classification; it is suspected that dropping out of school may be the dominant practice for explaining large downward shifts in counts of examinees. Whatever the reason for these shifts, it is presumed that students intending to obtain an Ohio high school diploma will test throughout their high school career until they are successful or no longer attending high school.

PART 3 – Discussion.

The best evidence for asserting that the Ohio Graduation Tests are suitable for making valid inferences would be some predicted schooling outcome such as other academic performance measures or work type measures (or other training measures). Presumably, students are being schooled for some purpose and if we knew that purpose and had measures of performance in the context of the purpose, that data would be the best evidence of validity – predictive validity.

In the absence of some variable to predict, some validity studies make use of similar tests, often national norm referenced tests, to inspect the nomological validity of test instruments like the OGT's. The primary reason that was not done is that the data were not available, in part, due to Ohio privacy laws. Should that data be made available, it will be used in filling in a gap in this validity study. However, it should be recognized that when comparing the OGT to a national norm referenced test, the comparison is (to some extent) a comparison of content standards for the two tests.

Ohio (like other states) relies on the expertise of educators in defining both content and suitable student performance. Most Ohio education prior to the introduction of testing to Ohio schools in the early 1990's relied on the individual expertise of classroom teachers to determine *what* his or her students should know or be able to do as well as determining *whether* the students knew and could do those things. With the introduction of the Ohio testing program and the Ohio Graduation Tests, assessment of student outcomes still rely on the judgments of teachers except the judgment of individual teachers is being supplemented by the collective judgment of a larger body of teachers, educational experts and other interested parties. This is central to the concept of validity as a community property.

Ohio's extensive use of committees in the determination of test content, the appropriateness of test questions, setting performance standards, and reviewing and participating in the formulation of scoring rubrics (i.e. rangefinding) results in test processes that are valid for making inferences about what Ohio students know and can do.

The quantitative analyses in Part Two of this paper failed to find evidence that the processes described in Part One have produced tests that were lacking validity by being poorly aligned or lacking convergent validity or exhibiting poor discriminant validity.

Finally, to some readers the extensive appendices may seem unnecessary. They are there only for the convenience of the reader. It is the intent of this study to be as "public" as possible …ergo… the appendices.

PART 4 - Conclusion.

According to Carnap (1936):

> "If by verification is meant a definitive and final establishment of truth, then no (synthetic) sentence is ever verifiable, … We can only confirm a sentence more and more. Therefore we shall speak of the problem of confirmation rather than verification."

Based on the data in this study there are more reasons to believe that the OGT can be used to make inferences about student ability than the opposite. Because the OGT lacks a variable to predict, the validity of the tests are based on input from the community of teachers, educators, parents, and the public; persons how know and care about education in Ohio. Testing in Ohio is under the surveillance of multiple committees (not just one or two committees) and thousands of eyes because of Ohio practices for public release of items from the Spring administration.

The quantitative data confirm the processes used in the creation and maintenance of the testing program. Alignment studies, correlation studies (congeneric studies, factor analysis studies and regression studies) and studies of decision consistency provide no data to indicate that the tests are not performing as they should.

The Ohio Graduation Tests are suitable for making valid inferences about what 10th grade students know and can in the context of Ohio's Academic Content Standards.

References

Academic Content Standards: K-12 English Language Arts, 2004. Columbus: Ohio Department of Education

Academic Content Standards: K-12 Mathematics, 2004. Columbus: Ohio Department of Education

Academic Content Standards: K-12 Science, 2003. Columbus: Ohio Department of Education.

Academic content Standards: K-12 Social Studies, 2003. Columbus: Ohio Department of Education

ACT, 2007. Technical Manual. Iowa City: ACT. 144 pp.

ACT, 2005. ACT Writing Test Preliminary Technical Report. Iowa City: ACT. 4 pp.

American Educational Research Association (AERA), 1999. Standards for Educational and Psychological Testing. Washington: American Educational Research Association. 194 pp.

Angoff, W. H., 1988. Validity: An Evolving Concept in Test Validity, H. Wainer & H. L. Braun, eds. Hillsdale: Erlbaum.

Bagozzi, R. P. & Yi, Y., 1988. On the Evaluation of Structural Equation Models. Journal of the Academy of Marketing Science, 16:1, p74-94.

Brennan, R. L., 1992. Elements of Generalizability Theory. Iowa City:ACT. 161 pp.

Bunch, M. B., 2006. Ohio Graduation Tests Standard Setting Report: Reading and Mathematics (T. Moore, ed.). Columbus: Ohio Department of Education. 99 pp.

Bunch. M. B., 2006 (2). Final Report for the Reliability Study of the OGT for Spring 2005. Durham: Measurement Incorporated.

Bunch, M. B., Inman, E., & Miles, J., 2006. Ohio Graduation Tests Standard Setting Report (T. Moore ed.). Columbus: Ohio Department of Education. 169 pp.

Byrne, B. M., 1989. A Primer of LISREL. New York: Springer-Verlag, 184 pp.

Byrne, B. N., 1998. Structural equation modeling with LISREL, PRELIS, and SIMPLIS: Basic Concepts, Applications and Programming. Mahwah: Erlbaum. 412pp

Carnap, R. (1936) Testability and Meaning a reprint from Philosophy of Science, v3 (1936) and v4 (1937) found in Readings in the Philosophy of Science, H. Feigl & M. Broadbeck eds. (1953) New York: Appleton-Century-Crofts, pp 47-92.

Cohen, J., 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement* XX:1

Cronbach, L. J. (1989). Construct Validation After Thirty Years in Intelligence: Measurement, Theory, and Public Policy, R.L. Linn, ed. Champaign: University of Illinois, 240 pp.

Cronbach, L. J. (1988). Five perspectives on Validity Argument in Test Validity, H. Wainer & H. L. Braun, eds. Hillsdale: Erlbaum.

Feil, J. (2006) Electronic evidence submitted in application for peer review and approval under NCLB. Columbus: Ohio Department of Education

Gerbing, D. W. & Anderson, J. C, 1993. Monte Carlo Evaluations of Goodness-of-Fit Indices for Structural Equation Models in Testing Structural Equation Models, Bollen & Long, eds. Newbury Park: Sage.

Item Development Process: Internal Item Review, udtd. Columbus: Ohio Department of Education.

Johnson, H. L. 2006. Letter to S.T. Zelman, Ohio Department of Education, dated 15 November 2006. Washington: US Department of Education.

Jöreskog, K.G. & Sörbom, D., 1988. LISREL7: A Guide to the Program and Applications, 2nd ed. Chicago: SPSS, 342 pp.

Kane, M. T., 2006. Validation in Educational Measurement (4th ed.), R.L. Brennan, ed. Praeger. 808pp.

Keene, J. 2006. Webb Alignment Study Report: Reading and Mathematics. Fair Oaks Ranch, Texas: Assessment and Evaluation Services, 35 pages plus appendices.

Landis, J. R. & Koch, G. R., 197. The Measurement of Observer Agreement for Categorical Data. Biometrics:33. pp. 159-174.

Linacre, J. M., 2005. A Users Guide to WINSTEPS. Chicago: WINSTEPS.COM, 313 pp.

Masters, G. N., 1982. A Rasch Model for Partial Credit Scoring. Psychometrica: 47, 2.

Messick, S.,1993. Validity in Educational Measurement (3rd ed), R.L. Linn ed. Phoenix: Oryx Press

Mitzel, H. C., Lewis, D. M., Patz, R. J. & Green, D. R., 2001. The Bookmark Procedure: Psychological Perspectives in Setting Performance Standards: Concepts, Methods, and Perspectives, G. J. Cizek, ed. Mahwah: Lawrence Earlbaum.

Nunnally, J. C. & Bernstein, I. H., 1992. Psychometric Theory. New York: McGraw-Hill, 752 pp.

ODE, 2003. OGT_Science_Blueprint_01_06_06[2].pdf. Columbus: Ohio Department of Education

ODE, 2003 (2). OGT_ Social_Studies_Blueprint_08_13_03[2].pdf. Columbus: Ohio Department of Education

ODE, 2003(3). Ohio Graduation Tests: Reading Item Specifications, 2003. Columbus: Ohio Department of Education.

ODE, 2003(4). Ohio Graduation Tests: Science Item Specifications, 2003. Columbus: Ohio Department of Education.

ODE, 2004. OGT_Writing_Blueprint_3-19-04[2].pdf. Columbus: Ohio Department of Education

ODE, 2004 (2). Ohio Graduation Tests: Mathematics Item Specifications, 2004. Columbus: Ohio Department of Education.

ODE, 2004 (3). Ohio Graduation Tests: Social Studies Item Specifications, 2004 (draft). Columbus: Ohio Department of Education.

ODE, 2006. OGT_Reading_Blueprint_01_06_06[2].pdf. Columbus: Ohio Department of Education

ODE, 2006 (2). Math_Blueprint_4_06[2].pdf. Columbus: Ohio Department of Education

ODE, 2007. 0607 AYP goals [2].pdf [web based]. Columbus: Ohio Department of Education

ODE, undated. ACSDevelopmentProcess[2].pdf [web based]. Columbus: ODE

SAS, 1999. OnlineDoc(TM), Version 7-1. Cary, NC:SAS Institute

Shavelson, R. J. and Web, N. J., 1991. Generalizability Theory. Thousand Oaks: Sage. 137 pp.

Webb, N. L. 2005. Web Alignment Tool (WAT); Training Manual (V1.1). Madison: Wisconsin Center for Education Research.

Wright, B. D. and Douglas, G. A., 1975. Research Memorandum Number 19: Best Test Design and Self-Tailored Testing. Chicago: University of Chicago

Wright, B. D. and Masters, G. N., 1982. Rating Scale Analysis. Chicago: MESA, 206 pp.

Wright, B. D. and Stone, M. H., 1979. Best Test Design. Chicago: MESA, 240 pp.

Appendix A: Preliminary results of the Science Alignment Study

The data from the Science alignment study are presented in Tables A1 through A5. The criteria provided for Table A1 though A4 is the default criteria of the Web Alignment Tool:

- for table A1, A score of less than 6 results in a negative or "no" classification
- for tables A2, and A3, a score between 40% and 50% results in a "weak" classification; a score below 40% would be a "no"
- for Table A4, a score between 0.60 and 0.70 results in a weak classification; a score below 0.60 would be classed as "no"

Table A1 – Categorical Concurrence for three forms of the Ohio Graduation Test

	(Want 6 or More Items per Standard)			
Standard	**Form A**	**Form B**	**Form C**	**Average**
1. Earth and Space Sciences	**12.5**	**12.33**	**11.33**	**12.05**
2. Life Sciences	**18.33**	**19**	**25.83**	**21.05**
3. Physical Sciences .	**17.83**	**20.33**	**15**	**17.72**
4. Scientific Methods and Applications	**11.67**	**14.67**	**14**	**13.45**

Notes: None of the cells are outside of the limits established for the WAT

Table A2 – Depth of Knowledge Consistency for three forms of the Ohio Graduation Test

	(Want 50% of items at or above Benchmark DOK)			
Standard	**Form A**	**Form B**	**Form C**	**Average**
1. Earth and Space Sciences	**62**	**79**	**88**	**76**
2. Life Sciences	**55**	**51**	**47**	**51**
3. Physical Sciences .	**66**	**85**	**75**	**75**
4. Scientific Methods and Applications	**64**	**42**	**80**	**62**

Notes: Two of the cells (47% & 42%) are classed as "weak"

Table A3 – Range Of Objectives for three forms of the Ohio Graduation Test

	(Want 50% or More)			
Standard	**Form A**	**Form B**	**Form C**	**Average**
1. Earth and Space Sciences	**84**	**81**	**83**	**83**
2. Life Sciences	**66**	**75**	**74**	**72**
3. Physical Sciences .	**85**	**75**	**69**	**76**
4. Scientific Methods and Applications	**49**	**63**	**63**	**58**

Notes: One of the cells (49%) is classed as weak

Table A4 – Balance Of Representation for three forms of the Ohio Graduation Test

	(Want index of 0.70 or greater)			
Standard	**Form A**	**Form B**	**Form C**	**Average**
1. Earth and Space Sciences	**0.70**	**0.75**	**0.75**	**0.73**
2. Life Sciences	**0.65**	**0.75**	**0.74**	**0.71**
3. Physical Sciences .	**0.74**	**0.74**	**0.76**	**0.75**
4. Scientific Methods and Applications	**0.80**	**0.80**	**0.76**	**0.79**

Notes: One of the cells (0.65) is classed as weak

Table A5 – Summary Comparison to other Ohio Alignment Studies

Parameter	**Reading July, 2006**	**Math July, 2006**	**Science March, 2007**
# cells	**256**	**300**	**48**
# cells classed as "No"	**1 (<1%)**	**3 (1%)**	**0 (0%)**
#cells classed as "Weak"	**15**	**27**	**4**
% errors	**6%**	**10%**	**8%**

Appendix B: Constructing a data file

Sampling:

Strata were identified as school typology and white+Asian or not. Race_Key = 1, else Race_Key = 0.

Resulting data:

type * Race_Key Crosstabulation

Count

		Race_Key		Total
		0	1	0
type	1	764	9251	10015
	2	538	14560	15098
	3	457	8736	9193
	4	3223	13791	17014
	5	10447	4539	14986
	6	2825	25669	28494
	7	2805	14566	17371
Total		21059	91112	112171

Also by type

Sample counts

Race_Key		Total
0	**1**	
68.1103	824.723	
47.96249	1298.018	
40.74137	778.8109	
287.3292	1229.462	
931.3459	404.65	
251.8476	2288.381	
250.0646	1298.553	
1877.401	8122.599	10000

Clusters were by school district
Files: Sample
With replacement

Summary for Stage 1

		Number of Units Sampled		Proportion of Units Sampled	
type	Race_Key	Requested	Actual	Requested	Actual
1	0	68	68	97.1%	97.1%
	1	825	825	896.7%	896.7%
2	0	48	48	41.4%	41.4%
	1	1298	1298	826.8%	826.8%
3	0	41	41	62.1%	62.1%
	1	779	779	973.8%	973.8%
4	0	287	287	292.9%	292.9%
	1	1229	1229	1216.8%	1216.8%
5	0	931	931	6206.7%	6206.7%
	1	252	252	1800.0%	1800.0%
6	0	250	250	252.5%	252.5%
	1	1299	1299	1261.2%	1261.2%
7	0	1877	1877	4080.4%	4080.4%
	1	8123	8123	17658.7%	17658.7%

Plan File: G:\Validity Study\plan1.csplan

Cluster: SIRN
With replacement
File: sampled2.sav

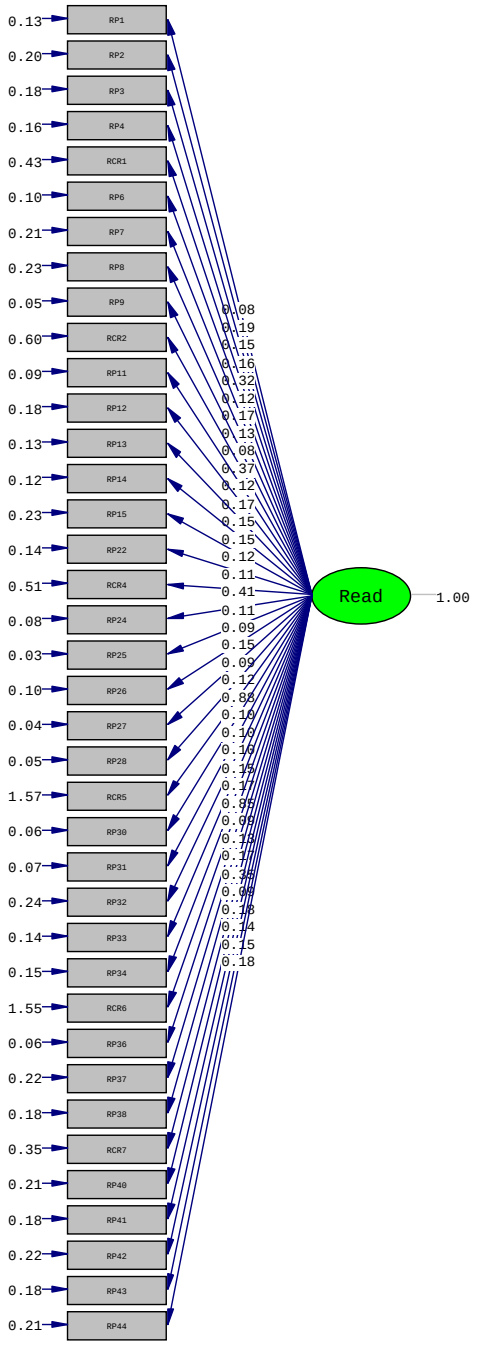
Summary for Stage 1

		Number of Units Sampled		Proportion of Units Sampled	
type	Race_Key	Requested	Actual	Requested	Actual
1	0	68	68	86.1%	86.1%
	1	825	825	743.2%	743.2%
2	0	48	48	38.4%	38.4%
	1	1298	1298	733.3%	733.3%
3	0	41	41	58.6%	58.6%
	1	779	779	875.3%	875.3%
4	0	287	287	254.0%	254.0%
	1	1229	1229	1032.8%	1032.8%
5	0	931	931	809.6%	809.6%
	1	252	252	240.0%	240.0%
6	0	250	250	215.5%	215.5%
	1	1299	1299	1047.6%	1047.6%
7	0	1877	1877	3077.0%	3077.0%
	1	8123	8123	13101.6%	13101.6%

Plan File: G:\Validity Study\plan1.csplan

Appendix C: Congeneric analyses

Figure C1 – Congeneric model for Reading



Analysis C1: Congeneric Reading model results

Goodness of Fit Statistics

Degrees of Freedom = 665
Minimum Fit Function Chi-Square = 5994.82 (P = 0.0)
Normal Theory Weighted Least Squares Chi-Square = 7348.40 (P = 0.0)
Estimated Non-centrality Parameter (NCP) = 6683.40
90 Percent Confidence Interval for NCP = (6411.30 ; 6962.67)

Minimum Fit Function Value = 0.60
Population Discrepancy Function Value (F0) = 0.67
90 Percent Confidence Interval for F0 = (0.64 ; 0.70)
Root Mean Square Error of Approximation (RMSEA) = 0.032
90 Percent Confidence Interval for RMSEA = (0.031 ; 0.032)
P-Value for Test of Close Fit (RMSEA < 0.05) = 1.00

Expected Cross-Validation Index (ECVI) = 0.75
90 Percent Confidence Interval for ECVI = (0.72 ; 0.78)
ECVI for Saturated Model = 0.15
ECVI for Independence Model = 15.27

Chi-Square for Independence Model with 703 Degrees of Freedom = 152611.33
Independence AIC = 152687.33
Model AIC = 7500.40
Saturated AIC = 1482.00
Independence CAIC = 152999.32
Model CAIC = 8124.39
Saturated CAIC = 7565.86

Normed Fit Index (NFI) = 0.96
Non-Normed Fit Index (NNFI) = 0.96
Parsimony Normed Fit Index (PNFI) = 0.91
Comparative Fit Index (CFI) = 0.96
Incremental Fit Index (IFI) = 0.96
Relative Fit Index (RFI) = 0.96

Critical N (CN) = 1256.59

Root Mean Square Residual (RMR) = 0.014
Standardized RMR = 0.026
Goodness of Fit Index (GFI) = 0.96
Adjusted Goodness of Fit Index (AGFI) = 0.96
Parsimony Goodness of Fit Index (PGFI) = 0.86

Analysis C2: Congeneric Mathematics model results

Goodness of Fit Statistics

Degrees of Freedom = 665
Minimum Fit Function Chi-Square = 5542.65 (P = 0.0)
Normal Theory Weighted Least Squares Chi-Square = 7000.59 (P = 0.0)
Estimated Non-centrality Parameter (NCP) = 6335.59
90 Percent Confidence Interval for NCP = (6070.43 ; 6607.95)

Minimum Fit Function Value = 0.55
Population Discrepancy Function Value (F0) = 0.63
90 Percent Confidence Interval for F0 = (0.61 ; 0.66)
Root Mean Square Error of Approximation (RMSEA) = 0.031
90 Percent Confidence Interval for RMSEA = (0.030 ; 0.032)
P-Value for Test of Close Fit (RMSEA < 0.05) = 1.00

Expected Cross-Validation Index (ECVI) = 0.72
90 Percent Confidence Interval for ECVI = (0.69 ; 0.74)
ECVI for Saturated Model = 0.15
ECVI for Independence Model = 24.51

Chi-Square for Independence Model with 703 Degrees of Freedom = 244997.00
Independence AIC = 245073.00
Model AIC = 7152.59
Saturated AIC = 1482.00
Independence CAIC = 245384.99
Model CAIC = 7776.57
Saturated CAIC = 7565.86

Normed Fit Index (NFI) = 0.98
Non-Normed Fit Index (NNFI) = 0.98
Parsimony Normed Fit Index (PNFI) = 0.92
Comparative Fit Index (CFI) = 0.98
Incremental Fit Index (IFI) = 0.98
Relative Fit Index (RFI) = 0.98

Critical N (CN) = 1359.02

Root Mean Square Residual (RMR) = 0.0072
Standardized RMR = 0.024
Goodness of Fit Index (GFI) = 0.96
Adjusted Goodness of Fit Index (AGFI) = 0.96
Parsimony Goodness of Fit Index (PGFI) = 0.87

All Path coefficients are large and significant.

Analysis C3: Congeneric Writing model results

Goodness of Fit Statistics

Degrees of Freedom = 90
Minimum Fit Function Chi-Square = 5586.34 (P = 0.0)
Normal Theory Weighted Least Squares Chi-Square = 5964.53 (P = 0.0)
Estimated Non-centrality Parameter (NCP) = 5874.53
90 Percent Confidence Interval for NCP = (5624.90 ; 6131.34)

Minimum Fit Function Value = 0.56
Population Discrepancy Function Value (F0) = 0.59
90 Percent Confidence Interval for F0 = (0.56 ; 0.61)
Root Mean Square Error of Approximation (RMSEA) = 0.081
90 Percent Confidence Interval for RMSEA = (0.079 ; 0.083)
P-Value for Test of Close Fit (RMSEA < 0.05) = 0.00

Expected Cross-Validation Index (ECVI) = 0.60
90 Percent Confidence Interval for ECVI = (0.58 ; 0.63)
ECVI for Saturated Model = 0.024
ECVI for Independence Model = 3.47

Chi-Square for Independence Model with 105 Degrees of Freedom = 34623.52
Independence AIC = 34653.52
Model AIC = 6024.53
Saturated AIC = 240.00
Independence CAIC = 34776.67
Model CAIC = 6270.84
Saturated CAIC = 1225.24

Normed Fit Index (NFI) = 0.84
Non-Normed Fit Index (NNFI) = 0.81
Parsimony Normed Fit Index (PNFI) = 0.72
Comparative Fit Index (CFI) = 0.84
Incremental Fit Index (IFI) = 0.84
Relative Fit Index (RFI) = 0.81

Critical N (CN) = 223.16

Root Mean Square Residual (RMR) = 0.068
Standardized RMR = 0.043
Goodness of Fit Index (GFI) = 0.93
Adjusted Goodness of Fit Index (AGFI) = 0.90
Parsimony Goodness of Fit Index (PGFI) = 0.69

Analysis C4: Congeneric Science model results

Goodness of Fit Statistics

Degrees of Freedom = 665
Minimum Fit Function Chi-Square = 3685.49 (P = 0.0)
Normal Theory Weighted Least Squares Chi-Square = 4266.54 (P = 0.0)
Estimated Non-centrality Parameter (NCP) = 3601.54
90 Percent Confidence Interval for NCP = (3398.81 ; 3811.64)

Minimum Fit Function Value = 0.37
Population Discrepancy Function Value (F0) = 0.36
90 Percent Confidence Interval for F0 = (0.34 ; 0.38)
Root Mean Square Error of Approximation (RMSEA) = 0.023
90 Percent Confidence Interval for RMSEA = (0.023 ; 0.024)
P-Value for Test of Close Fit (RMSEA < 0.05) = 1.00

Expected Cross-Validation Index (ECVI) = 0.44
90 Percent Confidence Interval for ECVI = (0.42 ; 0.46)
ECVI for Saturated Model = 0.15
ECVI for Independence Model = 14.76

Chi-Square for Independence Model with 703 Degrees of Freedom = 147542.26
Independence AIC = 147618.26
Model AIC = 4418.54
Saturated AIC = 1482.00
Independence CAIC = 147930.26
Model CAIC = 5042.53
Saturated CAIC = 7565.86

Normed Fit Index (NFI) = 0.98
Non-Normed Fit Index (NNFI) = 0.98
Parsimony Normed Fit Index (PNFI) = 0.92
Comparative Fit Index (CFI) = 0.98
Incremental Fit Index (IFI) = 0.98
Relative Fit Index (RFI) = 0.97

Critical N (CN) = 2043.34

Root Mean Square Residual (RMR) = 0.0061
Standardized RMR = 0.020
Goodness of Fit Index (GFI) = 0.98
Adjusted Goodness of Fit Index (AGFI) = 0.98
Parsimony Goodness of Fit Index (PGFI) = 0.88

Analysis C5: Congeneric Social Studies model results

Goodness of Fit Statistics

Degrees of Freedom = 665
Minimum Fit Function Chi-Square = 2965.18 (P = 0.0)
Normal Theory Weighted Least Squares Chi-Square = 3219.14 (P = 0.0)
Estimated Non-centrality Parameter (NCP) = 2554.14
90 Percent Confidence Interval for NCP = (2381.05 ; 2734.66)

Minimum Fit Function Value = 0.30
Population Discrepancy Function Value (F0) = 0.26
90 Percent Confidence Interval for F0 = (0.24 ; 0.27)
Root Mean Square Error of Approximation (RMSEA) = 0.020
90 Percent Confidence Interval for RMSEA = (0.019 ; 0.020)
P-Value for Test of Close Fit (RMSEA < 0.05) = 1.00

Expected Cross-Validation Index (ECVI) = 0.34
90 Percent Confidence Interval for ECVI = (0.32 ; 0.36)
ECVI for Saturated Model = 0.15
ECVI for Independence Model = 21.86

Chi-Square for Independence Model with 703 Degrees of Freedom = 218542.71
Independence AIC = 218618.71
Model AIC = 3371.14
Saturated AIC = 1482.00
Independence CAIC = 218930.70
Model CAIC = 3995.12
Saturated CAIC = 7565.86

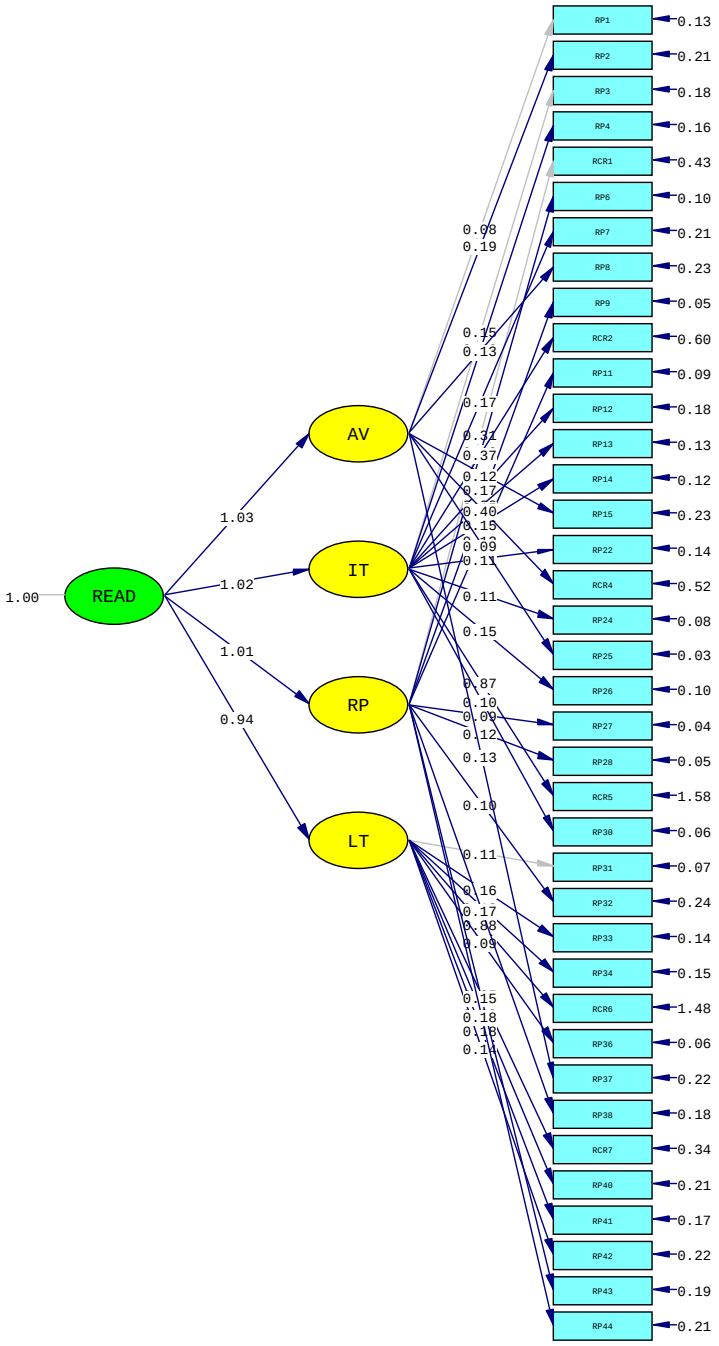
Normed Fit Index (NFI) = 0.99
Non-Normed Fit Index (NNFI) = 0.99
Parsimony Normed Fit Index (PNFI) = 0.93
Comparative Fit Index (CFI) = 0.99
Incremental Fit Index (IFI) = 0.99
Relative Fit Index (RFI) = 0.99

Critical N (CN) = 2539.47

Root Mean Square Residual (RMR) = 0.0044
Standardized RMR = 0.017
Goodness of Fit Index (GFI) = 0.98
Adjusted Goodness of Fit Index (AGFI) = 0.98
Parsimony Goodness of Fit Index (PGFI) = 0.88

Appendix D: Analysis of content areas by strand

Figure D1 – <u>Second order, confirmatory factor analysis, full information model for Reading</u>



Analysis D1: Second order, confirmatory factor analysis, full information model for Reading

Goodness of Fit Statistics

Degrees of Freedom = 661
Minimum Fit Function Chi-Square = 5859.16 (P = 0.0)
Normal Theory Weighted Least Squares Chi-Square = 7184.59 (P = 0.0)
Estimated Non-centrality Parameter (NCP) = 6523.59
90 Percent Confidence Interval for NCP = (6254.69 ; 6799.67)

Minimum Fit Function Value = 0.59
Population Discrepancy Function Value (F0) = 0.65
90 Percent Confidence Interval for F0 = (0.63 ; 0.68)
Root Mean Square Error of Approximation (RMSEA) = 0.031
90 Percent Confidence Interval for RMSEA = (0.031 ; 0.032)
P-Value for Test of Close Fit (RMSEA < 0.05) = 1.00

Expected Cross-Validation Index (ECVI) = 0.73
90 Percent Confidence Interval for ECVI = (0.71 ; 0.76)
ECVI for Saturated Model = 0.15
ECVI for Independence Model = 15.27

Chi-Square for Independence Model with 703 Degrees of Freedom = 152611.33
Independence AIC = 152687.33
Model AIC = 7344.59
Saturated AIC = 1482.00
Independence CAIC = 152999.32
Model CAIC = 8001.42
Saturated CAIC = 7565.86

Normed Fit Index (NFI) = 0.96
Non-Normed Fit Index (NNFI) = 0.96
Parsimony Normed Fit Index (PNFI) = 0.90
Comparative Fit Index (CFI) = 0.97
Incremental Fit Index (IFI) = 0.97
Relative Fit Index (RFI) = 0.96

Critical N (CN) = 1278.40

Root Mean Square Residual (RMR) = 0.014
Standardized RMR = 0.026
Goodness of Fit Index (GFI) = 0.96
Adjusted Goodness of Fit Index (AGFI) = 0.96
Parsimony Goodness of Fit Index (PGFI) = 0.86

Figure D2 – 2nd order confirmatory factor analysis model for Mathematics

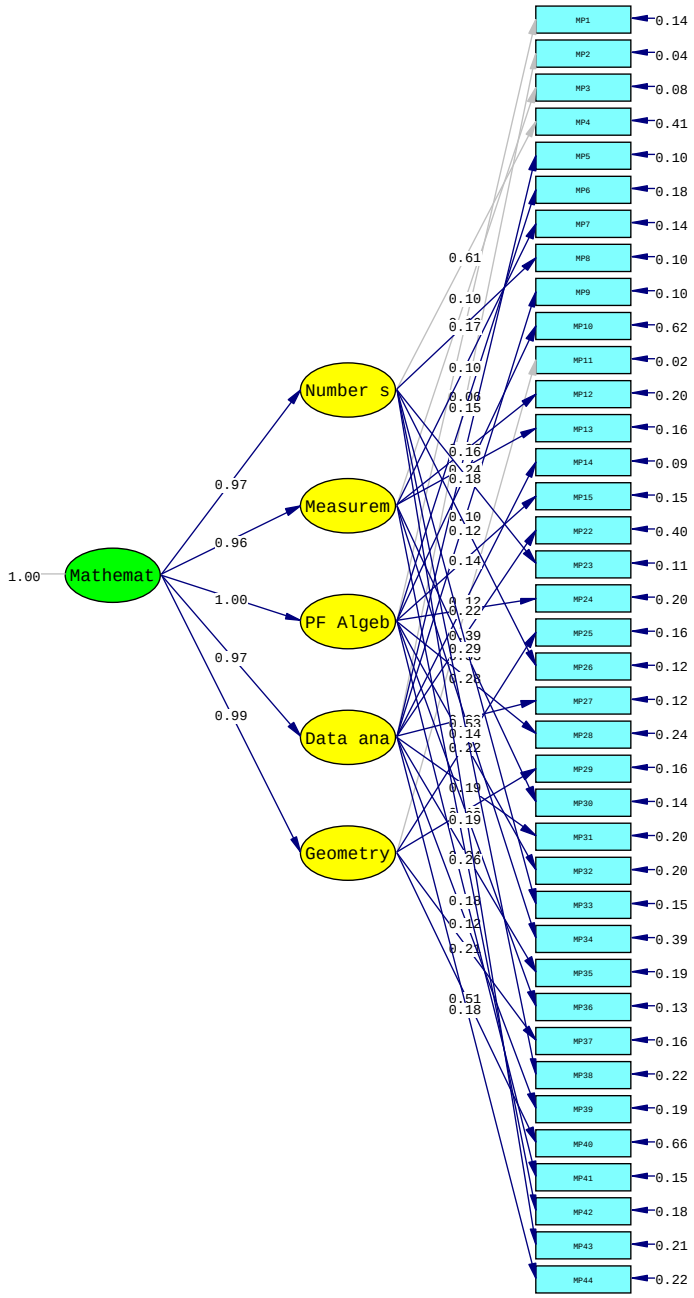


Table D2 - Data for 2nd order confirmatory factor analysis of Mathematics

Analysis 2: Second order, full information model for Mathematics

Goodness of Fit Statistics

Degrees of Freedom = 660
Minimum Fit Function Chi-Square = 5387.77 (P = 0.0)
Normal Theory Weighted Least Squares Chi-Square = 6755.54 (P = 0.0)
Estimated Non-centrality Parameter (NCP) = 6095.54
90 Percent Confidence Interval for NCP = (5835.32 ; 6362.96)

Minimum Fit Function Value = 0.54
Population Discrepancy Function Value (F0) = 0.61
90 Percent Confidence Interval for F0 = (0.58 ; 0.64)
Root Mean Square Error of Approximation (RMSEA) = 0.030
90 Percent Confidence Interval for RMSEA = (0.030 ; 0.031)
P-Value for Test of Close Fit (RMSEA < 0.05) = 1.00

Expected Cross-Validation Index (ECVI) = 0.69
90 Percent Confidence Interval for ECVI = (0.67 ; 0.72)
ECVI for Saturated Model = 0.15
ECVI for Independence Model = 24.51

Chi-Square for Independence Model with 703 Degrees of Freedom = 244997.00
Independence AIC = 245073.00
Model AIC = 6917.54
Saturated AIC = 1482.00
Independence CAIC = 245384.99
Model CAIC = 7582.57
Saturated CAIC = 7565.86

Normed Fit Index (NFI) = 0.98
Non-Normed Fit Index (NNFI) = 0.98
Parsimony Normed Fit Index (PNFI) = 0.92
Comparative Fit Index (CFI) = 0.98
Incremental Fit Index (IFI) = 0.98
Relative Fit Index (RFI) = 0.98

Critical N (CN) = 1388.18

Root Mean Square Residual (RMR) = 0.0072
Standardized RMR = 0.024
Goodness of Fit Index (GFI) = 0.97
Adjusted Goodness of Fit Index (AGFI) = 0.96
Parsimony Goodness of Fit Index (PGFI) = 0.86

Figure D3 – 2nd order confirmatory factor analysis model for Science

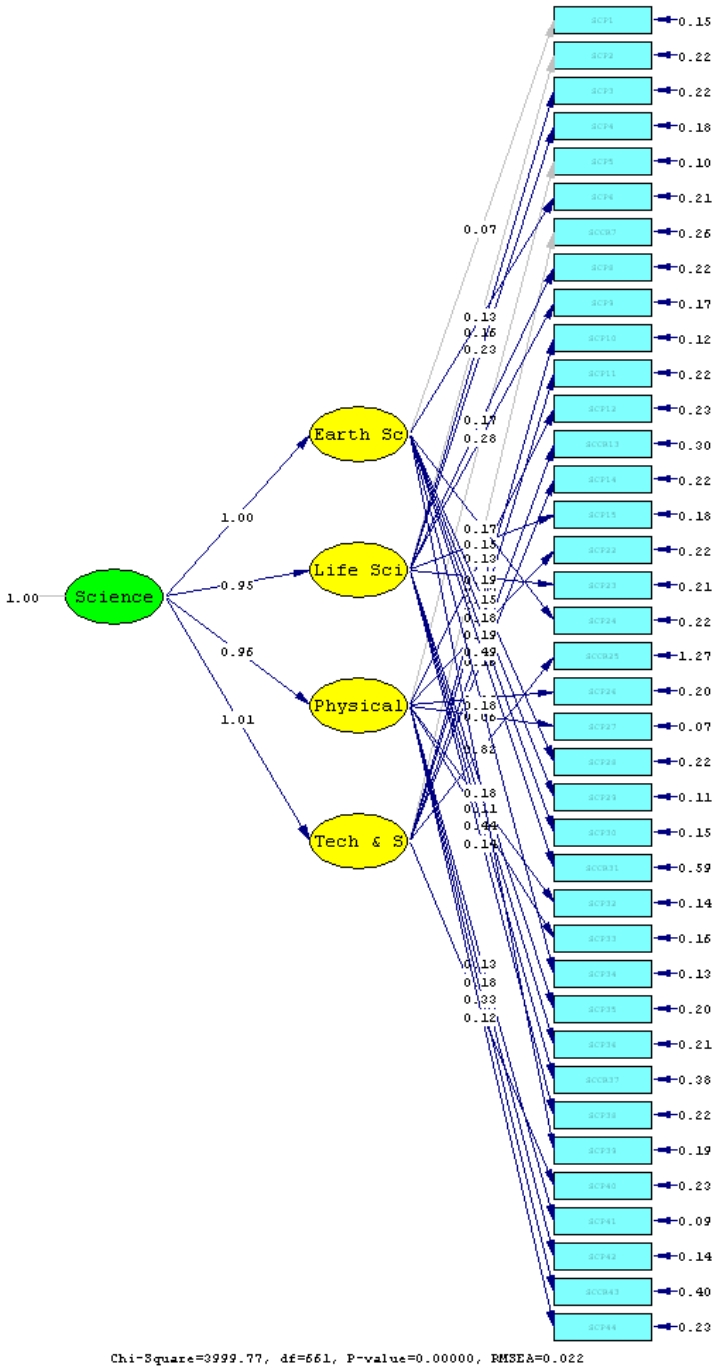


Table D3 - Data for 2nd order confirmatory factor analysis of Science

Goodness of Fit Statistics

Degrees of Freedom = 661
Minimum Fit Function Chi-Square = 3515.43 (P = 0.0)
Normal Theory Weighted Least Squares Chi-Square = 3999.77 (P = 0.0)
Estimated Non-centrality Parameter (NCP) = 3338.77
90 Percent Confidence Interval for NCP = (3143.11 ; 3541.81)

Minimum Fit Function Value = 0.35
Population Discrepancy Function Value (F0) = 0.33
90 Percent Confidence Interval for F0 = (0.31 ; 0.35)
Root Mean Square Error of Approximation (RMSEA) = 0.022
90 Percent Confidence Interval for RMSEA = (0.022 ; 0.023)
P-Value for Test of Close Fit (RMSEA < 0.05) = 1.00

Expected Cross-Validation Index (ECVI) = 0.42
90 Percent Confidence Interval for ECVI = (0.40 ; 0.44)
ECVI for Saturated Model = 0.15
ECVI for Independence Model = 14.76

Chi-Square for Independence Model with 703 Degrees of Freedom = 147542.26
Independence AIC = 147618.26
Model AIC = 4159.77
Saturated AIC = 1482.00
Independence CAIC = 147930.26
Model CAIC = 4816.60
Saturated CAIC = 7565.86

Normed Fit Index (NFI) = 0.98
Non-Normed Fit Index (NNFI) = 0.98
Parsimony Normed Fit Index (PNFI) = 0.92
Comparative Fit Index (CFI) = 0.98
Incremental Fit Index (IFI) = 0.98
Relative Fit Index (RFI) = 0.97

Critical N (CN) = 2130.04

Root Mean Square Residual (RMR) = 0.0060
Standardized RMR = 0.020
Goodness of Fit Index (GFI) = 0.98
Adjusted Goodness of Fit Index (AGFI) = 0.98
Parsimony Goodness of Fit Index (PGFI) = 0.87

Figure D4 – 2nd order confirmatory factor analysis model for Social Studies.

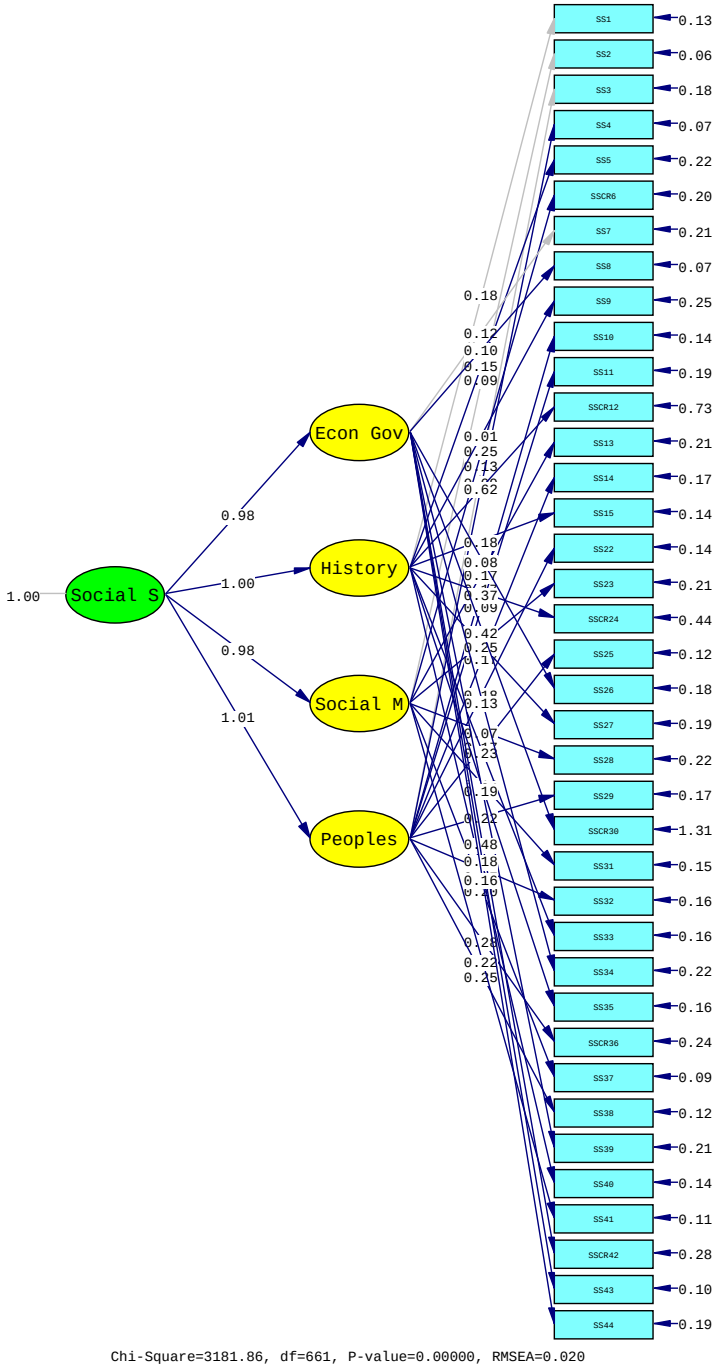


Table D4 - Data for 2nd order confirmatory factor analysis of Social Studies

Goodness of Fit Statistics

Degrees of Freedom = 661
Minimum Fit Function Chi-Square = 2932.82 (P = 0.0)
Normal Theory Weighted Least Squares Chi-Square = 3181.86 (P = 0.0)
Estimated Non-centrality Parameter (NCP) = 2520.86
90 Percent Confidence Interval for NCP = (2348.85 ; 2700.30)

Minimum Fit Function Value = 0.29
Population Discrepancy Function Value (F0) = 0.25
90 Percent Confidence Interval for F0 = (0.23 ; 0.27)
Root Mean Square Error of Approximation (RMSEA) = 0.020
90 Percent Confidence Interval for RMSEA = (0.019 ; 0.020)
P-Value for Test of Close Fit (RMSEA < 0.05) = 1.00

Expected Cross-Validation Index (ECVI) = 0.33
90 Percent Confidence Interval for ECVI = (0.32 ; 0.35)
ECVI for Saturated Model = 0.15
ECVI for Independence Model = 21.86

Chi-Square for Independence Model with 703 Degrees of Freedom = 218542.71
Independence AIC = 218618.71
Model AIC = 3341.86
Saturated AIC = 1482.00
Independence CAIC = 218930.70
Model CAIC = 3998.69
Saturated CAIC = 7565.86

Normed Fit Index (NFI) = 0.99
Non-Normed Fit Index (NNFI) = 0.99
Parsimony Normed Fit Index (PNFI) = 0.93
Comparative Fit Index (CFI) = 0.99
Incremental Fit Index (IFI) = 0.99
Relative Fit Index (RFI) = 0.99

Critical N (CN) = 2552.97

Root Mean Square Residual (RMR) = 0.0045
Standardized RMR = 0.017
Goodness of Fit Index (GFI) = 0.98
Adjusted Goodness of Fit Index (AGFI) = 0.98
Parsimony Goodness of Fit Index (PGFI) = 0.88

Figure D5 – 2nd order confirmatory factor analysis model for Writing.

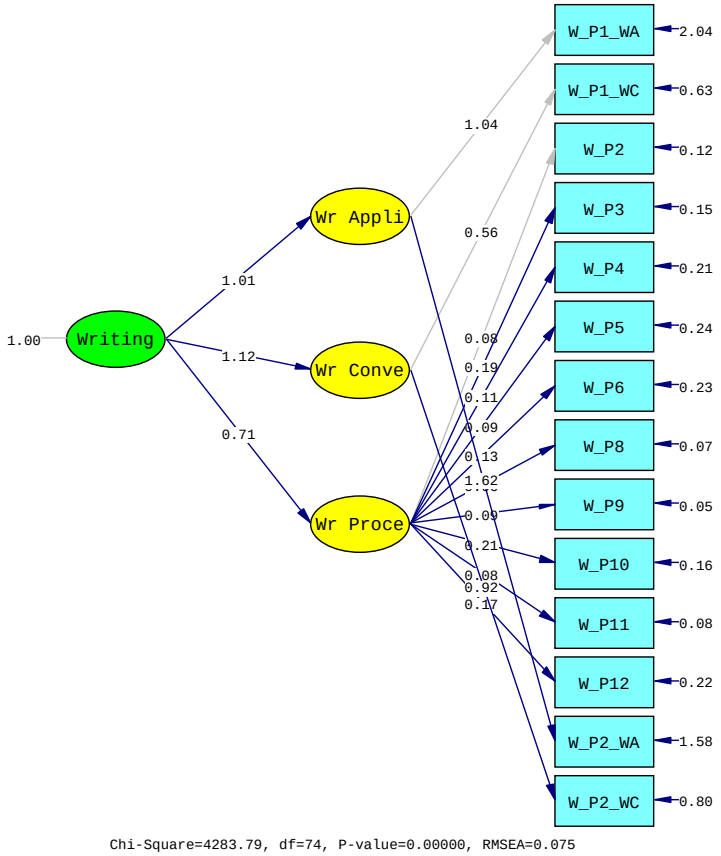


Table D5 - Data for 2nd order confirmatory factor analysis of Writing

Goodness of Fit Statistics

Degrees of Freedom = 74
Minimum Fit Function Chi-Square = 4486.31 (P = 0.0)
Normal Theory Weighted Least Squares Chi-Square = 4283.79 (P = 0.0)
Estimated Non-centrality Parameter (NCP) = 4209.79
90 Percent Confidence Interval for NCP = (3998.95 ; 4427.88)

Minimum Fit Function Value = 0.45
Population Discrepancy Function Value (F0) = 0.42
90 Percent Confidence Interval for F0 = (0.40 ; 0.44)
Root Mean Square Error of Approximation (RMSEA) = 0.075
90 Percent Confidence Interval for RMSEA = (0.074 ; 0.077)
P-Value for Test of Close Fit (RMSEA < 0.05) = 0.00

Expected Cross-Validation Index (ECVI) = 0.43
90 Percent Confidence Interval for ECVI = (0.41 ; 0.46)
ECVI for Saturated Model = 0.021
ECVI for Independence Model = 3.25

Chi-Square for Independence Model with 91 Degrees of Freedom = 32500.54
Independence AIC = 32528.54
Model AIC = 4345.79
Saturated AIC = 210.00
Independence CAIC = 32643.49
Model CAIC = 4600.31
Saturated CAIC = 1072.09

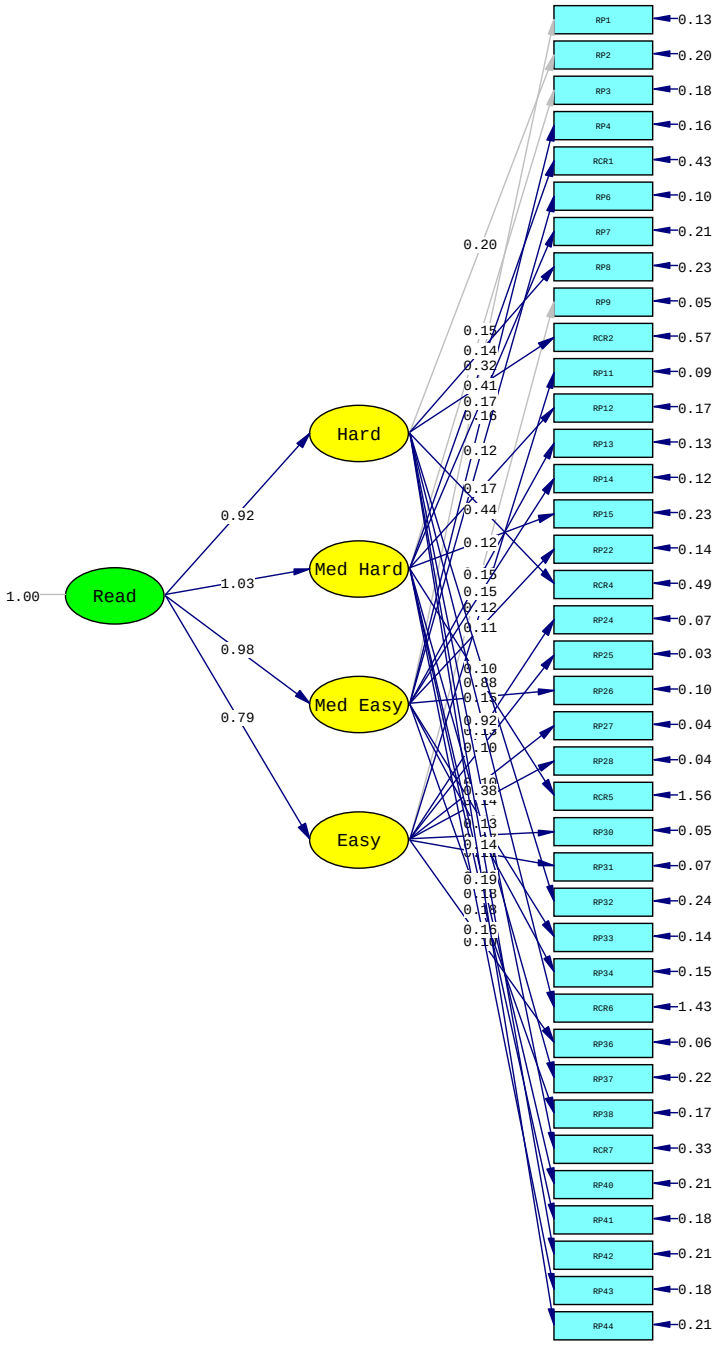
Normed Fit Index (NFI) = 0.86
Non-Normed Fit Index (NNFI) = 0.83
Parsimony Normed Fit Index (PNFI) = 0.70
Comparative Fit Index (CFI) = 0.86
Incremental Fit Index (IFI) = 0.86
Relative Fit Index (RFI) = 0.83

Critical N (CN) = 235.48

Root Mean Square Residual (RMR) = 0.063
Standardized RMR = 0.044
Goodness of Fit Index (GFI) = 0.94
Adjusted Goodness of Fit Index (AGFI) = 0.92
Parsimony Goodness of Fit Index (PGFI) = 0.66

Appendix E: 2nd Order Factor Analysis Models by Difficulty

Figure E1 – 2nd order confirmatory factor analysis model for Reading by difficulty



Chi-Square=4784.84, df=661, P-value=0.00000, RMSEA=0.025

Table E1 - Data for difficulty based 2nd order confirmatory factor analysis of Reading

Goodness of Fit Statistics

Degrees of Freedom = 661
Minimum Fit Function Chi-Square = 4327.76 (P = 0.0)
Normal Theory Weighted Least Squares Chi-Square = 4784.84 (P = 0.0)
Estimated Non-centrality Parameter (NCP) = 4123.84
90 Percent Confidence Interval for NCP = (3907.85 ; 4347.14)

Minimum Fit Function Value = 0.43
Population Discrepancy Function Value (F0) = 0.41
90 Percent Confidence Interval for F0 = (0.39 ; 0.43)
Root Mean Square Error of Approximation (RMSEA) = 0.025
90 Percent Confidence Interval for RMSEA = (0.024 ; 0.026)
P-Value for Test of Close Fit (RMSEA < 0.05) = 1.00

Expected Cross-Validation Index (ECVI) = 0.49
90 Percent Confidence Interval for ECVI = (0.47 ; 0.52)
ECVI for Saturated Model = 0.15
ECVI for Independence Model = 15.27

Chi-Square for Independence Model with 703 Degrees of Freedom = 152611.33
Independence AIC = 152687.33
Model AIC = 4944.84
Saturated AIC = 1482.00
Independence CAIC = 152999.32
Model CAIC = 5601.66
Saturated CAIC = 7565.86

Normed Fit Index (NFI) = 0.97
Non-Normed Fit Index (NNFI) = 0.97
Parsimony Normed Fit Index (PNFI) = 0.91
Comparative Fit Index (CFI) = 0.98
Incremental Fit Index (IFI) = 0.98
Relative Fit Index (RFI) = 0.97

Critical N (CN) = 1730.41

Root Mean Square Residual (RMR) = 0.011
Standardized RMR = 0.022
Goodness of Fit Index (GFI) = 0.98
Adjusted Goodness of Fit Index (AGFI) = 0.97
Parsimony Goodness of Fit Index (PGFI) = 0.87

Table E2 - Data for difficulty based 2nd order confirmatory factor analysis of Math

Goodness of Fit Statistics

Degrees of Freedom = 661
Minimum Fit Function Chi-Square = 4145.22 (P = 0.0)
Normal Theory Weighted Least Squares Chi-Square = 4568.03 (P = 0.0)
Estimated Non-centrality Parameter (NCP) = 3907.03
90 Percent Confidence Interval for NCP = (3696.47 ; 4124.93)

Minimum Fit Function Value = 0.41
Population Discrepancy Function Value (F0) = 0.39
90 Percent Confidence Interval for F0 = (0.37 ; 0.41)
Root Mean Square Error of Approximation (RMSEA) = 0.024
90 Percent Confidence Interval for RMSEA = (0.024 ; 0.025)
P-Value for Test of Close Fit (RMSEA < 0.05) = 1.00

Expected Cross-Validation Index (ECVI) = 0.47
90 Percent Confidence Interval for ECVI = (0.45 ; 0.49)
ECVI for Saturated Model = 0.15
ECVI for Independence Model = 24.51

Chi-Square for Independence Model with 703 Degrees of Freedom = 244997.00
Independence AIC = 245073.00
Model AIC = 4728.03
Saturated AIC = 1482.00
Independence CAIC = 245384.99
Model CAIC = 5384.86
Saturated CAIC = 7565.86

Normed Fit Index (NFI) = 0.98
Non-Normed Fit Index (NNFI) = 0.98
Parsimony Normed Fit Index (PNFI) = 0.92
Comparative Fit Index (CFI) = 0.99
Incremental Fit Index (IFI) = 0.99
Relative Fit Index (RFI) = 0.98

Critical N (CN) = 1806.57

Root Mean Square Residual (RMR) = 0.0057
Standardized RMR = 0.021
Goodness of Fit Index (GFI) = 0.98
Adjusted Goodness of Fit Index (AGFI) = 0.97
Parsimony Goodness of Fit Index (PGFI) = 0.87

Time used: 1.609 Seconds

Table E3 - Data for difficulty based 2nd order confirmatory factor analysis of Writing

Goodness of Fit Statistics

Degrees of Freedom = 86
Minimum Fit Function Chi-Square = 5185.56 (P = 0.0)
Normal Theory Weighted Least Squares Chi-Square = 5227.87 (P = 0.0)
Estimated Non-centrality Parameter (NCP) = 5141.87
90 Percent Confidence Interval for NCP = (4908.50 ; 5382.46)

Minimum Fit Function Value = 0.52
Population Discrepancy Function Value (F0) = 0.51
90 Percent Confidence Interval for F0 = (0.49 ; 0.54)
Root Mean Square Error of Approximation (RMSEA) = 0.077
90 Percent Confidence Interval for RMSEA = (0.076 ; 0.079)
P-Value for Test of Close Fit (RMSEA < 0.05) = 0.00

Expected Cross-Validation Index (ECVI) = 0.53
90 Percent Confidence Interval for ECVI = (0.51 ; 0.55)
ECVI for Saturated Model = 0.024
ECVI for Independence Model = 3.47

Chi-Square for Independence Model with 105 Degrees of Freedom = 34623.52
Independence AIC = 34653.52
Model AIC = 5295.87
Saturated AIC = 240.00
Independence CAIC = 34776.67
Model CAIC = 5575.02
Saturated CAIC = 1225.24

Normed Fit Index (NFI) = 0.85
Non-Normed Fit Index (NNFI) = 0.82
Parsimony Normed Fit Index (PNFI) = 0.70
Comparative Fit Index (CFI) = 0.85
Incremental Fit Index (IFI) = 0.85
Relative Fit Index (RFI) = 0.82

Critical N (CN) = 231.27

Root Mean Square Residual (RMR) = 0.063
Standardized RMR = 0.043
Goodness of Fit Index (GFI) = 0.93
Adjusted Goodness of Fit Index (AGFI) = 0.91
Parsimony Goodness of Fit Index (PGFI) = 0.67

Time used:0.953 Seconds

Table E4 - Data for difficulty based 2nd order confirmatory factor analysis of Science

Goodness of Fit Statistics

Degrees of Freedom = 661
Minimum Fit Function Chi-Square = 3287.09 (P = 0.0)
Normal Theory Weighted Least Squares Chi-Square = 3617.49 (P = 0.0)
Estimated Non-centrality Parameter (NCP) = 2956.49
90 Percent Confidence Interval for NCP = (2771.52 ; 3148.87)

Minimum Fit Function Value = 0.33
Population Discrepancy Function Value (F0) = 0.30
90 Percent Confidence Interval for F0 = (0.28 ; 0.31)
Root Mean Square Error of Approximation (RMSEA) = 0.021
90 Percent Confidence Interval for RMSEA = (0.020 ; 0.022)
P-Value for Test of Close Fit (RMSEA < 0.05) = 1.00

Expected Cross-Validation Index (ECVI) = 0.38
90 Percent Confidence Interval for ECVI = (0.36 ; 0.40)
ECVI for Saturated Model = 0.15
ECVI for Independence Model = 14.76

Chi-Square for Independence Model with 703 Degrees of Freedom = 147542.26
Independence AIC = 147618.26
Model AIC = 3777.49
Saturated AIC = 1482.00
Independence CAIC = 147930.26
Model CAIC = 4434.32
Saturated CAIC = 7565.86

Normed Fit Index (NFI) = 0.98
Non-Normed Fit Index (NNFI) = 0.98
Parsimony Normed Fit Index (PNFI) = 0.92
Comparative Fit Index (CFI) = 0.98
Incremental Fit Index (IFI) = 0.98
Relative Fit Index (RFI) = 0.98

Critical N (CN) = 2277.93

Root Mean Square Residual (RMR) = 0.0062
Standardized RMR = 0.019
Goodness of Fit Index (GFI) = 0.98
Adjusted Goodness of Fit Index (AGFI) = 0.98
Parsimony Goodness of Fit Index (PGFI) = 0.88

Time used:1.656 Seconds

Table E5 - Data for difficulty based 2nd order confirmatory factor analysis of Social Studies

Goodness of Fit Statistics

Degrees of Freedom = 661
Minimum Fit Function Chi-Square = 2785.71 (P = 0.0)
Normal Theory Weighted Least Squares Chi-Square = 2995.55 (P = 0.0)
Estimated Non-centrality Parameter (NCP) = 2334.55
90 Percent Confidence Interval for NCP = (2168.40 ; 2508.16)

Minimum Fit Function Value = 0.28
Population Discrepancy Function Value (F0) = 0.23
90 Percent Confidence Interval for F0 = (0.22 ; 0.25)
Root Mean Square Error of Approximation (RMSEA) = 0.019
90 Percent Confidence Interval for RMSEA = (0.018 ; 0.019)
P-Value for Test of Close Fit (RMSEA < 0.05) = 1.00

Expected Cross-Validation Index (ECVI) = 0.32
90 Percent Confidence Interval for ECVI = (0.30 ; 0.33)
ECVI for Saturated Model = 0.15
ECVI for Independence Model = 21.86

Chi-Square for Independence Model with 703 Degrees of Freedom = 218542.71
Independence AIC = 218618.71
Model AIC = 3155.55
Saturated AIC = 1482.00
Independence CAIC = 218930.70
Model CAIC = 3812.38
Saturated CAIC = 7565.86

Normed Fit Index (NFI) = 0.99
Non-Normed Fit Index (NNFI) = 0.99
Parsimony Normed Fit Index (PNFI) = 0.93
Comparative Fit Index (CFI) = 0.99
Incremental Fit Index (IFI) = 0.99
Relative Fit Index (RFI) = 0.99

Critical N (CN) = 2687.74

Root Mean Square Residual (RMR) = 0.0039
Standardized RMR = 0.016
Goodness of Fit Index (GFI) = 0.98
Adjusted Goodness of Fit Index (AGFI) = 0.98
Parsimony Goodness of Fit Index (PGFI) = 0.88

Time used:1.672 Seconds

Table E6 – Classifications of items by difficulty

Key to the strand data in Table E6

Content area	Code	Description from Ohio Academic Content Standards
Mathematics	ALG	Patterns, Functions and Algebra
	DATA	Data analysis and Probability
	GEO	Geometry and Spatial Sense
	MEAS	Measurement
	NUMS	Number, Number Sense and Operations
Reading	AV	Acquisition of Vocabulary
	IT	Informational, Technical and Persuasive text
	LT	Literary Text
	RP	Reading Process
Science	ES	Earth Science
	LS	Life Science
	PS	Physical Science
	SK	Scientific Ways of Knowing
	SI	Scientific Inquiry
	ST	Science and Technology
Social Studies	ECON	Economics
	GEO	Geography
	GOV	Government
	HIST	History
	PEOP	People in Societies
	SO	Social Studies Skills and Methods
Writing		All categories are self explanatory

Table E6 – Classifications of items by difficulty

MATH			
Classification	Strand	Item	Rasch
E	GEO	11	-3.059
E	DATA	2	-2.219
E	DATA	5	-1.376
E	MEAS	3	-1.332
E	DATA	9	-1.152
E	DATA	14	-0.925
E	NUMS	8	-0.892
E	DATA	27	-0.771
E	NUMS	23	-0.606
ME	NUMS	26	-0.445
ME	ALG	1	-0.435
ME	ALG	15	-0.359
ME	MEAS	7	-0.326
ME	ALG	41	-0.253
ME	ALG	36	-0.23
ME	GEO	29	-0.13
ME	GEO	37	-0.089
ME	ALG	6	-0.075
Field test		16	0
Field test		17	0
Field test		18	0
Field test		19	0
Field test		20	0
Field test		21	0
ME	DATA	39	0.073
MH	GEO	25	0.104
MH	MEAS	13	0.151
MH	MEAS	12	0.42
MH	NUMS	33	0.509
MH	DATA	22	0.574
MH	DATA	31	0.696
MH	NUMS	38	0.788
MH	NUMS	4	0.943
MH	ALG	10	0.998
MH	MEAS	30	1.084
H	DATA	35	1.092
H	NUMS	43	1.151
H	ALG	24	1.183
H	ALG	28	1.27
H	MEAS	42	1.31
H	DATA	44	1.356
H	ALG	32	1.442
H	MEAS	34	2.151
H	GEO	40	3.191

READING			
Classification	Strand	Item	Rasch
E	AV	25	-1.889
E	RP	27	-1.79
E	RP	9	-1.75
E	LT	36	-1.557
E	IT	30	-1.547
E	RP	28	-1.532
E	LT	31	-1.205
E	IT	24	-1.168
E	RP	11	-0.996
ME	RP	6	-0.988
ME	IT	26	-0.639
ME	IT	14	-0.427
ME	AV	1	-0.415
ME	IT	13	-0.376
ME	IT	4	-0.311
ME	IT	22	-0.258
ME	LT	33	-0.159
ME	LT	34	-0.051
ME	RP	38	-0.033
Field test		16	0
Field test		17	0
Field test		18	0
Field test		19	0
Field test		20	0
Field test		21	0
MH	LT	41	0.172
MH	RP	43	0.188
MH	IT	3	0.203
MH	IT	12	0.313
MH	LT	40	0.414
MH	IT	29	0.735
MH	RP	5	0.738
MH	AV	37	0.754
MH	IT	7	0.783
MH	AV	15	0.787
H	AV	2	0.79
H	LT	39	0.904
H	RP	44	0.92
H	AV	23	0.929
H	LT	35	0.997
H	AV	8	1.018
H	RP	32	1.142
H	IT	10	1.37
H	LT	42	1.808

Table E6 – Classifications of items by difficulty (continued)

Science Classification	Strand	Item	Rasch
E	PS	27	-1.871
E	PS	5	-1.412
E	PS	41	-1.363
E	ES	29	-0.915
E	ES	1	-0.745
E	SK	10	-0.735
E	ES	34	-0.603
E	PS	42	-0.601
E	PS	32	-0.589
ME	PS	33	-0.552
ME	ES	30	-0.349
ME	LS	15	-0.323
Field test		16	0
Field test		17	0
Field test		18	0
Field test		19	0
Field test		20	0
Field test		21	0
ME	SK	7	0.009
ME	LS	4	0.076
ME	ES	6	0.157
ME	PS	43	0.223
ME	ES	39	0.249
ME	SI	14	0.294
ME	PS	26	0.325
MH	ST	40	0.333
MH	ES	28	0.416
MH	LS	2	0.429
MH	ST	25	0.449
MH	SI	11	0.471
MH	LS	37	0.491
MH	PS	22	0.614
MH	LS	8	0.647
MH	LS	23	0.665
MH	PS	12	0.755
H	LS	3	0.883
H	LS	38	0.908
H	ES	24	0.991
H	PS	44	1.015
H	LS	9	1.076
H	LS	36	1.334
H	LS	35	1.385
H	ES	31	1.396
H	SI	13	1.96

Social Studies Classification	Strand	Item	Rasch
E	GOV	8	-1.599
E	GEOG	4	-1.572
E	GOV	43	-1.108
E	SO	37	-0.847
E	SO	31	-0.642
E	GEOG	38	-0.593
E	PEOP	25	-0.467
E	HIST	40	-0.443
E	GEOG	22	-0.383
ME	HIST	15	-0.3711
ME	PEOP	10	-0.348
ME	SO	41	-0.348
ME	HIST	1	-0.252
ME	SO	2	-0.179
ME	GEOG	11	-0.178
ME	PEOP	3	-0.125
ME	CIT	44	-0.018
Field test		16	0
Field test		17	0
Field test		18	0
Field test		19	0
Field test		20	0
Field test		21	0
ME	GOV	26	0.057
ME	HIST	33	0.058
MH	CIT	42	0.116
MH	HIST	35	0.139
MH	PEOP	29	0.188
MH	PEOP	14	0.201
MH	PEOP	32	0.221
MH	ECON	7	0.225
MH	CIT	34	0.329
MH	SO	23	0.606
MH	HIST	9	0.667
MH	HIST	27	0.815
H	HIST	24	0.917
H	ECON	39	0.987
H	SO	13	1.22
H	HIST	5	1.241
H	SO	6	1.274
H	HIST	12	1.592
H	SO	28	1.628
H	PEOP	36	1.949
H	ECON	30	1.99

Table E6 – Classifications of items by difficulty (continued)

Writing			
Classification	**Item**	**Rasch**	**Strand**
E	9	-1.477	Process
E	8	-1.395	Process
E	11	-1.136	Process
E	2	-0.969	Process
ME	7	-0.936	Applications
ME	13	-0.8291	Conventions
ME	1	-0.268	Conventions
ME	3	-0.161	Process
MH	4	0.353	Process
MH	10	0.685	Process
MH	5	1.057	Process
H	13	1.097	Applications
H	12	1.254	Process
H	6	1.323	Process
H	1	1.873	Applications

Appendix F: Split of content standards

Table F1 – The split of Mathematics items into two equivalent groups

SP06			GROUP 1			GROUP 2		
Item	Rasch	Points	PTS	Applied?	Rasch	PTS	Applied?	Rasch
2	-2.219	1	1	1*	-2.219	0		0
3	-1.332	1	1	1	-1.332	0		0
6	-0.075	1	1	1	-0.075	0		0
8	-0.892	1	1	1	-0.892	0		0
9	-1.152	1	1	1	-1.152	0		0
10	0.998	2	2	1	0.998	0		0
12	0.42	1	1	1	0.42	0		0
15	-0.359	1	1	1	-0.359	0		0
23	-0.606	1	1	1	-0.606	0		0
24	1.183	1	1	1	1.183	0		0
25	0.104	1	1	1	0.104	0		0
27	-0.771	1	1	1	-0.771	0		0
29	-0.13	1	1	1	-0.13	0		0
33	0.509	1	1	1	0.509	0		0
35	1.092	1	1	1	1.092	0		0
38	0.788	1	1	1	0.788	0		0
40	3.191	4	4	1	3.191	0		0
41	-0.253	1	1	1	-0.253	0		0
44	1.356	1	1	1	1.356	0		0
1	-0.435	1	0		0	1	1	-0.435
4	0.943	2	0		0	2	1	0.943
5	-1.376	1	0		0	1	1	-1.376
7	-0.326	1	0		0	1	1	-0.326
11	-3.059	1	0		0	1	1	-3.059
13	0.151	1	0		0	1	1	0.151
14	-0.925	1	0		0	1	1	-0.925
22	0.574	2	0		0	2	1	0.574
26	-0.445	1	0		0	1	1	-0.445
28	1.27	2	0		0	2	1	1.27
30	1.084	1	0		0	1	1	1.084
31	0.696	1	0		0	1	1	0.696
32	1.442	1	0		0	1	1	1.442
34	2.151	2	0		0	2	1	2.151
36	-0.23	1	0		0	1	1	-0.23
37	-0.089	1	0		0	1	1	-0.089
39	0.073	1	0		0	1	1	0.073
42	1.31	1	0		0	1	1	1.31
43	1.151	1	0		0	1	1	1.151
	SUMS:	46	23	19		23	19	
	AVERAGE RASCH:				0.042091			0.09

* The value "1" in the "Applied?" column indicates that the item was included in the group. A blank indicates omission from the group.

Table F2 – The split of Reading items into two equivalent groups

SP06			GROUP 1			GROUP 2		
Item	**Rasch**	**Points**	**PTS**	**Applied?**	**Rasch**	**PTS**	**Applied?**	**Rasch**
2	0.79	1	1	1	0.79	0		0
3	0.203	1	1	1	0.203	0		0
6	-0.988	1	1	1	-0.988	0		0
7	0.783	1	1	1	0.783	0		0
10	1.37	2	2	1	1.37	0		0
13	-0.376	1	1	1	-0.376	0		0
22	-0.258	1	1	1	-0.258	0		0
24	-1.168	1	1	1	-1.168	0		0
25	-1.889	1	1	1	-1.889	0		0
26	-0.639	1	1	1	-0.639	0		0
28	-1.532	1	1	1	-1.532	0		0
29	0.735	4	4	1	0.735	0		0
32	1.142	1	1	1	1.142	0		0
34	-0.051	1	1	1	-0.051	0		0
36	-1.557	1	1	1	-1.557	0		0
39	0.904	2	2	1	0.904	0		0
40	0.414	1	1	1	0.414	0		0
41	0.172	1	1	1	0.172	0		0
44	0.92	1	1	1	0.92	0		0
1	-0.415	1	0		0	1	1	-0.415
4	-0.311	1	0		0	1	1	-0.311
5	0.738	2	0		0	2	1	0.738
8	1.018	1	0		0	1	1	1.018
9	-1.75	1	0		0	1	1	-1.75
11	-0.996	1	0		0	1	1	-0.996
12	0.313	1	0		0	1	1	0.313
14	-0.427	1	0		0	1	1	-0.427
15	0.787	1	0		0	1	1	0.787
23	0.929	2	0		0	2	1	0.929
27	-1.79	1	0		0	1	1	-1.79
30	-1.547	1	0		0	1	1	-1.547
31	-1.205	1	0		0	1	1	-1.205
33	-0.159	1	0		0	1	1	-0.159
35	0.997	4	0		0	4	1	0.997
37	0.754	1	0		0	1	1	0.754
38	-0.033	1	0		0	1	1	-0.033
42	1.808	1	0		0	1	1	1.808
43	0.188	1	0		0	1	1	0.188
	SUMS:	48	24	19		24	19	
	AVERAGE RASCH:				-0.0233			-0.02502

Table F3 – The split of Writing items into two equivalent groups

SP06				GROUP 1			GROUP 2		
Item	**Rasch**	**Points**	**Type**	**PTS**	**Applied?**	**Rasch**	**PTS**	**Applied?**	**Rasch**
1	1.873	12	Applications	12	1	1.873	0		0
3	-0.161	1		1	1	-0.161	0		0
5	1.057	1		1	1	1.057	0		0
6	1.323	1		1	1	1.323	0		0
7	-0.936	2	Applications	2	1	-0.936	0		0
11	-1.136	1		1	1	-1.136	0		0
13	-0.8291	6	Conventions	6	1	-0.8291	0		0
1	-0.268	6	Conventions	0		0	6	1	-0.268
2	-0.969	1		0		0	1	1	-0.969
4	0.353	1		0		0	1	1	0.353
8	-1.395	1		0		0	1	1	-1.395
9	-1.477	1		0		0	1	1	-1.477
10	0.685	1		0		0	1	1	0.685
12	1.254	1		0		0	1	1	1.254
13	1.097	12	Applications	0		0	12	1	1.097
	SUMS:	48		24	7		24	8	
	AVERAGE RASCH:					-0.03789			-0.04

Table F4 – The split of Science items into two equivalent groups

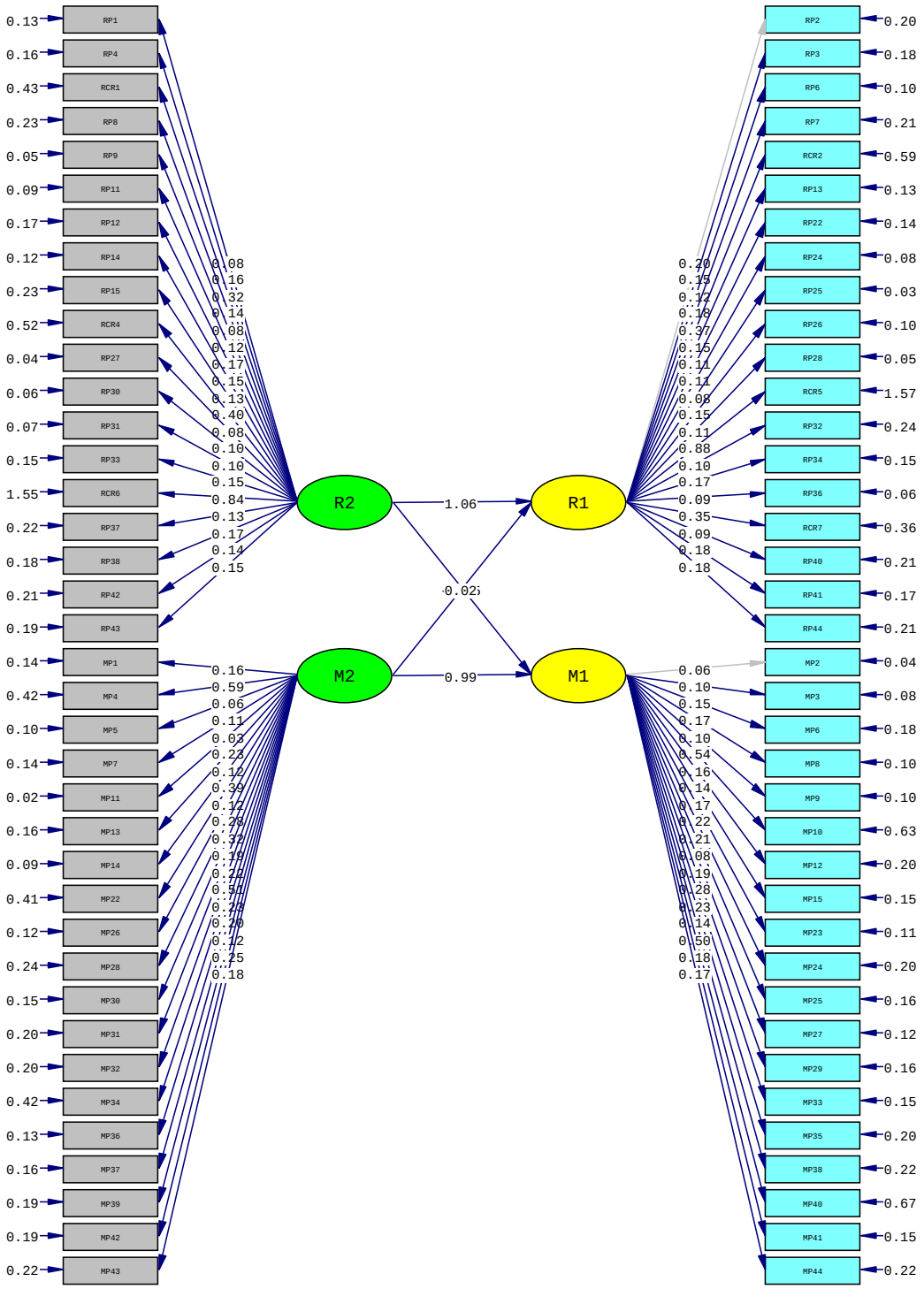
SP06			GROUP 1			GROUP 2		
Item	**Rasch**	**Points**	**PTS**	**Applied?**	**Rasch**	**PTS**	**Applied?**	**Rasch**
1	-0.745	1	1	1	-0.745	0		0
4	0.076	1	1	1	0.076	0		0
5	-1.412	1	1	1	-1.412	0		0
8	0.647	1	1	1	0.647	0		0
11	0.471	1	1	1	0.471	0		0
12	0.755	1	1	1	0.755	0		0
13	1.96	2	2	1	1.96	0		0
25	0.449	4	4	1	0.449	0		0
26	0.325	1	1	1	0.325	0		0
28	0.416	1	1	1	0.416	0		0
30	-0.349	1	1	1	-0.349	0		0
32	-0.589	1	1	1	-0.589	0		0
34	-0.603	1	1	1	-0.603	0		0
36	1.334	1	1	1	1.334	0		0
38	0.908	1	1	1	0.908	0		0
39	0.249	1	1	1	0.249	0		0
41	-1.363	1	1	1	-1.363	0		0
43	0.223	2	2	1	0.223	0		0
44	1.015	1	1	1	1.015	0		0
2	0.429	1	0		0	1	1	0.429
3	0.883	1	0		0	1	1	0.883
6	0.157	1	0		0	1	1	0.157
7	0.009	2	0		0	2	1	0.009
9	1.076	1	0		0	1	1	1.076
10	-0.735	1	0		0	1	1	-0.735
14	0.294	1	0		0	1	1	0.294
15	-0.323	1	0		0	1	1	-0.323
22	0.614	1	0		0	1	1	0.614
23	0.665	1	0		0	1	1	0.665
24	0.991	1	0		0	1	1	0.991
27	-1.871	1	0		0	1	1	-1.871
29	-0.915	1	0		0	1	1	-0.915
31	1.396	4	0		0	4	1	1.396
33	-0.552	1	0		0	1	1	-0.552
35	1.385	1	0		0	1	1	1.385
37	0.491	2	0		0	2	1	0.491
40	0.333	1	0		0	1	1	0.333
42	-0.601	1	0		0	1	1	-0.601
	SUMS:	48	24	19		24	19	
	AVERAGE RASCH:				0.085614			0.084682

Table F5 – The split of Social Studies items into two equivalent groups

SP06			GROUP 1			GROUP 2		
Item	**Rasch**	**Points**	**PTS**	**Applied?**	**Rasch**	**PTS**	**Applied?**	**Rasch**
2	-0.179	1	1	1	-0.179	0		0
8	-1.599	1	1	1	-1.599	0		0
9	0.667	1	1	1	0.667	0		0
10	-0.348	1	1	1	-0.348	0		0
11	-0.178	1	1	1	-0.178	0		0
12	1.592	4	4	1	1.592	0		0
13	1.22	1	1	1	1.22	0		0
22	-0.383	1	1	1	-0.383	0		0
23	0.606	1	1	1	0.606	0		0
25	-0.467	1	1	1	-0.467	0		0
28	1.628	1	1	1	1.628	0		0
29	0.188	1	1	1	0.188	0		0
31	-0.642	1	1	1	-0.642	0		0
32	0.221	1	1	1	0.221	0		0
33	0.058	1	1	1	0.058	0		0
36	1.949	2	2	1	1.949	0		0
42	0.116	2	2	1	0.116	0		0
43	-1.108	1	1	1	-1.108	0		0
44	-0.018	1	1	1	-0.018	0		0
1	-0.252	1	0		0	1	1	-0.252
3	-0.125	1	0		0	1	1	-0.125
4	-1.572	1	0		0	1	1	-1.572
5	1.241	1	0		0	1	1	1.241
6	1.274	2	0		0	2	1	1.274
7	0.225	1	0		0	1	1	0.225
14	0.201	1	0		0	1	1	0.201
15	-0.3711	1	0		0	1	1	-0.3711
24	0.917	2	0		0	2	1	0.917
26	0.057	1	0		0	1	1	0.057
27	0.815	1	0		0	1	1	0.815
30	1.99	4	0		0	4	1	1.99
34	0.329	1	0		0	1	1	0.329
35	0.139	1	0		0	1	1	0.139
37	-0.847	1	0		0	1	1	-0.847
38	-0.593	1	0		0	1	1	-0.593
39	0.987	1	0		0	1	1	0.987
40	-0.443	1	0		0	1	1	-0.443
41	-0.348	1	0		0	1	1	-0.348
	SUMS:	48	24	19		24	19	
	AVERAGE RASCH:				0.075523			0.082361

Appendix G: Regression models

Figure G-1 – Full information regression model of the Reading/Mathematics dyad



Chi-Square=20928.55, df=2769, P-value=0.00000, RMSEA=0.026

Table G-1a – Results of the Full information regression model of the Reading/Mathematics dyad

Structural Equations

```
   R1 = 1.06*R2 - 0.051*M2, Errorvar.= -0.044 , R² = 1.04
       (0.031)   (0.017)               (0.0089)
       34.64     -2.93                 -4.91

W_A_R_N_I_N_G : Error variance is negative.

   M1 = 0.018*R2 + 0.99*M2, Errorvar.= -0.016 , R² = 1.02
       (0.013)    (0.035)              (0.0060)
        1.40      28.79                -2.58

W_A_R_N_I_N_G : Error variance is negative.


    Correlation Matrix of Independent Variables

        R2     M2
     --------  --------
   R2    1.00
   M2    0.77    1.00
        (0.01)
       113.05

    Covariance Matrix of Latent Variables

        R1     M1     R2     M2
     --------  --------  --------  --------
   R1    1.00
   M1    0.78    1.00
   R2    1.02    0.78    1.00
   M2    0.76    1.01    0.77    1.00
```

Table G-1b – Results of the Full information regression model of the Reading/Mathematics dyad

Goodness of Fit Statistics

Degrees of Freedom = 2769
Minimum Fit Function Chi-Square = 15577.58 (P = 0.0)
Normal Theory Weighted Least Squares Chi-Square = 20928.55 (P = 0.0)
Estimated Non-centrality Parameter (NCP) = 18159.55
90 Percent Confidence Interval for NCP = (17702.58 ; 18623.15)

Minimum Fit Function Value = 1.56
Population Discrepancy Function Value (F0) = 1.82
90 Percent Confidence Interval for F0 = (1.77 ; 1.86)
Root Mean Square Error of Approximation (RMSEA) = 0.026
90 Percent Confidence Interval for RMSEA = (0.025 ; 0.026)
P-Value for Test of Close Fit (RMSEA < 0.05) = 1.00

Expected Cross-Validation Index (ECVI) = 2.12
90 Percent Confidence Interval for ECVI = (2.08 ; 2.17)
ECVI for Saturated Model = 0.59
ECVI for Independence Model = 63.91

Chi-Square for Independence Model with 2850 Degrees of Freedom = 638914.02
Independence AIC = 639066.02
Model AIC = 21242.55
Saturated AIC = 5852.00
Independence CAIC = 639690.01
Model CAIC = 22531.58
Saturated CAIC = 29875.46

Normed Fit Index (NFI) = 0.98
Non-Normed Fit Index (NNFI) = 0.98
Parsimony Normed Fit Index (PNFI) = 0.95
Comparative Fit Index (CFI) = 0.98
Incremental Fit Index (IFI) = 0.98
Relative Fit Index (RFI) = 0.97

Critical N (CN) = 1891.39

Root Mean Square Residual (RMR) = 0.0089
Standardized RMR = 0.025
Goodness of Fit Index (GFI) = 0.95
Adjusted Goodness of Fit Index (AGFI) = 0.94
Parsimony Goodness of Fit Index (PGFI) = 0.90

Table G-2a – Results of the Full information regression model of the Mathematics/Writing dyad

Structural Equations

```
   M1 = 0.99*M2 + 0.022*W2, Errorvar.= -0.015 , R² = 1.02
       (0.034)   (0.010)               (0.0060)
       28.83      2.17                 -2.55

W_A_R_N_I_N_G : Error variance is negative.

   W1 = - 0.33*M2 + 1.51*W2, Errorvar.= -0.65 , R² = 1.65
         (0.028)   (0.031)              (0.023)
        -12.07     49.03               -27.84

W_A_R_N_I_N_G : Error variance is negative.


    Correlation Matrix of Independent Variables

        M2     W2
     --------  --------
   M2    1.00
   W2    0.74    1.00
        (0.01)
        90.02

    Covariance Matrix of Latent Variables

        M1     W1     M2     W2
     --------  --------  --------  --------
   M1    1.00
   W1    0.80    1.00
   M2    1.01    0.78    1.00
   W2    0.75    1.26    0.74    1.00
```

Table G-2b – Results of the Full information regression model of the Mathematics/Writing dyad

Goodness of Fit Statistics

Degrees of Freedom = 1320
Minimum Fit Function Chi-Square = 12741.64 (P = 0.0)
Normal Theory Weighted Least Squares Chi-Square = 15491.30 (P = 0.0)
Estimated Non-centrality Parameter (NCP) = 14171.30
90 Percent Confidence Interval for NCP = (13773.78 ; 14575.63)

Minimum Fit Function Value = 1.27
Population Discrepancy Function Value (F0) = 1.42
90 Percent Confidence Interval for F0 = (1.38 ; 1.46)
Root Mean Square Error of Approximation (RMSEA) = 0.033
90 Percent Confidence Interval for RMSEA = (0.032 ; 0.033)
P-Value for Test of Close Fit (RMSEA < 0.05) = 1.00

Expected Cross-Validation Index (ECVI) = 1.57
90 Percent Confidence Interval for ECVI = (1.53 ; 1.61)
ECVI for Saturated Model = 0.29
ECVI for Independence Model = 39.63

Chi-Square for Independence Model with 1378 Degrees of Freedom = 396113.53
Independence AIC = 396219.53
Model AIC = 15713.30
Saturated AIC = 2862.00
Independence CAIC = 396654.68
Model CAIC = 16624.65
Saturated CAIC = 14611.00

Normed Fit Index (NFI) = 0.97
Non-Normed Fit Index (NNFI) = 0.97
Parsimony Normed Fit Index (PNFI) = 0.93
Comparative Fit Index (CFI) = 0.97
Incremental Fit Index (IFI) = 0.97
Relative Fit Index (RFI) = 0.97

Critical N (CN) = 1132.97

Root Mean Square Residual (RMR) = 0.024
Standardized RMR = 0.033
Goodness of Fit Index (GFI) = 0.94
Adjusted Goodness of Fit Index (AGFI) = 0.94
Parsimony Goodness of Fit Index (PGFI) = 0.87

Table G-3a – Results of the Full information regression model of the Mathematics/Science dyad

Structural Equations

```
M1 = 1.05*M2 - 0.016*SC2, Errorvar.= -0.016  , R² = 1.01
    (0.044)   (0.026)               (0.0066)
    24.10     -0.62                 -2.41

W_A_R_N_I_N_G : Error variance is negative.

SC1 =  - 0.12*M2 + 1.17*SC2, Errorvar.= -0.057 , R² = 1.05
       (0.038)   (0.075)               (0.012)
       -3.22     15.62                 -4.92

W_A_R_N_I_N_G : Error variance is negative.
```

Correlation Matrix of Independent Variables

	M2	SC2
M2	1.00	
SC2	0.90	1.00
	(0.00)	
	196.51	

Covariance Matrix of Latent Variables

	M1	SC1	M2	SC2
M1	1.06			
SC1	0.97	1.08		
M2	1.04	0.94	1.00	
SC2	0.94	1.06	0.90	1.00

Table G-3b – Results of the Full information regression model of the Mathematics/Science dyad

Goodness of Fit Statistics

Degrees of Freedom = 2769
Minimum Fit Function Chi-Square = 13059.86 (P = 0.0)
Normal Theory Weighted Least Squares Chi-Square = 17211.70 (P = 0.0)
Estimated Non-centrality Parameter (NCP) = 14442.70
90 Percent Confidence Interval for NCP = (14031.94 ; 14860.28)

Minimum Fit Function Value = 1.31
Population Discrepancy Function Value (F0) = 1.44
90 Percent Confidence Interval for F0 = (1.40 ; 1.49)
Root Mean Square Error of Approximation (RMSEA) = 0.023
90 Percent Confidence Interval for RMSEA = (0.023 ; 0.023)
P-Value for Test of Close Fit (RMSEA < 0.05) = 1.00

Expected Cross-Validation Index (ECVI) = 1.75
90 Percent Confidence Interval for ECVI = (1.71 ; 1.79)
ECVI for Saturated Model = 0.59
ECVI for Independence Model = 70.71

Chi-Square for Independence Model with 2850 Degrees of Freedom = 706897.08
Independence AIC = 707049.08
Model AIC = 17525.70
Saturated AIC = 5852.00
Independence CAIC = 707673.07
Model CAIC = 18814.72
Saturated CAIC = 29875.46

Normed Fit Index (NFI) = 0.98
Non-Normed Fit Index (NNFI) = 0.98
Parsimony Normed Fit Index (PNFI) = 0.95
Comparative Fit Index (CFI) = 0.99
Incremental Fit Index (IFI) = 0.99
Relative Fit Index (RFI) = 0.98

Critical N (CN) = 2255.82

Root Mean Square Residual (RMR) = 0.0060
Standardized RMR = 0.020
Goodness of Fit Index (GFI) = 0.96
Adjusted Goodness of Fit Index (AGFI) = 0.95
Parsimony Goodness of Fit Index (PGFI) = 0.91

Table G-4a – Results of the Full information regression model of the Mathematics/Social Studies dyad

Structural Equations

```
 M1 = 1.27*M2 - 0.27*SSt2, Errorvar.= -0.058 , R² = 1.06
     (0.062)   (0.047)                (0.0097)
     20.65     -5.70                  -5.98

W_A_R_N_I_N_G : Error variance is negative.

SSt1 = - 0.35*M2 + 1.33*SSt2, Errorvar.= -0.019 , R² = 1.02
        (0.035)   (0.049)                (0.0062)
        -9.86     27.14                  -3.02

W_A_R_N_I_N_G : Error variance is negative.
```

Correlation Matrix of Independent Variables

	M2	SSt2
M2	1.00	
SSt2	0.94	1.00
	(0.00)	
	272.57	

Covariance Matrix of Latent Variables

	M1	SSt1	M2	SSt2
M1	1.00			
SSt1	0.88	1.00		
M2	1.02	0.90	1.00	
SSt2	0.93	1.00	0.94	1.00

Table G-4b – Results of the Full information regression model of the Mathematics/Social Studies dyad

Goodness of Fit Statistics

Degrees of Freedom = 2769
Minimum Fit Function Chi-Square = 110707.71 (P = 0.0)
Normal Theory Weighted Least Squares Chi-Square = 454577.59 (P = 0.0)
Estimated Non-centrality Parameter (NCP) = 451808.59
90 Percent Confidence Interval for NCP = (0.0 ; 0.0)

Minimum Fit Function Value = 11.07
Population Discrepancy Function Value (F0) = 45.19
90 Percent Confidence Interval for F0 = (0.0 ; 0.0)
Root Mean Square Error of Approximation (RMSEA) = 0.13
90 Percent Confidence Interval for RMSEA = (0.0 ; 0.0)
P-Value for Test of Close Fit (RMSEA < 0.05) = 1.00

Expected Cross-Validation Index (ECVI) = 45.49
90 Percent Confidence Interval for ECVI = (0.31 ; 0.31)
ECVI for Saturated Model = 0.59
ECVI for Independence Model = 124.71

Chi-Square for Independence Model with 2850 Degrees of Freedom = 1246789.21
Independence AIC = 1246941.21
Model AIC = 454891.59
Saturated AIC = 5852.00
Independence CAIC = 1247565.20
Model CAIC = 456180.61
Saturated CAIC = 29875.46

Normed Fit Index (NFI) = 0.91
Non-Normed Fit Index (NNFI) = 0.91
Parsimony Normed Fit Index (PNFI) = 0.89
Comparative Fit Index (CFI) = 0.91
Incremental Fit Index (IFI) = 0.91
Relative Fit Index (RFI) = 0.91

Critical N (CN) = 266.99

Root Mean Square Residual (RMR) = 0.040
Standardized RMR = 0.11
Goodness of Fit Index (GFI) = 0.46
Adjusted Goodness of Fit Index (AGFI) = 0.42
Parsimony Goodness of Fit Index (PGFI) = 0.43

Table G-5a – Results of the Full information regression model of the Reading/Writing dyad

Structural Equations

```
   R1 = 1.04*R2 - 0.024*W2, Errorvar.= -0.041 , R² = 1.04
       (0.032)   (0.017)               (0.0081)
       32.95     -1.42                 -5.00

W_A_R_N_I_N_G : Error variance is negative.

   W1 = - 0.77*R2 + 1.92*W2, Errorvar.= -0.74 , R² = 1.74
         (0.065)   (0.066)              (0.033)
        -11.85     28.95               -22.04

W_A_R_N_I_N_G : Error variance is negative.


    Correlation Matrix of Independent Variables

        R2     W2
     --------  --------
   R2    1.00
   W2    0.86    1.00
        (0.01)
       126.94

    Covariance Matrix of Latent Variables

        R1     W1     R2     W2
     --------  --------  --------  --------
   R1    1.00
   W1    0.90    1.00
   R2    1.02    0.89    1.00
   W2    0.87    1.26    0.86    1.00

W_A_R_N_I_N_G: Matrix above is not positive definite
```

Table G-5b – Results of the Full information regression model of the Reading/Writing dyad

Goodness of Fit Statistics

Degrees of Freedom = 1320
Minimum Fit Function Chi-Square = 12771.72 (P = 0.0)
Normal Theory Weighted Least Squares Chi-Square = 15328.49 (P = 0.0)
Estimated Non-centrality Parameter (NCP) = 14008.49
90 Percent Confidence Interval for NCP = (13613.19 ; 14410.62)

Minimum Fit Function Value = 1.28
Population Discrepancy Function Value (F0) = 1.40
90 Percent Confidence Interval for F0 = (1.36 ; 1.44)
Root Mean Square Error of Approximation (RMSEA) = 0.033
90 Percent Confidence Interval for RMSEA = (0.032 ; 0.033)
P-Value for Test of Close Fit (RMSEA < 0.05) = 1.00

Expected Cross-Validation Index (ECVI) = 1.56
90 Percent Confidence Interval for ECVI = (1.52 ; 1.60)
ECVI for Saturated Model = 0.29
ECVI for Independence Model = 29.74

Chi-Square for Independence Model with 1378 Degrees of Freedom = 297304.60
Independence AIC = 297410.60
Model AIC = 15550.49
Saturated AIC = 2862.00
Independence CAIC = 297845.75
Model CAIC = 16461.84
Saturated CAIC = 14611.00

Normed Fit Index (NFI) = 0.96
Non-Normed Fit Index (NNFI) = 0.96
Parsimony Normed Fit Index (PNFI) = 0.92
Comparative Fit Index (CFI) = 0.96
Incremental Fit Index (IFI) = 0.96
Relative Fit Index (RFI) = 0.96

Critical N (CN) = 1130.31

Root Mean Square Residual (RMR) = 0.024
Standardized RMR = 0.029
Goodness of Fit Index (GFI) = 0.95
Adjusted Goodness of Fit Index (AGFI) = 0.94
Parsimony Goodness of Fit Index (PGFI) = 0.87

Table G-6a – Results of the Full information regression model of the Reading/Science dyad

Structural Equations

```
   R1 = 1.15*R2 - 0.15*SC2, Errorvar.= -0.049 , R² = 1.05
       (0.040)   (0.030)               (0.0097)
       28.83     -5.13                 -5.03

W_A_R_N_I_N_G : Error variance is negative.

  SC1 = - 0.16*R2 + 1.16*SC2, Errorvar.= -0.056 , R² = 1.06
         (0.030)   (0.070)               (0.011)
         -5.28     16.53                 -5.06

W_A_R_N_I_N_G : Error variance is negative.


    Correlation Matrix of Independent Variables

        R2    SC2
     --------  --------
   R2    1.00
  SC2    0.88    1.00
        (0.01)
       157.09

    Covariance Matrix of Latent Variables

        R1    SC1     R2    SC2
     --------  --------  --------  --------
   R1    1.00
  SC1    0.84    1.00
   R2    1.02    0.86    1.00
  SC2    0.86    1.02    0.88    1.00

W_A_R_N_I_N_G: Matrix above is not positive definite
```

Table G-6b – Results of the Full information regression model of the Reading/Science dyad

Goodness of Fit Statistics

Degrees of Freedom = 2769
Minimum Fit Function Chi-Square = 14263.82 (P = 0.0)
Normal Theory Weighted Least Squares Chi-Square = 18676.63 (P = 0.0)
Estimated Non-centrality Parameter (NCP) = 15907.63
90 Percent Confidence Interval for NCP = (15478.07 ; 16343.93)

Minimum Fit Function Value = 1.43
Population Discrepancy Function Value (F0) = 1.59
90 Percent Confidence Interval for F0 = (1.55 ; 1.63)
Root Mean Square Error of Approximation (RMSEA) = 0.024
90 Percent Confidence Interval for RMSEA = (0.024 ; 0.024)
P-Value for Test of Close Fit (RMSEA < 0.05) = 1.00

Expected Cross-Validation Index (ECVI) = 1.90
90 Percent Confidence Interval for ECVI = (1.86 ; 1.94)
ECVI for Saturated Model = 0.59
ECVI for Independence Model = 52.41

Chi-Square for Independence Model with 2850 Degrees of Freedom = 523916.87
Independence AIC = 524068.87
Model AIC = 18990.63
Saturated AIC = 5852.00
Independence CAIC = 524692.86
Model CAIC = 20279.65
Saturated CAIC = 29875.46

Normed Fit Index (NFI) = 0.97
Non-Normed Fit Index (NNFI) = 0.98
Parsimony Normed Fit Index (PNFI) = 0.95
Comparative Fit Index (CFI) = 0.98
Incremental Fit Index (IFI) = 0.98
Relative Fit Index (RFI) = 0.97

Critical N (CN) = 2065.50

Root Mean Square Residual (RMR) = 0.0096
Standardized RMR = 0.023
Goodness of Fit Index (GFI) = 0.95
Adjusted Goodness of Fit Index (AGFI) = 0.95
Parsimony Goodness of Fit Index (PGFI) = 0.90

Table G-7a – Results of the Full information regression model of the Reading/Social Studies dyad

Structural Equations

```
    R1 = 1.09*R2 - 0.079*SSt2, Errorvar.= -0.042 , R² = 1.04
        (0.043)   (0.035)                (0.0092)
        25.18     -2.23                  -4.60

 W_A_R_N_I_N_G : Error variance is negative.

   SSt1 = - 0.059*R2 + 1.05*SSt2, Errorvar.= 0.0077 , R² = 0.99
           (0.028)    (0.040)               (0.0070)
           -2.09      26.15                  1.11


     Correlation Matrix of Independent Variables

          R2    SSt2
      --------  --------
    R2    1.00
   SSt2   0.90    1.00
         (0.01)
        179.22

     Covariance Matrix of Latent Variables

          R1    SSt1     R2    SSt2
      --------  --------  --------  --------
    R1    1.00
   SSt1   0.89    1.00
    R2    1.02    0.89    1.00
   SSt2   0.90    1.00    0.90    1.00
```

Table G-7b – Results of the Full information regression model of the Reading/Social Studies

Goodness of Fit Statistics

Degrees of Freedom = 2769
Minimum Fit Function Chi-Square = 13286.97 (P = 0.0)
Normal Theory Weighted Least Squares Chi-Square = 16882.34 (P = 0.0)
Estimated Non-centrality Parameter (NCP) = 14113.34
90 Percent Confidence Interval for NCP = (13706.93 ; 14526.59)

Minimum Fit Function Value = 1.33
Population Discrepancy Function Value (F0) = 1.41
90 Percent Confidence Interval for F0 = (1.37 ; 1.45)
Root Mean Square Error of Approximation (RMSEA) = 0.023
90 Percent Confidence Interval for RMSEA = (0.022 ; 0.023)
P-Value for Test of Close Fit (RMSEA < 0.05) = 1.00

Expected Cross-Validation Index (ECVI) = 1.72
90 Percent Confidence Interval for ECVI = (1.68 ; 1.76)
ECVI for Saturated Model = 0.59
ECVI for Independence Model = 67.13

Chi-Square for Independence Model with 2850 Degrees of Freedom = 671060.49
Independence AIC = 671212.49
Model AIC = 17196.34
Saturated AIC = 5852.00
Independence CAIC = 671836.48
Model CAIC = 18485.36
Saturated CAIC = 29875.46

Normed Fit Index (NFI) = 0.98
Non-Normed Fit Index (NNFI) = 0.98
Parsimony Normed Fit Index (PNFI) = 0.95
Comparative Fit Index (CFI) = 0.98
Incremental Fit Index (IFI) = 0.98
Relative Fit Index (RFI) = 0.98

Critical N (CN) = 2217.28

Root Mean Square Residual (RMR) = 0.0086
Standardized RMR = 0.021
Goodness of Fit Index (GFI) = 0.96
Adjusted Goodness of Fit Index (AGFI) = 0.96
Parsimony Goodness of Fit Index (PGFI) = 0.91

Table G-8a – Results of the Full information regression model of the Writing/Science

Structural Equations

```
   W1 = 1.58*W2 - 0.40*SC2, Errorvar.= -0.66 , R² = 1.66
       (0.036)   (0.033)               (0.024)
       44.22    -12.28                -27.27

W_A_R_N_I_N_G : Error variance is negative.

  SC1 = - 0.10*W2 + 1.10*SC2, Errorvar.= -0.049 , R² = 1.05
         (0.015)   (0.063)               (0.010)
         -6.71     17.46                 -4.90

W_A_R_N_I_N_G : Error variance is negative.
```

Correlation Matrix of Independent Variables

	W2	SC2
W2	1.00	
SC2	0.78	1.00
	(0.01)	
	102.27	

Covariance Matrix of Latent Variables

	W1	SC1	W2	SC2
W1	1.00			
SC1	0.79	1.00		
W2	1.26	0.76	1.00	
SC2	0.84	1.02	0.78	1.00

Table G-8b – Results of the Full information regression model of the Writing/Science

Goodness of Fit Statistics

Degrees of Freedom = 1320
Minimum Fit Function Chi-Square = 10586.81 (P = 0.0)
Normal Theory Weighted Least Squares Chi-Square = 12224.33 (P = 0.0)
Estimated Non-centrality Parameter (NCP) = 10904.33
90 Percent Confidence Interval for NCP = (10553.83 ; 11261.81)

Minimum Fit Function Value = 1.06
Population Discrepancy Function Value (F0) = 1.09
90 Percent Confidence Interval for F0 = (1.06 ; 1.13)
Root Mean Square Error of Approximation (RMSEA) = 0.029
90 Percent Confidence Interval for RMSEA = (0.028 ; 0.029)
P-Value for Test of Close Fit (RMSEA < 0.05) = 1.00

Expected Cross-Validation Index (ECVI) = 1.24
90 Percent Confidence Interval for ECVI = (1.21 ; 1.28)
ECVI for Saturated Model = 0.29
ECVI for Independence Model = 27.54

Chi-Square for Independence Model with 1378 Degrees of Freedom = 275218.25
Independence AIC = 275324.25
Model AIC = 12446.33
Saturated AIC = 2862.00
Independence CAIC = 275759.40
Model CAIC = 13357.68
Saturated CAIC = 14611.00

Normed Fit Index (NFI) = 0.96
Non-Normed Fit Index (NNFI) = 0.96
Parsimony Normed Fit Index (PNFI) = 0.92
Comparative Fit Index (CFI) = 0.97
Incremental Fit Index (IFI) = 0.97
Relative Fit Index (RFI) = 0.96

Critical N (CN) = 1363.37

Root Mean Square Residual (RMR) = 0.023
Standardized RMR = 0.029
Goodness of Fit Index (GFI) = 0.96
Adjusted Goodness of Fit Index (AGFI) = 0.95
Parsimony Goodness of Fit Index (PGFI) = 0.88

Table G-9a – Results of the Full information regression model of the Writing/Social Studies

Structural Equations

```
   W1 = 1.59*W2 - 0.42*SSt2, Errorvar.= -0.67 , R² = 1.67
       (0.036)   (0.034)                (0.025)
       43.61    -12.48                 -26.74

W_A_R_N_I_N_G : Error variance is negative.

  SSt1 = - 0.026*W2 + 1.02*SSt2, Errorvar.= 0.0092 , R² = 0.99
          (0.012)    (0.031)               (0.0066)
          -2.15      33.11                  1.40


    Correlation Matrix of Independent Variables

         W2    SSt2
      --------  --------
   W2    1.00
  SSt2   0.78    1.00
        (0.01)
       101.01

    Covariance Matrix of Latent Variables

         W1    SSt1     W2    SSt2
      --------  --------  --------  --------
   W1    1.00
  SSt1   0.80    1.00
   W2    1.26    0.77    1.00
  SSt2   0.82    1.00    0.78    1.00
```

Table G-9b – Results of the Full information regression model of the Writing/Social Studies

Goodness of Fit Statistics

Degrees of Freedom = 1320
Minimum Fit Function Chi-Square = 9772.66 (P = 0.0)
Normal Theory Weighted Least Squares Chi-Square = 10915.95 (P = 0.0)
Estimated Non-centrality Parameter (NCP) = 9595.95
90 Percent Confidence Interval for NCP = (9266.12 ; 9932.83)

Minimum Fit Function Value = 0.98
Population Discrepancy Function Value (F0) = 0.96
90 Percent Confidence Interval for F0 = (0.93 ; 0.99)
Root Mean Square Error of Approximation (RMSEA) = 0.027
90 Percent Confidence Interval for RMSEA = (0.026 ; 0.027)
P-Value for Test of Close Fit (RMSEA < 0.05) = 1.00

Expected Cross-Validation Index (ECVI) = 1.11
90 Percent Confidence Interval for ECVI = (1.08 ; 1.15)
ECVI for Saturated Model = 0.29
ECVI for Independence Model = 36.69

Chi-Square for Independence Model with 1378 Degrees of Freedom = 366791.34
Independence AIC = 366897.34
Model AIC = 11137.95
Saturated AIC = 2862.00
Independence CAIC = 367332.48
Model CAIC = 12049.30
Saturated CAIC = 14611.00

Normed Fit Index (NFI) = 0.97
Non-Normed Fit Index (NNFI) = 0.98
Parsimony Normed Fit Index (PNFI) = 0.93
Comparative Fit Index (CFI) = 0.98
Incremental Fit Index (IFI) = 0.98
Relative Fit Index (RFI) = 0.97

Critical N (CN) = 1476.87

Root Mean Square Residual (RMR) = 0.022
Standardized RMR = 0.028
Goodness of Fit Index (GFI) = 0.96
Adjusted Goodness of Fit Index (AGFI) = 0.96
Parsimony Goodness of Fit Index (PGFI) = 0.89

Table G-10a – Results of the Full information regression model of the Science/Social Studies

Structural Equations

```
SC1 = 1.26*SC2 - 0.25*SSt2, Errorvar.= -0.057 , R² = 1.06
     (0.092)    (0.061)                (0.012)
     13.75      -4.13                  -4.69

W_A_R_N_I_N_G : Error variance is negative.

SSt1 = 0.020*SC2 + 0.98*SSt2, Errorvar.= 0.0052 , R² = 0.99
      (0.037)   (0.046)                 (0.0064)
       0.56     21.41                    0.81


  Correlation Matrix of Independent Variables

       SC2    SSt2
   --------  --------
SC2    1.00
SSt2   0.94    1.00
      (0.00)
     216.34

  Covariance Matrix of Latent Variables

       SC1    SSt1    SC2    SSt2
   --------  --------  --------  --------
SC1    1.00
SSt1   0.93    1.00
SC2    1.02    0.94    1.00
SSt2   0.93    1.00    0.94    1.00
```

Table G-10b – Results of the Full information regression model of the Science/Social Studies

Goodness of Fit Statistics

Degrees of Freedom = 2769
Minimum Fit Function Chi-Square = 10552.28 (P = 0.0)
Normal Theory Weighted Least Squares Chi-Square = 13041.90 (P = 0.0)
Estimated Non-centrality Parameter (NCP) = 10272.90
90 Percent Confidence Interval for NCP = (9921.03 ; 10631.84)

Minimum Fit Function Value = 1.06
Population Discrepancy Function Value (F0) = 1.03
90 Percent Confidence Interval for F0 = (0.99 ; 1.06)
Root Mean Square Error of Approximation (RMSEA) = 0.019
90 Percent Confidence Interval for RMSEA = (0.019 ; 0.020)
P-Value for Test of Close Fit (RMSEA < 0.05) = 1.00

Expected Cross-Validation Index (ECVI) = 1.34
90 Percent Confidence Interval for ECVI = (1.30 ; 1.37)
ECVI for Saturated Model = 0.59
ECVI for Independence Model = 68.44

Chi-Square for Independence Model with 2850 Degrees of Freedom = 684221.22
Independence AIC = 684373.22
Model AIC = 13355.90
Saturated AIC = 5852.00
Independence CAIC = 684997.20
Model CAIC = 14644.92
Saturated CAIC = 29875.46

Normed Fit Index (NFI) = 0.98
Non-Normed Fit Index (NNFI) = 0.99
Parsimony Normed Fit Index (PNFI) = 0.96
Comparative Fit Index (CFI) = 0.99
Incremental Fit Index (IFI) = 0.99
Relative Fit Index (RFI) = 0.98

Critical N (CN) = 2791.64

Root Mean Square Residual (RMR) = 0.0052
Standardized RMR = 0.018
Goodness of Fit Index (GFI) = 0.97
Adjusted Goodness of Fit Index (AGFI) = 0.96
Parsimony Goodness of Fit Index (PGFI) = 0.91

Figure G2 – Figure of the regression model with Reading and Science the dependent variables and Mathematics and Social Studies as the independent variables (Equations 3a and 3b).

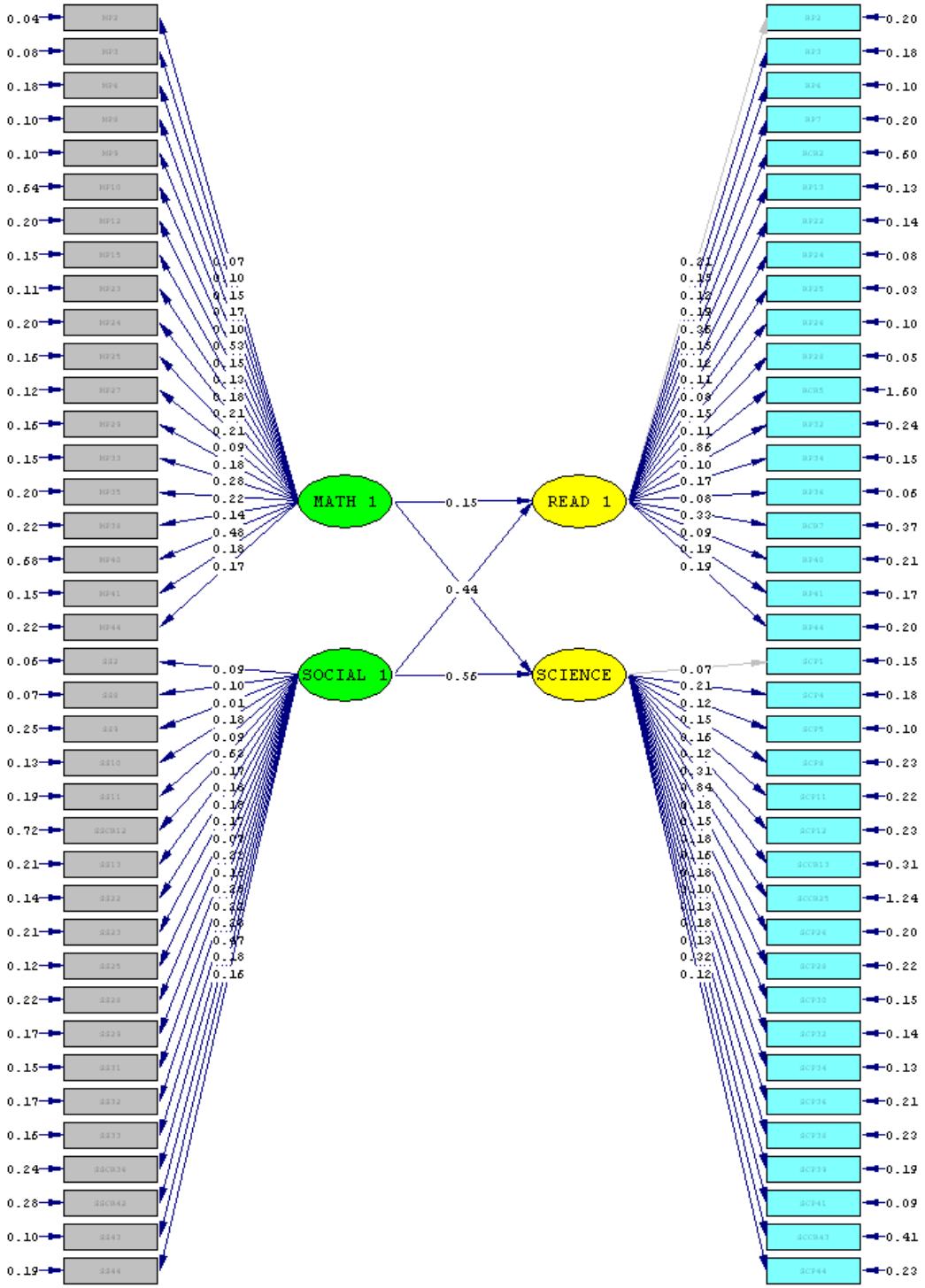


Table G-11a – Results of the regression model with Reading and Science the dependent variables and Mathematics and Social Studies as the independent variables.

```
        Structural Equations


  READ 1 = 0.15*MATH 1 + 0.77*SOCIAL 1, Errorvar.= 0.20  , R² = 0.80
          (0.021)       (0.027)                   (0.012)
           7.07          28.46                     16.22

SCIENCE  = 0.44*MATH 1 + 0.56*SOCIAL 1, Errorvar.= 0.078 , R² = 0.92
          (0.031)       (0.036)                   (0.011)
           14.50         15.60                     7.06


        Correlation Matrix of Independent Variables

             MATH 1   SOCIAL 1
           --------   --------
  MATH 1       1.00
SOCIAL 1       0.82       1.00
             (0.01)
             139.17

        Covariance Matrix of Latent Variables

             READ 1   SCIENCE      MATH 1   SOCIAL 1
           --------   --------   --------   --------
  READ 1       1.00
SCIENCE        0.85       1.00
  MATH 1       0.78       0.91       1.00
SOCIAL 1       0.89       0.93       0.82       1.00
```

Table G-11b – Results of the regression model with Reading and Science the dependent variables and Mathematics and Social Studies as the independent variables.

Goodness of Fit Statistics

Degrees of Freedom = 2769
Minimum Fit Function Chi-Square = 10961.54 (P = 0.0)
Normal Theory Weighted Least Squares Chi-Square = 13597.01 (P = 0.0)
Estimated Non-centrality Parameter (NCP) = 10828.01
90 Percent Confidence Interval for NCP = (10467.76 ; 11195.30)

Minimum Fit Function Value = 1.10
Population Discrepancy Function Value (F0) = 1.08
90 Percent Confidence Interval for F0 = (1.05 ; 1.12)
Root Mean Square Error of Approximation (RMSEA) = 0.020
90 Percent Confidence Interval for RMSEA = (0.019 ; 0.020)
P-Value for Test of Close Fit (RMSEA < 0.05) = 1.00

Expected Cross-Validation Index (ECVI) = 1.39
90 Percent Confidence Interval for ECVI = (1.36 ; 1.43)
ECVI for Saturated Model = 0.59
ECVI for Independence Model = 59.19

Chi-Square for Independence Model with 2850 Degrees of Freedom = 591652.91
Independence AIC = 591804.91
Model AIC = 13911.01
Saturated AIC = 5852.00
Independence CAIC = 592428.89
Model CAIC = 15200.03
Saturated CAIC = 29875.46

Normed Fit Index (NFI) = 0.98
Non-Normed Fit Index (NNFI) = 0.99
Parsimony Normed Fit Index (PNFI) = 0.95
Comparative Fit Index (CFI) = 0.99
Incremental Fit Index (IFI) = 0.99
Relative Fit Index (RFI) = 0.98

Critical N (CN) = 2687.45

Root Mean Square Residual (RMR) = 0.0067
Standardized RMR = 0.019
Goodness of Fit Index (GFI) = 0.97
Adjusted Goodness of Fit Index (AGFI) = 0.96
Parsimony Goodness of Fit Index (PGFI) = 0.91

Figure G3 – Figure of the regression model with Science and Social Studies as the dependent variables and Mathematics and Reading as the independent variables (Equations 4a and 4b).

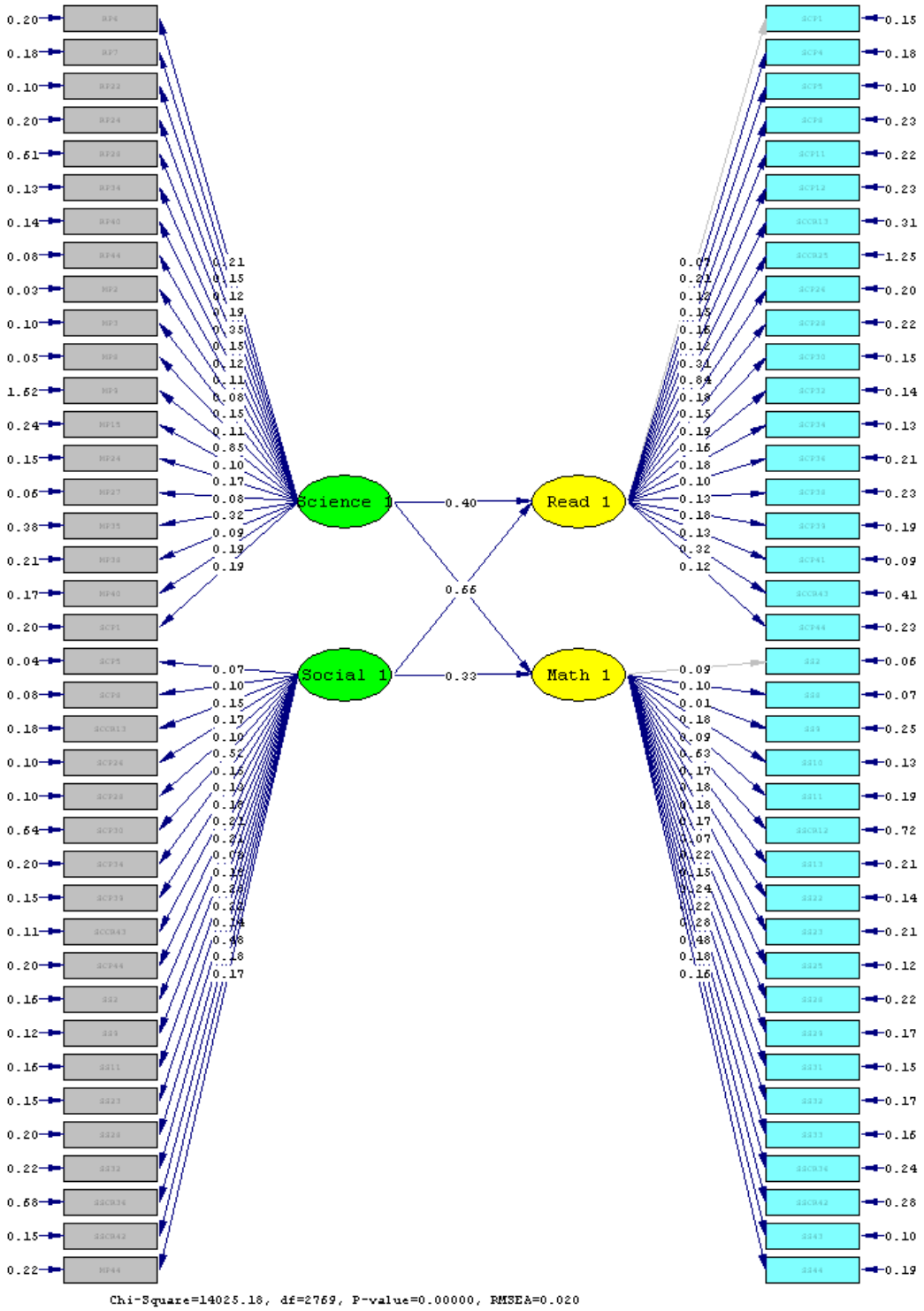


Table G-11c – Results of the regression model with Science and Social Studies as the dependent variables and Reading and Mathematics as the independent variables.

```
Structural Equations


 Science  = 0.40*Read 1 + 0.61*Math 1, Errorvar.= 0.086 , R² = 0.91
           (0.028)       (0.038)                 (0.012)
            14.50         16.16                   7.24

 Social 1 = 0.66*Read 1 + 0.33*Math 1, Errorvar.= 0.12   , R² = 0.88
           (0.024)       (0.019)                 (0.0092)
            27.27         17.84                   12.99


         Correlation Matrix of Independent Variables

              Read 1     Math 1
            --------   --------
   Read 1       1.00
   Math 1       0.78       1.00
              (0.01)
              111.90

         Covariance Matrix of Latent Variables

            Science    Social 1     Read 1     Math 1
            --------   --------   --------   --------
 Science        1.00
 Social 1       0.88       1.00
   Read 1       0.88       0.92       1.00
   Math 1       0.92       0.84       0.78       1.00
```

Table G-11d – Results of the regression model with Science and Social Studies as the dependent variables and Reading and Mathematics as the independent variables.

Goodness of Fit Statistics

Degrees of Freedom = 2769
Minimum Fit Function Chi-Square = 11277.22 (P = 0.0)
Normal Theory Weighted Least Squares Chi-Square = 14025.18 (P = 0.0)
Estimated Non-centrality Parameter (NCP) = 11256.18
90 Percent Confidence Interval for NCP = (10889.59 ; 11629.79)

Minimum Fit Function Value = 1.13
Population Discrepancy Function Value (F0) = 1.13
90 Percent Confidence Interval for F0 = (1.09 ; 1.16)
Root Mean Square Error of Approximation (RMSEA) = 0.020
90 Percent Confidence Interval for RMSEA = (0.020 ; 0.020)
P-Value for Test of Close Fit (RMSEA < 0.05) = 1.00

Expected Cross-Validation Index (ECVI) = 1.43
90 Percent Confidence Interval for ECVI = (1.40 ; 1.47)
ECVI for Saturated Model = 0.59
ECVI for Independence Model = 59.19

Chi-Square for Independence Model with 2850 Degrees of Freedom = 591652.91
Independence AIC = 591804.91
Model AIC = 14339.18
Saturated AIC = 5852.00
Independence CAIC = 592428.89
Model CAIC = 15628.20
Saturated CAIC = 29875.46

Normed Fit Index (NFI) = 0.98
Non-Normed Fit Index (NNFI) = 0.99
Parsimony Normed Fit Index (PNFI) = 0.95
Comparative Fit Index (CFI) = 0.99
Incremental Fit Index (IFI) = 0.99
Relative Fit Index (RFI) = 0.98

Critical N (CN) = 2612.25

Root Mean Square Residual (RMR) = 0.0068
Standardized RMR = 0.019
Goodness of Fit Index (GFI) = 0.96
Adjusted Goodness of Fit Index (AGFI) = 0.96
Parsimony Goodness of Fit Index (PGFI) = 0.91

Appendix H: Comparison of traditionally computed reliabilities (α) to G coefficients

Table H-1 – Summary of Analysis of Variance Results for the Spring 2005 Administration of Reading, Mathematics, Science, and Social Studies, G study.

	Reading	Mathematics	Science	Social Studies
Variance due to Students (S)	.05067	.05645	.05149	.06409
Variance due to Items (I)	.00590	.00164	.00358	.00256
Interaction Variance (SI)	.00774	.00629	.00745	.00813
Number of Persons	45,006	45,007	45,021	45,000
Number of Items	38	38	38	38

Note: Tables H-1 through H-3 are from Bunch, 2006(2).

Table H-2 – Summary of Generalizability/Dependability Analyses for Spring 2005 Reading, Mathematics, Science, and Social Studies (Cut scores are shown in parentheses)

Index	Reading	Mathematics	Science	Social Studies
Generalizability	.87	.90	.87	.89
Phi	.79	.88	.82	.86
D for Basic	.96 (12.0)	.96 (12.5)	.94 (14.5)	.95 (15.0)
D for Proficient	.94 (18.0)	.94 (18.0)	.85 (23.5)	.92 (21.5)
D for Accelerated	.77 (32.0)	.87 (27.5)	.84 (32.0)	.86 (33.0)
D for Advanced	.84 (40.0)	.92 (34.5)	.90 (37.5)	.91 (39.0)
Mean Raw Score	33.74	27.69	28.43	30.45

Table H-3 – Summary of Generalizability and Dependability Analyses for Spring 2005 Writing[13] (Cut scores are shown in parentheses)

Source	Value	Index	Value
Variance due to Students (S)	.13837	Generalizability	.89
Variance due to Items (I)	.06251	Phi	.64
Interaction Variance (SI)	.01665	D for Basic	.88
Number of Persons	45,007	D for Proficient	.70
Number of Items	19	D for Accelerated	.53
		D for Advanced	.80
		Mean Raw Score	31.86

[13] According to Bunch, 2006 (2), "…because so much of the score on that test derives from two essay prompts. Each prompt is scored for content and for conventions by two independent readers. The four scores for each prompt are summed for a final score that ranges from 0 to 18 for each prompt and from 0 to 36 for the two prompts combined, as opposed to 12 points for all other parts of the test (10 multiple-choice editing items and one 2-point short-answer question). Because we were not able to cross every reader with every student, we were not able to count reader as a separate source of variance. We did, however, treat the two readers for each prompt as separate items. Thus, for example, the first essay prompt yielded four separate scores, treated as responses to four separate items (Reader 1 X Content, Reader 2 X Content, Reader 1 X Conventions, and Reader 2 X Conventions). Thus, the generalizability analysis of the Writing test considered 19 separate items, rather than the 13 actual items."

Table H-4 – Summary of Generalizability Analyses for Reading

Group	N	Mean	VarS	VarI	VarSI	G	α[14]	Phi	D2	D3	D4	D5
Female	22007	34.77	0.04447	0.00695	0.00735	0.86	.884277	0.76	0.97	0.94	0.75	0.80
Male	22954	32.74	0.05514	0.00501	0.00802	0.87	.900034	0.81	0.96	0.94	0.79	0.87
American Indian	78	31.54	0.06417	0.00491	0.00827	0.89	.906663	0.83	0.96	0.93	0.82	0.89
Asian Pacific Islander	544	35.22	0.05368	0.00692	0.00717	0.88	.892179	0.79	0.97	0.95	0.79	0.82
Black/African American	6359	27.36	0.06118	0.00346	0.00899	0.87	.894147	0.83	0.95	0.90	0.84	0.93
Hispanic	890	29.31	0.06801	0.00378	0.00862	0.89	.898683	0.85	0.96	0.92	0.85	0.92
White	36419	34.95	0.04200	0.00648	0.00740	0.85	.879711	0.75	0.97	0.94	0.75	0.79
Multiracial	590	33.20	0.05061	0.00555	0.00784	0.87	.890257	0.79	0.96	0.94	0.77	0.85
Other	126	32.03	0.08926	0.00524	0.00793	0.92	.911883	0.87	0.96	0.94	0.86	0.91
Public	41239	33.28	0.05200	0.00571	0.00787	0.87	.894127	0.79	0.96	0.94	0.78	0.85
Nonpublic	3767	38.67	0.01791	0.00836	0.00603	0.75	.792816	0.55	0.97	0.95	0.74	0.43

Table H-5 – Summary of Generalizability Analyses for Mathematics

Group	N	Mean	VarS	VarI	VarSI	G	α*	Phi	D2	D3	D4	D5
Female	22079	27.21	0.05316	0.00187	0.00633	0.89	.900182	0.87	0.96	0.93	0.86	0.91
Male	22873	28.18	0.05926	0.00147	0.00621	0.91	.912237	0.89	0.97	0.94	0.88	0.92
American Indian	86	25.48	0.05561	0.00122	0.00669	0.89	.907005	0.88	0.96	0.92	0.88	0.93
Asian Pacific Islander	545	32.48	0.04983	0.00227	0.00560	0.90	.902129	0.86	0.98	0.96	0.89	0.86
Black/African American	6323	19.90	0.04296	0.00089	0.00643	0.87	.880001	0.85	0.92	0.86	0.92	0.96
Hispanic	927	22.73	0.05421	0.00102	0.00650	0.89	.893330	0.88	0.94	0.90	0.90	0.95
White	36419	29.13	0.04942	0.00188	0.00617	0.89	.897175	0.86	0.97	0.94	0.86	0.89
Multiracial	581	26.52	0.06165	0.00141	0.00630	0.91	.899104	0.89	0.96	0.93	0.89	0.93
Other	126	25.27	0.08359	0.00108	0.00625	0.93	.920369	0.92	0.96	0.94	0.92	0.95
Public	41461	27.27	0.05686	0.00160	0.00633	0.90	.906928	0.88	0.96	0.94	0.87	0.92
Nonpublic	3546	32.70	0.03290	0.00220	0.00573	0.85	.862183	0.81	0.98	0.96	0.86	0.81

* The reliability, α, is per footnote for the table H-4.

[14] Reliabilities are from Tables 12a through 12e and are shown for comparison to the G coefficients produced by the G studies. The G study was performed on a sample of about 45,000 examinees from the Spring 2005 administration, while the reported α is for a population of more than 140,000 examinees for the Spring 2006 Administration.

Table H-6 – Summary of Generalizability Analyses for Science

Group	N	Mean	VarS	VarI	VarSI	G	α*	Phi	D2	D3	D4	D5
Female	22075	27.97	0.04885	0.00412	0.00760	0.87	.859913	0.81	0.94	0.83	0.83	0.90
Male	22891	28.92	0.05368	0.00312	0.00725	0.88	.884617	0.84	0.95	0.87	0.85	0.91
American Indian	86	26.14	0.05768	0.00259	0.00775	0.88	.864770	0.85	0.93	0.85	0.88	0.93
Asian Pacific Islander	545	31.09	0.05711	0.00477	0.00750	0.88	.880738	0.82	0.95	0.88	0.81	0.87
Black/African American	6323	19.94	0.04256	0.00151	0.00747	0.85	.822907	0.83	0.87	0.85	0.94	0.97
Hispanic	927	22.88	0.05306	0.00207	0.00772	0.87	.854540	0.84	0.91	0.84	0.92	0.95
White	36419	30.06	0.04184	0.00415	0.00726	0.85	.860955	0.79	0.95	0.86	0.78	0.87
Multiracial	58	25.50	0.04955	0.00309	0.00802	0.86	.869357	0.82	0.92	0.81	0.87	0.93
Other	67	18.22	0.06252	0.00063	0.00690	0.90	.887821	0.89	0.90	0.91	0.96	0.98
Public	41869	28.10	0.05245	0.00348	0.00748	0.88	.874545	0.83	0.94	0.85	0.84	0.91
Nonpublic	3152	32.94	0.02374	0.00498	0.00690	0.77	.821762	0.67	0.96	0.87	0.62	0.74

* The reliability, α, is per footnote for the table H-4.

Table H-7 – Summary of Generalizability Analyses for Social Studies

Group	N	Mean	VarS	VarI	VarSI	G	α*	Phi	D2	D3	D4	D5
Female	22132	29.94	0.05992	0.00244	0.00821	0.88	.884276	0.85	0.95	0.91	0.86	0.91
Male	22818	30.98	0.06766	0.00275	0.00799	0.89	.904736	0.86	0.96	0.92	0.86	0.91
American Indian	73	29.32	0.05841	0.00230	0.00863	0.87	.884972	0.84	0.95	0.90	0.86	0.92
Asian Pacific Islander	545	33.91	0.06487	0.00421	0.00736	0.90	.898048	0.85	0.96	0.94	0.84	0.87
Black/African American	6345	22.17	0.05658	0.00076	0.00798	0.88	.879989	0.87	0.91	0.87	0.94	0.97
Hispanic	910	24.24	0.07021	0.00101	0.00824	0.89	.891999	0.88	0.93	0.89	0.93	0.96
White	36419	32.03	0.05454	0.00311	0.00796	0.87	.886874	0.83	0.96	0.92	0.82	0.88
Multiracial	601	29.55	0.06240	0.00218	0.00820	0.88	.894910	0.86	0.95	0.91	0.87	0.92
Other	107	26.09	0.10037	0.00161	0.00826	0.92	.909049	0.91	0.95	0.92	0.93	0.96
Public	41773	30.07	0.06550	0.00249	0.00820	0.89	.896036	0.86	0.95	0.91	0.87	0.92
Nonpublic	3227	35.40	0.02274	0.00361	0.00707	0.80	.827808	0.72	0.97	0.94	0.72	0.76

* The reliability, α, is per footnote for the table H-4.

Table H-8 - Summary of Generalizability Analyses for Writing

Group	N	Mean	VarS	VarI	VarSI	G[15]	α*	Phi	D2	D3	D4	D5
Female	22158	31.68	0.10130	0.07887	0.01462	0.87	.810283	0.52	0.85	0.58	0.28	0.74
Male	22799	28.93	0.14550	0.06284	0.01735	0.89	.840334	0.64	0.84	0.59	0.66	0.86
American Indian	86	28.66	0.11272	0.06009	0.01859	0.86	.840156	0.59	0.82	0.50	0.62	0.86
Asian Pacific Islander	545	31.74	0.13970	0.08030	0.01679	0.89	.829518	0.59	0.85	0.63	0.43	0.75
Black/African American	6323	25.46	0.14938	0.04956	0.01875	0.89	.814615	0.69	0.79	0.59	0.82	0.92
Hispanic	927	26.31	0.17355	0.05373	0.01975	0.90	.862971	0.70	0.81	0.62	0.79	0.91
White	36419	31.21	0.10952	0.07489	0.01527	0.88	.833219	0.55	0.85	0.58	0.38	0.77
Multiracial	581	29.90	0.12971	0.06868	0.01645	0.89	.829369	0.60	0.84	0.57	0.56	0.83
Other	126	28.12	0.21480	0.05933	0.01876	0.92	.855658	0.73	0.85	0.69	0.76	0.89
Public	41952	29.95	0.13038	0.06879	0.01646	0.89	.833254	0.60	0.84	0.58	0.56	0.82
Nonpublic	3055	34.87	0.04982	0.09532	0.01142	0.81	.698433	0.32	0.87	0.65	---	0.36

* The reliability, α, is per footnote for the table H-4.

[15] The G study for Writing was based on handling the scores from different readers as independent items while the computation of α was based on the sum of the scores for both readers of the same item. The difference in the methods may explain why the computational result for G differs from the computation of α.

Appendix I: Test Success Data

Table I-1 – Counts of examinees of the graduating class cohort for Spring 2007.
OGT Writing Test

	SP05 Grade 10		SU05 Grade 11		FA05 Grade 11		SP06 Grade 11		SU06 Grade 12		FA06 Grade 12		SP07 Grade 12	
	Attempt	Passed	Attempt	Passed	Attempt	Passed	Attempt	Passed	Attempt	Passed	Attempt	Passed	Attempt	Passed
Ethnicity														
*	2382	1778												
Blank	212	144												
Am Ind	247	181	3	2	52	25	25	13	1		19	9	5	2
Asian	1870	1654	6	3	268	167	134	84	3	1	137	84	62	34
Black	18946	12172	202	88	5174	2544	3026	1457	44	19	1667	612	931	400
Hisp	2594	1801	26	12	633	289	370	154	9	3	267	111	141	65
White	118902	103564	501	314	13979	7246	6850	3218	75	50	4229	1861	1798	678
Multi	2083	1659	19	9	411	211	213	102	7	6	162	76	62	29
Other	511	433	12	6	148	75	72	36	1		69	40	56	29
Gender														
Blank	917	703	4	2	101	47	27	14	2	1	26	7	40	20
Female	72348	64703	274	172	7101	4189	3521	1938	46	24	2307	1162	978	452
Male	74482	57980	491	260	13463	6321	7142	3112	92	54	4441	1710	2037	765

Table I-2 – Counts of examinees of the graduating class cohort for Spring 2007.
OGT Reading Test

	SP05 Grade 10		SU05 Grade 11		FA05 Grade 11		SP06 Grade 11		SU06 Grade 12		FA06 Grade 12		SP07 Grade 12	
	Attempt	Passed	Attempt	Passed	Attempt	Passed	Attempt	Passed	Attempt	Passed	Attempt	Passed	Attempt	Passed
Ethnicity														
*	2423	2056												
Blank	220	180												
Am Ind	253	218	6	2	41	13	17	7			16	9	4	1
Asian	1873	1757	2	1	228	100	144	87	3	1	138	81	67	32
Black	19278	15280	218	92	3666	1063	2492	911	35	11	1549	517	849	241
Hisp	2626	2135	22	10	456	135	317	130	9	5	227	104	118	44
White	119218	111886	307	156	8781	3345	4777	1840	59	36	3533	1499	1438	426
Multi	2109	1870	21	11	301	124	170	74	6	5	144	75	50	24
Other	514	448	13	6	111	45	53	23	7	4	63	30	58	25
Gender														
Blank	941	788	8	4	67	17	20	8	2	0	23	6	38	11
Female	72646	68305	262	133	5525	2127	3092	1316	53	29	2206	985	1015	357
Male	74927	66737	319	141	7992	2681	4858	1748	64	33	3441	1324	1531	425

Table I-3 – Counts of examinees of the graduating class cohort for Spring 2007.
OGT Mathematics Test

	SP05 Grade 10		SU05 Grade 11		FA05 Grade 11		SP06 Grade 11		SU06 Grade 12		FA06 Grade 12		SP07 Grade 12	
	Attempt	Passed	Attempt	Passed	Attempt	Passed	Attempt	Passed	Attempt	Passed	Attempt	Passed	Attempt	Passed
Ethnicity														
*		1648												
Blank	215	132												
Am Ind	253	169	10	4	61	16	35	14	1	1	26	11	9	3
Asian	1876	1704	9	8	245	154	123	82			121	81	50	32
Black	19421	10521	467	166	6842	1991	4873	1599	255	40	2846	503	2217	607
Hisp	2638	1726	32	11	693	256	464	181	22	5	309	98	189	67
White	119149	102483	758	382	15095	6057	8790	3303	197	57	5489	1702	2896	885
Multi	2097	1580	44	14	484	180	314	117	18	7	207	70	104	40
Other	517	410	22	14	175	86	81	30	17	2	77	35	98	34
Gender														
Blank	951	647	10	2	108	34	23	6	6	0	41	9	68	17
Female	72736	59121	782	365	11791	4504	7244	2697	321	76	4378	1085	2904	906
Male	74923	60605	550	232	11696	4202	7413	2623	183	36	4656	1406	2591	745

Table I-4 – Counts of examinees of the graduating class cohort for Spring 2007.
OGT Social Studies

	SP05 Grade 10		SU05 Grade 11		FA05 Grade 11		SP06 Grade 11		SU06 Grade 12		FA06 Grade 12		SP07 Grade 12	
	Attempt	Passed	Attempt	Passed	Attempt	Passed	Attempt	Passed	Attempt	Passed	Attempt	Passed	Attempt	Passed
Ethnicity														
*	2362	1582												
Blank	206	134												
Am Ind	247	175	4	3	60	13	38	14	1		29	19	7	1
Asian	1862	1613	9	3	304	123	204	113	3	1	168	105	73	25
Black	18775	10011	283	82	6850	1428	5308	1624	129	48	3190	1209	1961	417
Hisp	2590	1591	36	9	789	193	621	217	9	3	399	207	182	54
White	118674	99383	746	388	17468	5359	11580	4153	257	133	7198	3129	3208	876
Multi	2077	1553	34	14	477	150	319	99	16	6	228	106	96	28
Other	509	409	15	8	160	55	99	39	15	8	98	51	89	30
Gender														
Blank	921	631	8	1	118	27	39	12	5	2	45	18	64	19
Female	72138	57108	631	320	13116	3540	931	3371	267	123	5559	2338	2926	724
Male	74243	58712	488	186	12874	3754	8819	2876	158	74	5706	2470	2626	688

Table I-5 – Counts of examinees of the graduating class cohort for Spring 2007.
OGT Science Test

	SP05 Grade 10		SU05 Grade 11		FA05 Grade 11		SP06 Grade 11		SU06 Grade 12		FA06 Grade 12		SP07 Grade 12	
	Attempt	Passed	Attempt	Passed	Attempt	Passed	Attempt	Passed	Attempt	Passed	Attempt	Passed	Attempt	Passed
Ethnicity														
*	2399	1344												
Blank	208	107												
Am Ind	251	158	12	1	75	14	50	16	3	1	34	12	14	4
Asian	1875	1527	15	6	383	143	254	113	7		216	95	130	42
Black	19005	6883	576	155	9751	2165	7181	1440	320	64	4923	1115	3806	937
Hisp	2614	1338	54	17	1021	265	736	203	20	5	523	175	325	99
White	118830	94302	1107	511	22184	8203	13146	3944	435	171	8683	2994	4655	1390
Multi	2086	1379	43	10	635	221	404	92	29	13	300	96	163	51
Other	518	380	28	11	238	94	124	37	23	5	118	45	148	39
Gender														
Blank	933	549	19	5	150	42	44	16	9	3	53	11	105	28
Female	74504	54840	1090	450	17862	5663	11608	3044	564	167	7800	2119	5308	1509
Male	72349	52029	726	256	16275	11105	10243	2785	264	89	6944	2402	3828	1025