

# Appendix D: Research

- |  |    |
|--|----|
| 1. Person Fit by Disability Study        | 2  |
| 2. Science Item Cognitive Validity Study | 20 |

Georgia Alternate Assessment 2.0 (GAA 2.0)

# Person Fit by Disability Category

2023 Spring

Data Recognition Corporation Psychometrics Services  
11-6-2023

## **Background**

The GAA 2.0 is designed to measure the degree to which students with significant cognitive disabilities have mastered alternate achievement standards in the core content areas of English language arts (ELA), mathematics, science, and social studies. To be eligible for the GAA 2.0, a student must meet all the participation guidelines. The decision to assess using the GAA 2.0 is made by the IEP team after considering, responding to, and providing a rationale for the eligibility questions. There are 13 primary disability categories with the GAA 2.0 student population. Three categories (SRC 05: Mild Intellectual Disability; SRC 07: Moderate/Severe/Profound Intellectual Disabilities; 08: Autism) have relatively large sample sizes among the categories across all grades.

Measurement invariance assesses the psychometric equivalence of a construct across groups. Measurement non-invariance suggests that a construct has a different structure or meaning to different groups or on different measurement occasions in the same group. Under measurement non-invariance, the construct cannot be meaningfully tested or construed across groups. Measurement invariance with person residual is examined for the GAA 2.0 assessment and the person fit indices from Winsteps output were used as the person residual indicator. With the spring equating document for the 2023 administration, person infit and outfit mean squared error (MSQ) was used to flag students by person fit for demographic subgroups including gender, ethnicity (white and non-White), and primary disability (SRC05, SRC07, SRC08). Also, the person fit of infit Z-values was used to plot and examine the invariance among subgroups. The results of this analysis indicate no patterned differences in person fit residuals by gender and ethnicity. While no substantive difference was observed by primary disability as well, potential patterns are explored. This study further examined the differences in person fit among the primary disabilities.

### **Person Fit Z-values Across Subgroups**

Figure 1 is the same as the plots shown in the spring equating document. The plots include gender and ethnicity subgroups along with the primary disability groups. As seen in Figure 1, the distribution of gender and ethnicity are very similar among subgroups. Among the disability subgroups, the distribution of the SRC05 tends to have a narrower range than SRC07 or SRC08. The mean of the SRC05 is within the Q1 and Q3 range of SRC07 and SRC08, so this implies no statistically significant differences among the disability subgroups. However, the tendency of SRC05 to have a narrower range of person fit indices was examined closely by regression model and the visual presentation of the infit Z-value distribution across the scale score by disability categories.

### **Regression Analysis**

Regression of person fit by demographic variables was conducted to examine which variables have a more significant effect on the size of infit Z-values controlling for multiple variables. The dependent variable was the absolute values of the infit Z-values because both directions of the magnitude can happen in one demographic category. We are interested in how infit Z-values are affected by predictors. The mean square error could be used, instead, however, the consistency of using the same variable with Figure 1 was considered. The predictors were Gender, Ethnicity (White/non-White), SRC05, SRC07, SRC08, and Scale Score.

A benefit of multiple regression is that the effects of the multiple variables are examined at the same time. By including multiple variables, essentially other predictors are controlled to examine the effect of each predictor, unlike Figure 1 where only one variable is examined at once. One limitation of this model is that the coefficient slopes for SRC05, SRC07, and SRC08 are difference from students with other primary disability.

Across all grades and content areas, scale score was a significant predictor. This means that the magnitude of the person fit indices depended on the scale score levels. As can be observed from Figure 2, the infit values are large in magnitude in the mid-ranges of the scale score.

Ethnicity was not a significant predictor across all grades and content areas. Gender and disability categories were mostly not statistically significant. They were significant in the following grades and content areas. There is no consistent clear pattern of the statistical significance, but higher grades tend to have the disability predictors statistically significant.

Gender: ELA and Mathematics Grade 5

SRC05: Mathematics grades 6 and 8, and science grade 8

SRC07: ELA grades 7 and 8, Mathematics grade 6, and social studies grade HS

SRC08: ELA grades 7 and 8, HS, Mathematics grade 6

### **Person Infit Z and Scale Score**

Figure 2 shows the distribution of person infit Z-values across scale scores. As it was described above, the scale score was a significant predictor of the magnitude of the person infit Z indices, and the range of Person fit is wider in the mid-range of the scale score. Figure 3 shows the distribution of person infit Z-values across scale scores by disability categories. The scale score range differs by disability category, thus the distribution of the person infit appears different due to the scale score range for each disability category. With the mid-range of the scale score, the SRC05 tends to have less variability of the person fit Z-values.

### Conclusion

The analysis conducted did not show strong evidence that person fit indices differ by disability categories. The narrower range of the person fit Z-values distribution with SRC05 than SRC07 or SRC08 was observed with some grades and content areas, however, it appears there is no consistent statistical difference of SRC05 from other primary disability categories. From the analysis conducted, we conclude that the assumption of measurement invariance based on person residuals is met for the GAA 2.0 assessments.

### Tables and Figures

**Table 1. ELA Grade 3 Regression Analysis**

	<b>Coefficient</b>	<b>SE</b>	<b>T-statistics</b>	<b>P-value</b>
<b>Intercept</b>	5.1298	0.5023	10.2122	0.00
<b>Gender (M=1)</b>	0.0027	0.0364	0.0730	0.94
<b>White</b>	-0.0354	0.0371	-0.9519	0.34
<b>SRC05</b>	-0.0797	0.0553	-1.4401	0.15
<b>SRC07</b>	0.0648	0.0528	1.2280	0.22
<b>SRC08</b>	-0.0001	0.0385	-0.0016	1.00
<b>Scale Score</b>	-0.0030	0.0004	-8.6110	0.00

**Table 2. ELA Grade 4 Regression Analysis**

	<b>Coefficient</b>	<b>SE</b>	<b>T-statistics</b>	<b>P-value</b>
<b>Intercept</b>	5.1289	0.4703	10.9058	0.00
<b>Gender (M=1)</b>	-0.0339	0.0347	-0.9759	0.33
<b>White</b>	0.0097	0.0340	0.2863	0.77
<b>SRC05</b>	-0.1099	0.0571	-1.9239	0.05
<b>SRC07</b>	-0.0319	0.0534	-0.5967	0.55
<b>SRC08</b>	0.0345	0.0482	0.7162	0.47
<b>Scale Score</b>	-0.0030	0.0003	-9.2042	0.00

Note: SRC05 p-value > 0.05

**Table 3. ELA Grade 5 Regression Analysis**

	<b>Coefficient</b>	<b>SE</b>	<b>T-statistics</b>	<b>P-value</b>
<b>Intercept</b>	5.8161	0.4930	11.7977	0.00
<b>Gender (M=1)</b>	-0.0691	0.0346	-1.9967	0.05
<b>White</b>	-0.0125	0.0346	-0.3606	0.72
<b>SRC05</b>	-0.0891	0.0636	-1.4007	0.16
<b>SRC07</b>	0.1133	0.0591	1.9182	0.06
<b>SRC08</b>	-0.0252	0.0566	-0.4458	0.66
<b>Scale Score</b>	-0.0035	0.0003	-10.0871	0.00

**Table 4. ELA Grade 6 Regression Analysis**

	<b>Coefficient</b>	<b>SE</b>	<b>T-statistics</b>	<b>P-value</b>
<b>Intercept</b>	8.8210	0.5465	16.1421	0.00
<b>Gender (M=1)</b>	-0.0264	0.0338	-0.7816	0.43
<b>White</b>	-0.0366	0.0336	-1.0908	0.28
<b>SRC05</b>	-0.1143	0.0638	-1.7914	0.07
<b>SRC07</b>	-0.0918	0.0625	-1.4696	0.14
<b>SRC08</b>	-0.0413	0.0604	-0.6843	0.49
<b>Scale Score</b>	-0.0056	0.0004	-14.7178	0.00

**Table 5. ELA Grade 7 Regression Analysis**

	<b>Coefficient</b>	<b>SE</b>	<b>T-statistics</b>	<b>P-value</b>
<b>Intercept</b>	4.7099	0.3234	14.5619	0.00
<b>Gender (M=1)</b>	-0.0086	0.0329	-0.2621	0.79
<b>White</b>	0.0293	0.0334	0.8784	0.38
<b>SRC05</b>	0.0420	0.0576	0.7298	0.47
<b>SRC07</b>	0.1135	0.0568	1.9975	0.05
<b>SRC08</b>	0.1454	0.0548	2.6522	0.01
<b>Scale Score</b>	-0.0028	0.0002	-12.8589	0.00

**Table 6. ELA Grade 8 Regression Analysis**

	<b>Coefficient</b>	<b>SE</b>	<b>T-statistics</b>	<b>P-value</b>
<b>Intercept</b>	5.4579	0.3690	14.7893	0.00
<b>Gender (M=1)</b>	-0.0212	0.0309	-0.6874	0.49
<b>White</b>	-0.0404	0.0311	-1.2994	0.19
<b>SRC05</b>	-0.0120	0.0551	-0.2173	0.83
<b>SRC07</b>	0.1177	0.0545	2.1590	0.03
<b>SRC08</b>	0.1201	0.0529	2.2725	0.02
<b>Scale Score</b>	-0.0033	0.0003	-13.1130	0.00

**Table 7. ELA Grade HS Regression Analysis**

	<b>Coefficient</b>	<b>SE</b>	<b>T-statistics</b>	<b>P-value</b>
<b>Intercept</b>	5.1771	0.8413	6.1536	0.00
<b>Gender (M=1)</b>	0.0262	0.0597	0.4392	0.66
<b>White</b>	0.0063	0.0576	0.1098	0.91
<b>SRC05</b>	-0.0201	0.0983	-0.2044	0.84
<b>SRC07</b>	0.1531	0.0970	1.5787	0.11
<b>SRC08</b>	0.2939	0.0957	3.0701	0.00
<b>Scale Score</b>	-0.0038	0.0006	-6.4587	0.00

**Table 8. Mathematics Grade 3 Regression Analysis**

	<b>Coefficient</b>	<b>SE</b>	<b>T-statistics</b>	<b>P-value</b>
<b>Intercept</b>	3.2896	0.3683	8.9314	0.00
<b>Gender (M=1)</b>	-0.0227	0.0421	-0.5390	0.59
<b>White</b>	0.0274	0.0429	0.6392	0.52
<b>SRC05</b>	-0.0770	0.0648	-1.1878	0.24
<b>SRC07</b>	0.0073	0.0609	0.1200	0.90
<b>SRC08</b>	-0.0096	0.0447	-0.2157	0.83
<b>Scale Score</b>	-0.0017	0.0003	-6.4838	0.00

**Table 9. Mathematics Grade 4 Regression Analysis**

	<b>Coefficient</b>	<b>SE</b>	<b>T-statistics</b>	<b>P-value</b>
<b>Intercept</b>	8.7061	0.5703	15.2662	0.00
<b>Gender (M=1)</b>	0.0419	0.0357	1.1738	0.24
<b>White</b>	-0.0086	0.0352	-0.2443	0.81
<b>SRC05</b>	0.0199	0.0585	0.3411	0.73
<b>SRC07</b>	0.1057	0.0551	1.9194	0.06
<b>SRC08</b>	0.0576	0.0500	1.1531	0.25
<b>Scale Score</b>	-0.0056	0.0004	-13.8579	0.00

**Table 10. Mathematics Grade 5 Regression Analysis**

	<b>Coefficient</b>	<b>SE</b>	<b>T-statistics</b>	<b>P-value</b>
<b>Intercept</b>	5.7968	0.5689	10.1903	0.00
<b>Gender (M=1)</b>	-0.0784	0.0387	-2.0237	0.04
<b>White</b>	-0.0016	0.0389	-0.0414	0.97
<b>SRC05</b>	-0.0220	0.0708	-0.3105	0.76
<b>SRC07</b>	-0.0448	0.0660	-0.6789	0.50
<b>SRC08</b>	-0.0478	0.0635	-0.7521	0.45
<b>Scale Score</b>	-0.0034	0.0004	-8.4108	0.00

**Table 11. Mathematics Grade 6 Regression Analysis**

	<b>Coefficient</b>	<b>SE</b>	<b>T-statistics</b>	<b>P-value</b>
<b>Intercept</b>	8.6934	0.5979	14.5394	0.00
<b>Gender (M=1)</b>	0.0016	0.0344	0.0462	0.96
<b>White</b>	0.0075	0.0340	0.2195	0.83
<b>SRC05</b>	0.1467	0.0642	2.2828	0.02
<b>SRC07</b>	0.1745	0.0632	2.7627	0.01
<b>SRC08</b>	0.1428	0.0610	2.3415	0.02
<b>Scale Score</b>	-0.0056	0.0004	-13.4033	0.00

**Table 12. Mathematics Grade 7 Regression Analysis**

	<b>Coefficient</b>	<b>SE</b>	<b>T-statistics</b>	<b>P-value</b>
<b>Intercept</b>	5.4249	0.4929	11.0064	0.00
<b>Gender (M=1)</b>	0.0464	0.0349	1.3278	0.18
<b>White</b>	0.0575	0.0356	1.6147	0.11
<b>SRC05</b>	-0.0493	0.0617	-0.7993	0.42
<b>SRC07</b>	0.0831	0.0600	1.3836	0.17
<b>SRC08</b>	-0.0157	0.0585	-0.2692	0.79
<b>Scale Score</b>	-0.0033	0.0003	-9.3993	0.00

**Table 13. Mathematics Grade 8 Regression Analysis**

	<b>Coefficient</b>	<b>SE</b>	<b>T-statistics</b>	<b>P-value</b>
<b>Intercept</b>	7.5459	0.5182	14.5605	0.00
<b>Gender (M=1)</b>	-0.0208	0.0364	-0.5710	0.57
<b>White</b>	-0.0636	0.0365	-1.7448	0.08
<b>SRC05</b>	-0.1801	0.0637	-2.8286	0.00
<b>SRC07</b>	-0.0733	0.0634	-1.1548	0.25
<b>SRC08</b>	-0.0052	0.0614	-0.0846	0.93
<b>Scale Score</b>	-0.0047	0.0004	-12.8621	0.00

**Table 14. Mathematics Grade HS Regression Analysis**

	<b>Coefficient</b>	<b>SE</b>	<b>T-statistics</b>	<b>P-value</b>
<b>Intercept</b>	6.2257	0.6371	9.7717	0.00
<b>Gender (M=1)</b>	-0.0436	0.0381	-1.1423	0.25
<b>White</b>	-0.0173	0.0372	-0.4664	0.64
<b>SRC05</b>	-0.0845	0.0627	-1.3481	0.18
<b>SRC07</b>	0.0402	0.0623	0.6451	0.52
<b>SRC08</b>	0.0059	0.0614	0.0957	0.92
<b>Scale Score</b>	-0.0038	0.0004	-8.4325	0.00

**Table 15. Science Grade 5 Regression Analysis**

	<b>Coefficient</b>	<b>SE</b>	<b>T-statistics</b>	<b>P-value</b>
<b>Intercept</b>	4.8822	0.3729	13.0922	0.00
<b>Gender (M=1)</b>	-0.0373	0.0358	-1.0422	0.30
<b>White</b>	0.0479	0.0359	1.3351	0.18
<b>SRC05</b>	-0.0970	0.0656	-1.4785	0.14
<b>SRC07</b>	0.0454	0.0612	0.7431	0.46
<b>SRC08</b>	0.0390	0.0587	0.6636	0.51
<b>Scale Score</b>	-0.0028	0.0003	-10.8592	0.00

**Table 16. Science Grade 8 Regression Analysis**

	<b>Coefficient</b>	<b>SE</b>	<b>T-statistics</b>	<b>P-value</b>
<b>Intercept</b>	6.5322	0.5828	11.2078	0.00
<b>Gender (M=1)</b>	-0.0403	0.0346	-1.1662	0.24
<b>White</b>	-0.0397	0.0349	-1.1370	0.26
<b>SRC05</b>	-0.1553	0.0610	-2.5446	0.01
<b>SRC07</b>	0.0404	0.0606	0.6676	0.50
<b>SRC08</b>	0.0529	0.0587	0.9018	0.37
<b>Scale Score</b>	-0.0040	0.0004	-9.7410	0.00

**Table 17. Science Grade HS Regression Analysis**

	<b>Coefficient</b>	<b>SE</b>	<b>T-statistics</b>	<b>P-value</b>
<b>Intercept</b>	4.3474	0.4160	10.4507	0.00
<b>Gender (M=1)</b>	-0.0471	0.0411	-1.1451	0.25
<b>White</b>	-0.0121	0.0400	-0.3024	0.76
<b>SRC05</b>	-0.1296	0.0679	-1.9081	0.06
<b>SRC07</b>	0.0343	0.0675	0.5076	0.61
<b>SRC08</b>	0.0672	0.0664	1.0122	0.31
<b>Scale Score</b>	-0.0024	0.0003	-8.3992	0.00

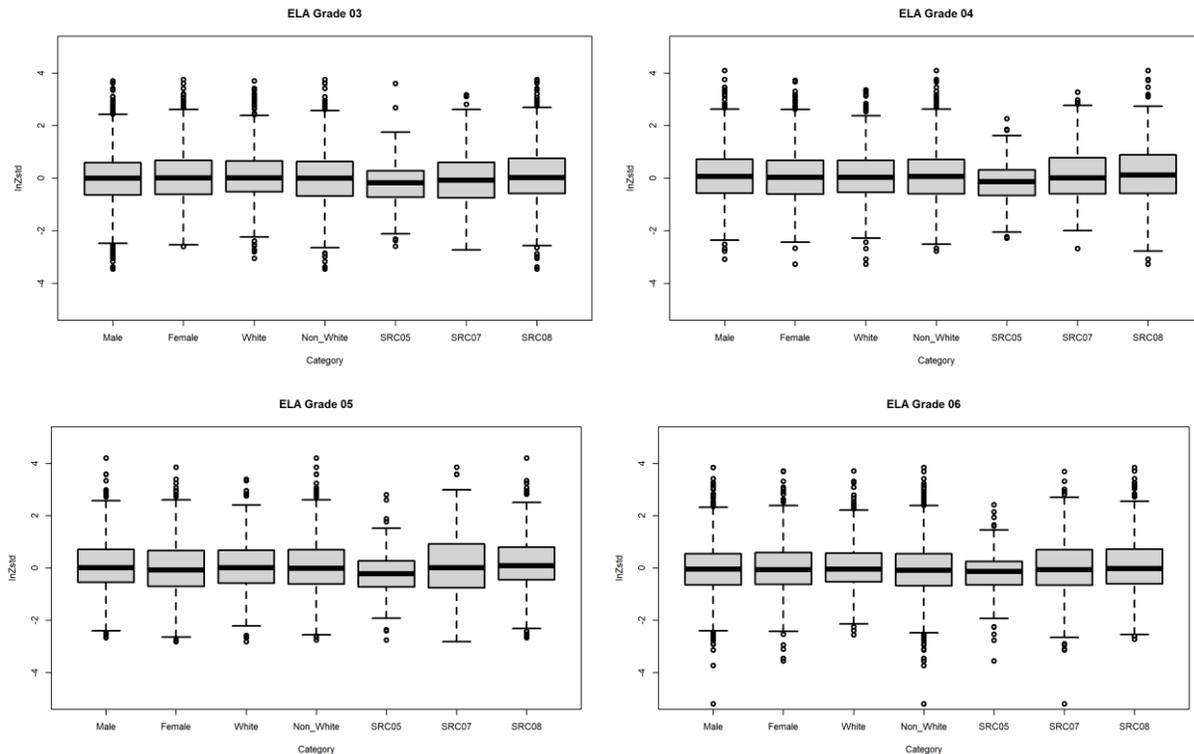
**Table 18. Social Studies Grade 8 Regression Analysis**

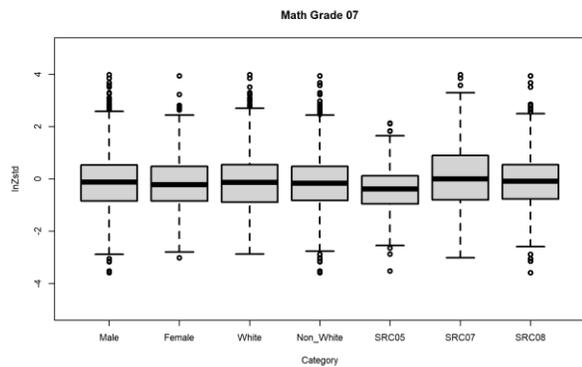
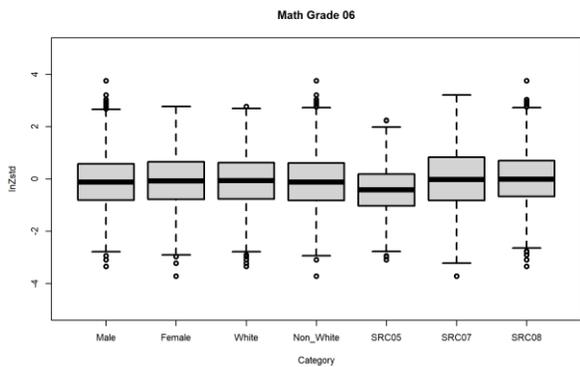
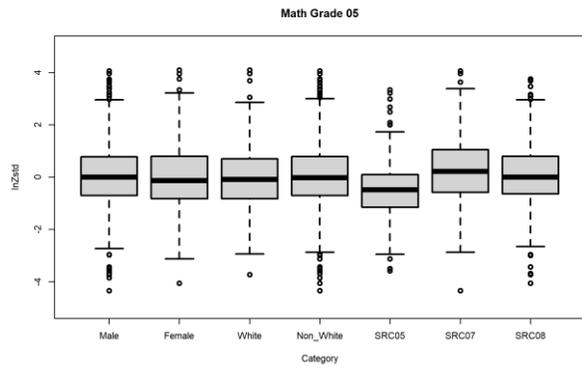
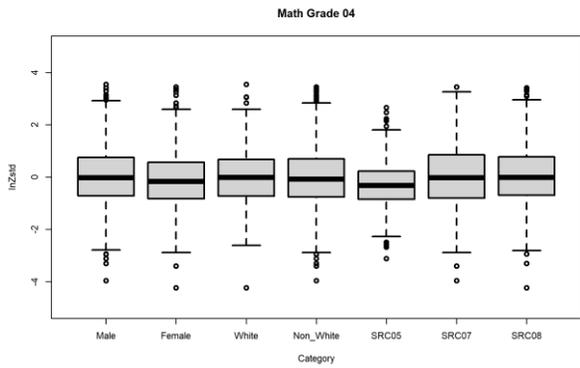
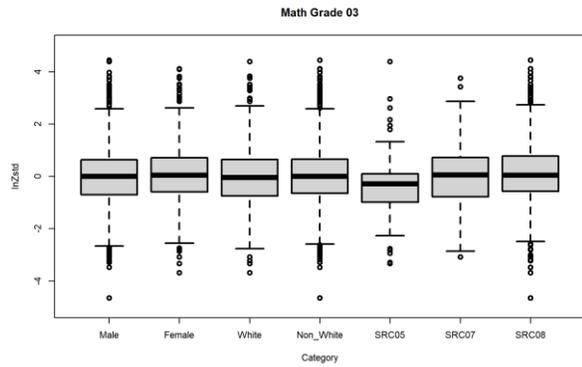
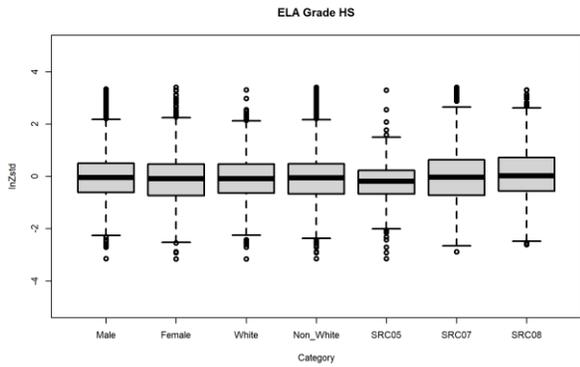
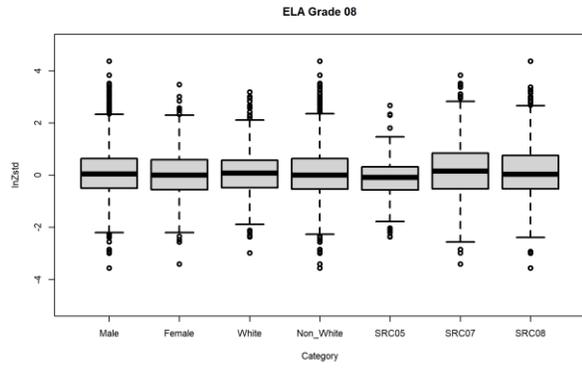
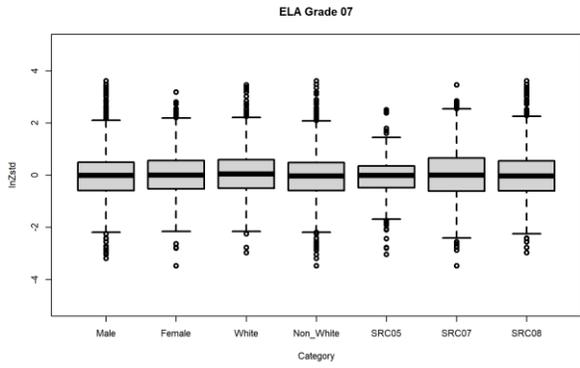
	Coefficient	SE	T-statistics	P-value
<b>Intercept</b>	5.9261	0.3861	15.3484	0.00
<b>Gender (M=1)</b>	-0.0281	0.0339	-0.8298	0.41
<b>White</b>	0.0180	0.0340	0.5298	0.60
<b>SRC05</b>	-0.0299	0.0594	-0.5041	0.61
<b>SRC07</b>	0.0990	0.0595	1.6641	0.10
<b>SRC08</b>	-0.0093	0.0575	-0.1610	0.87
<b>Scale Score</b>	-0.0036	0.0003	-13.3179	0.00

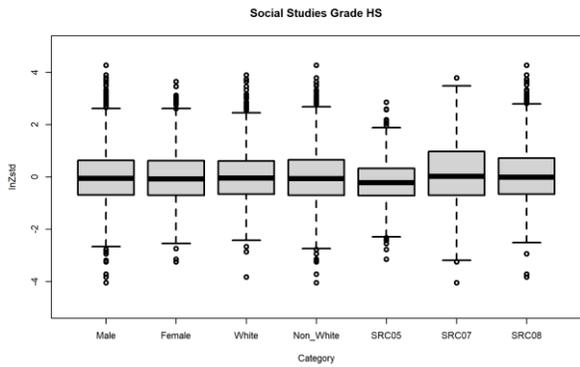
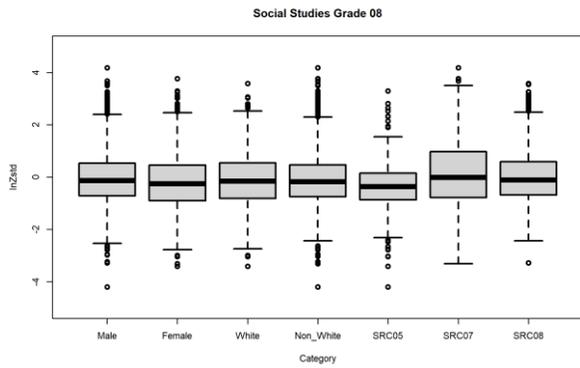
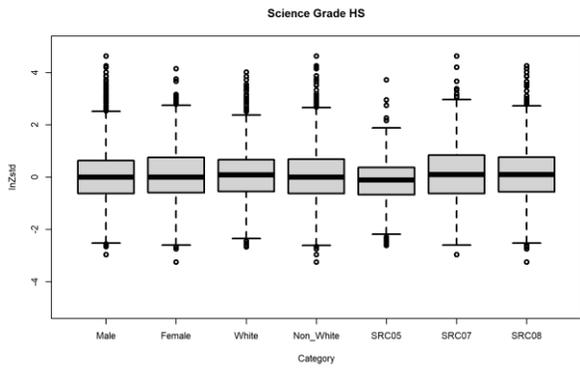
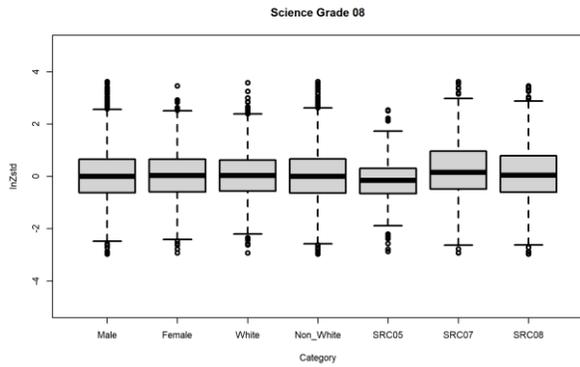
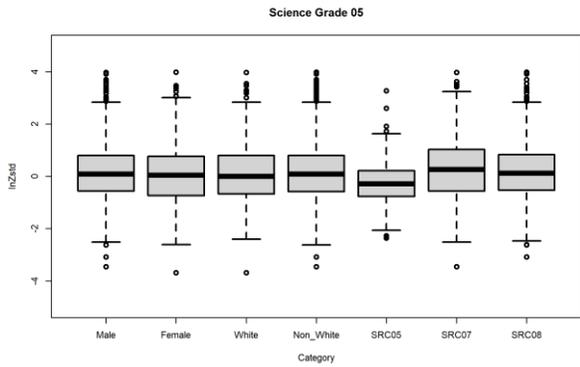
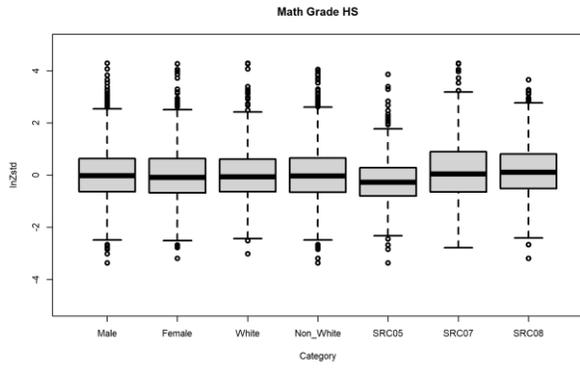
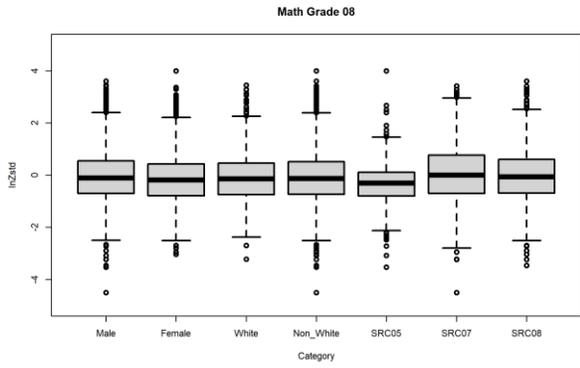
**Table 19. Social Studies Grade HS Regression Analysis**

	Coefficient	SE	T-statistics	P-value
<b>Intercept</b>	3.6169	0.4819	7.5051	0.00
<b>Gender (M=1)</b>	-0.0024	0.0420	-0.0569	0.95
<b>White</b>	-0.0273	0.0410	-0.6655	0.51
<b>SRC05</b>	-0.1007	0.0689	-1.4614	0.14
<b>SRC07</b>	0.1390	0.0677	2.0530	0.04
<b>SRC08</b>	0.0961	0.0672	1.4289	0.15
<b>Scale Score</b>	-0.0019	0.0003	-5.7742	0.00

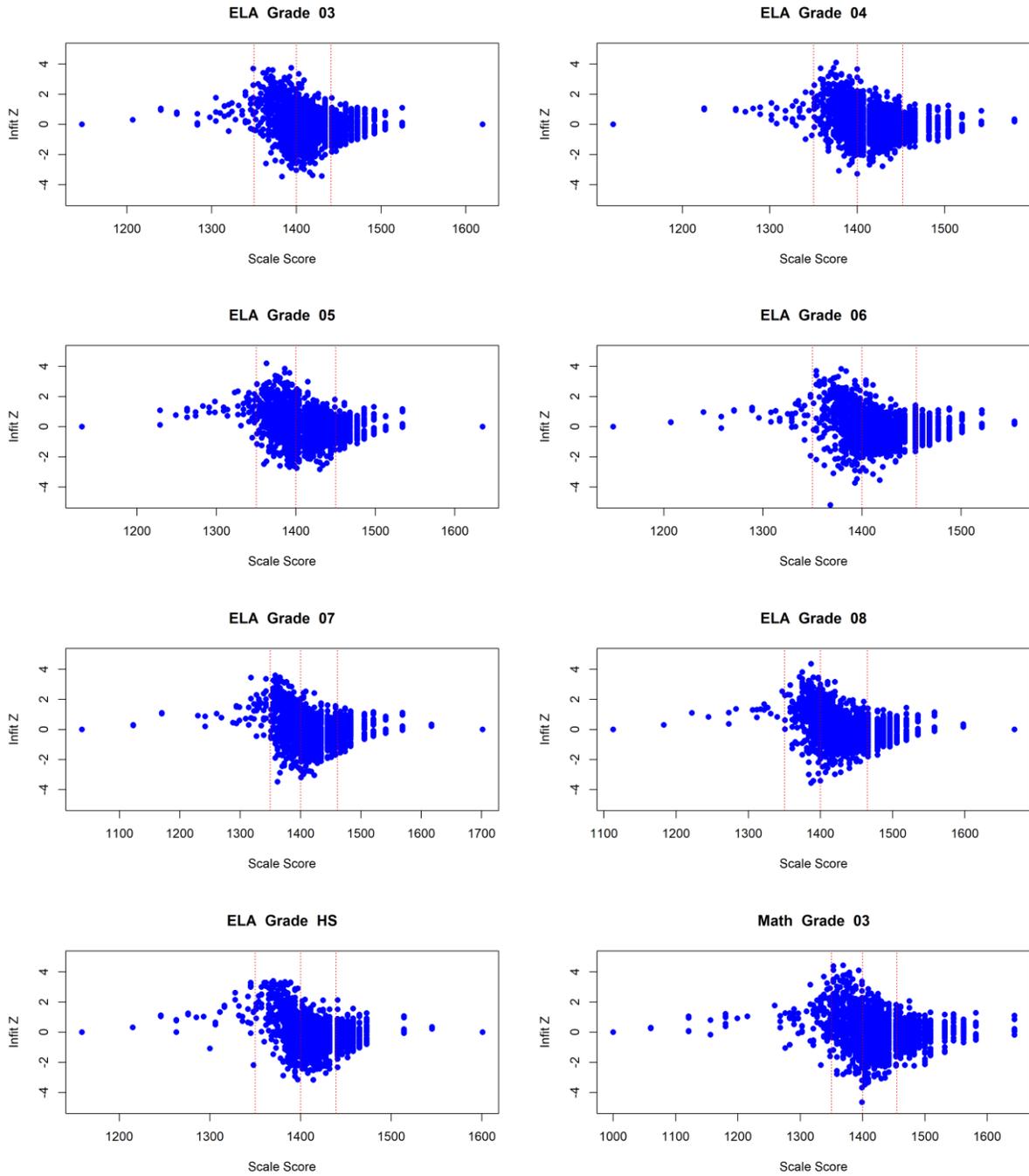
**Figure 1. Person Fit of Infit Z-Values for Subgroups**

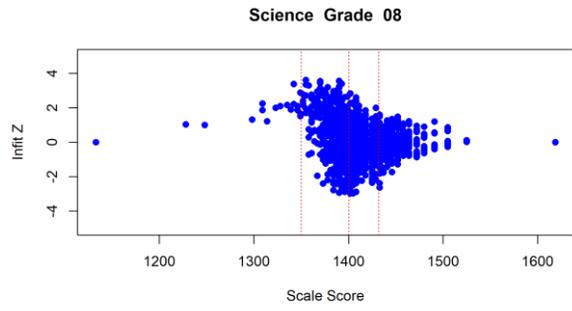
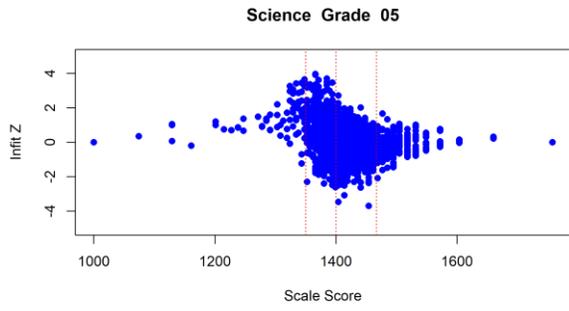
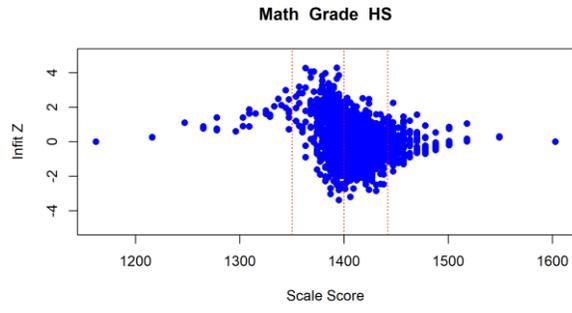
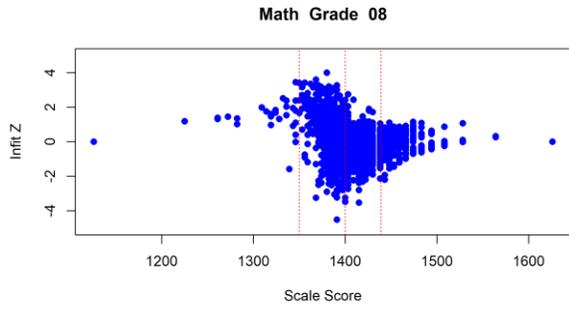
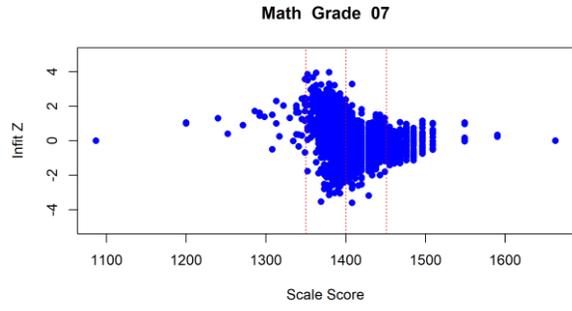
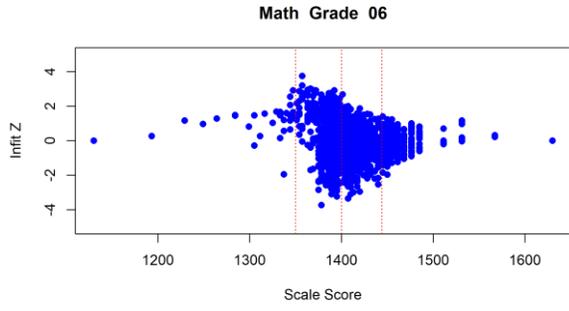
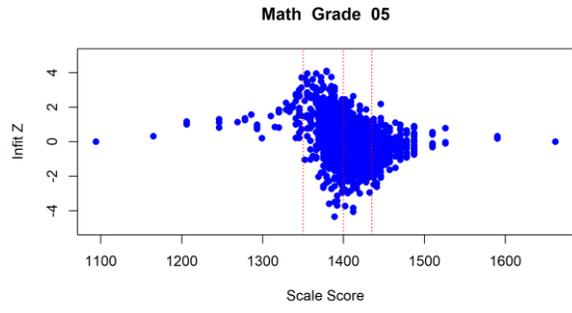
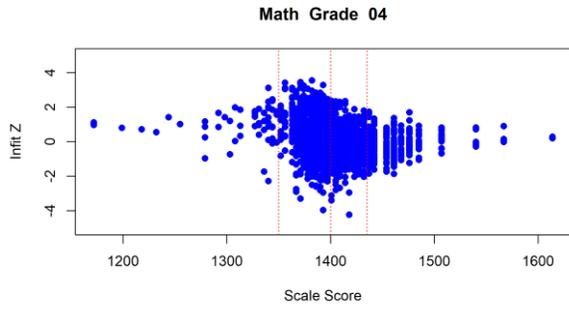




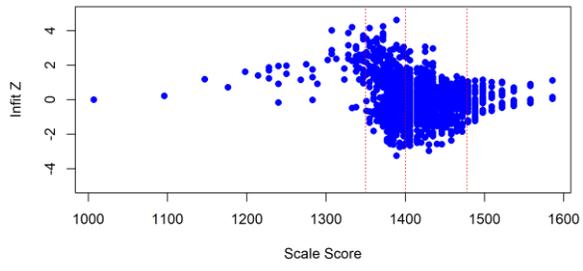


**Figure 2. Person Infit Z and Scale Score**

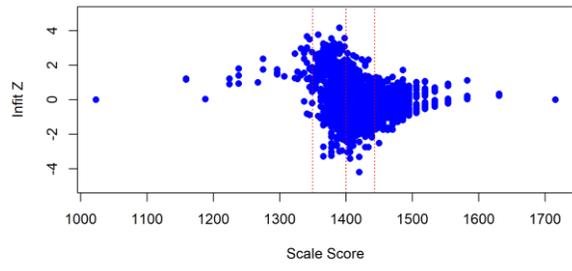




**Science Grade HS**



**Social Studies Grade 08**



**Social Studies Grade HS**

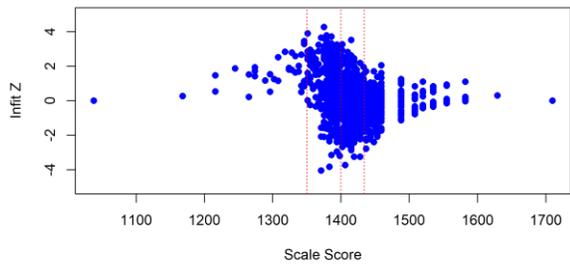
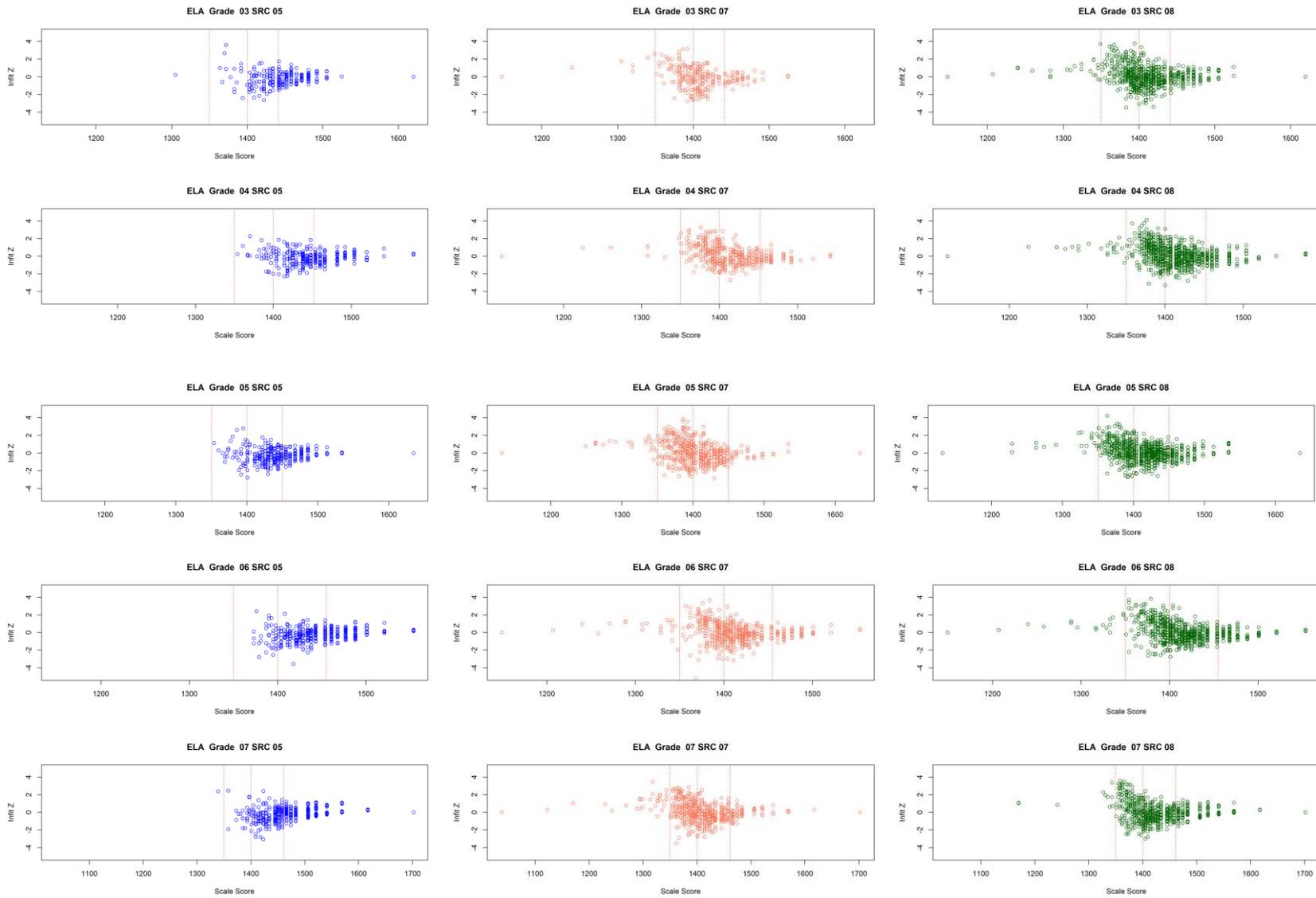
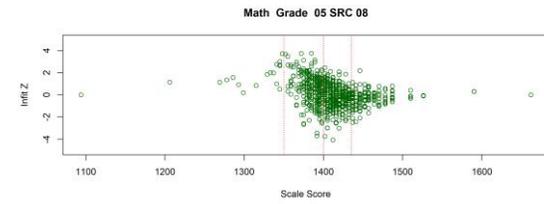
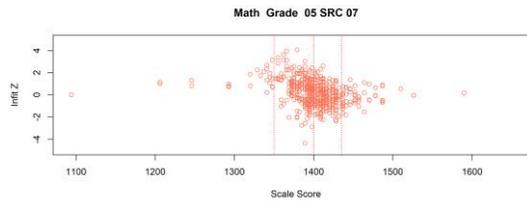
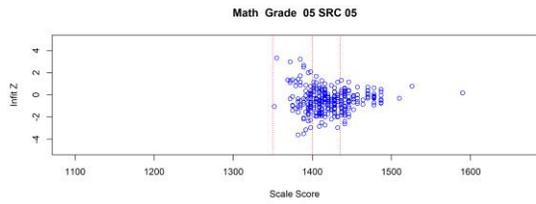
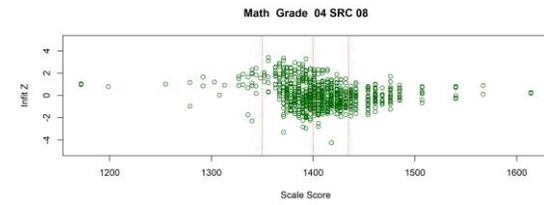
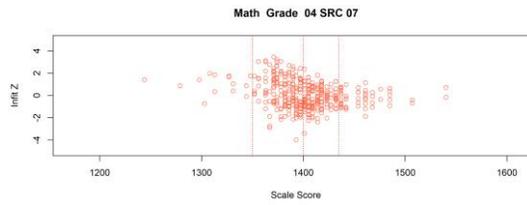
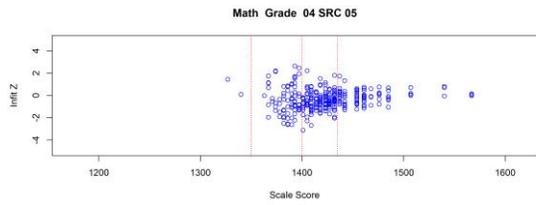
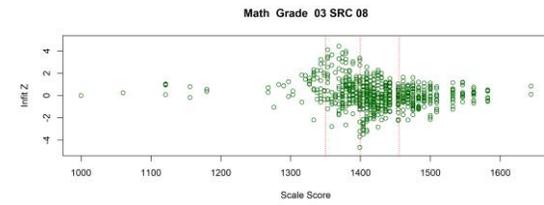
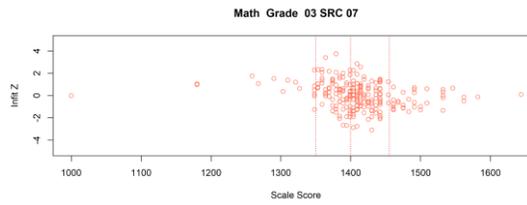
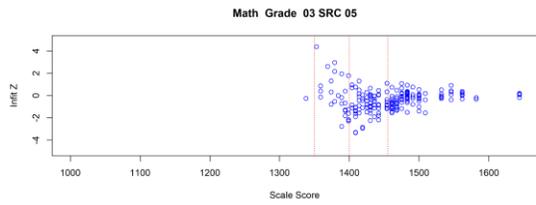
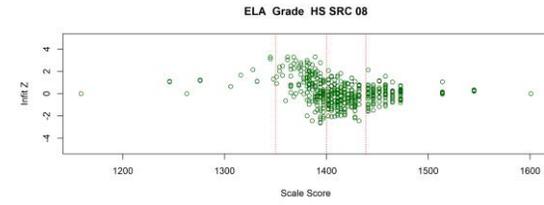
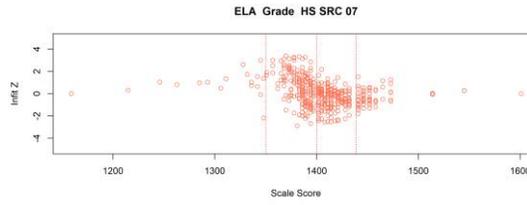
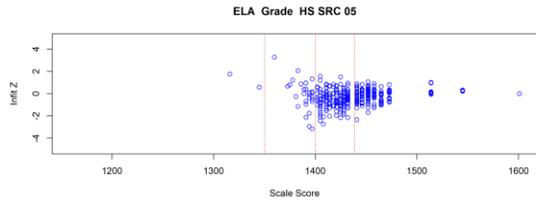
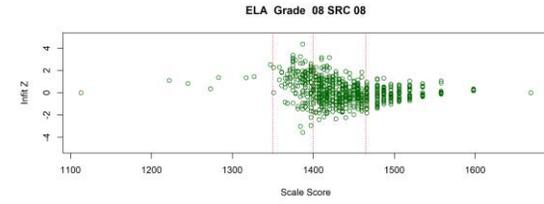
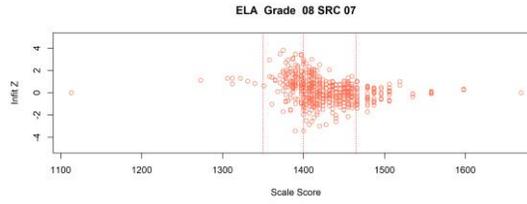
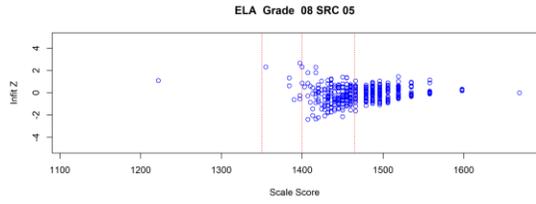
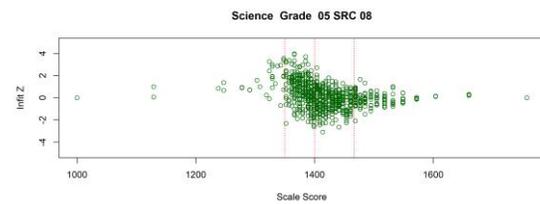
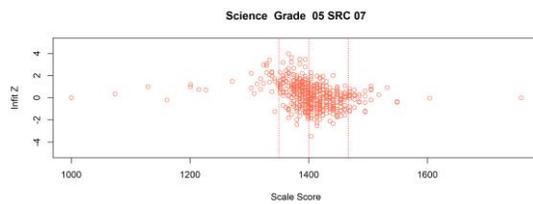
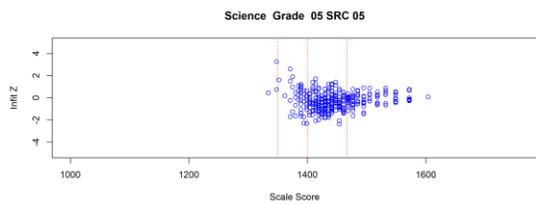
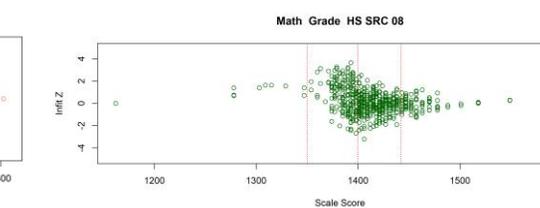
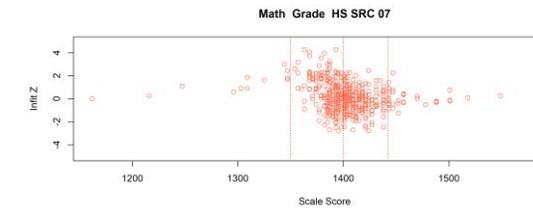
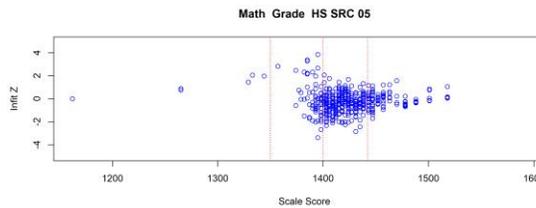
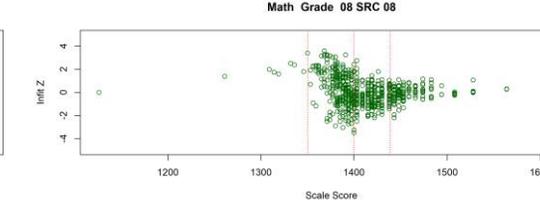
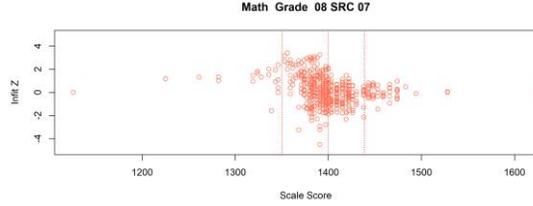
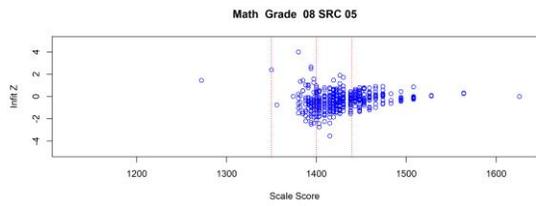
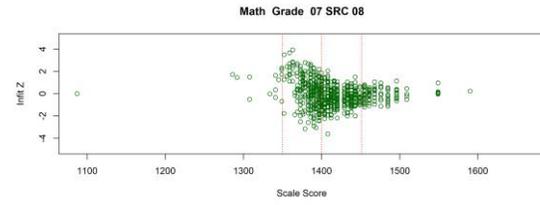
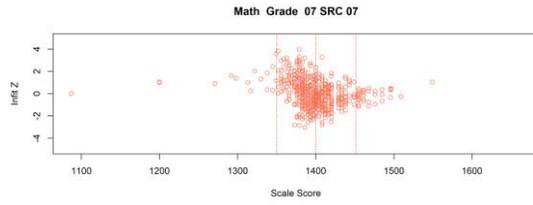
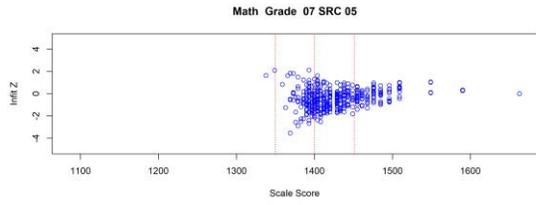
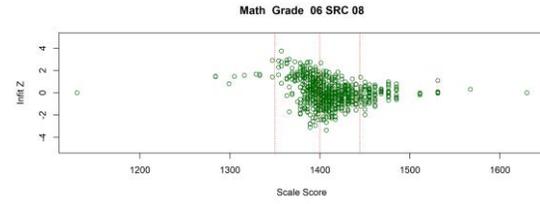
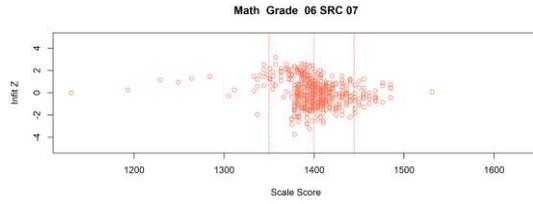
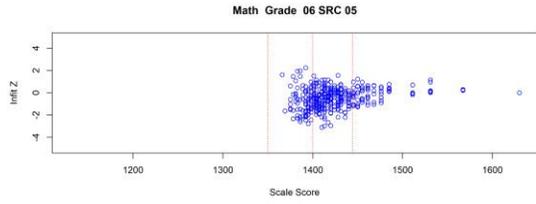
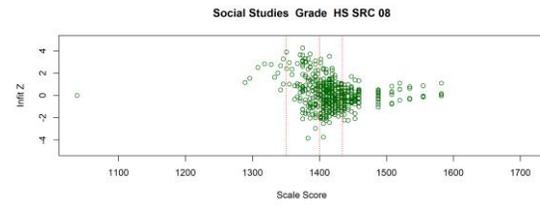
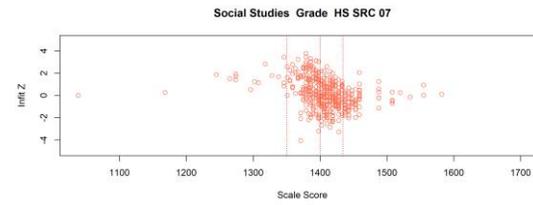
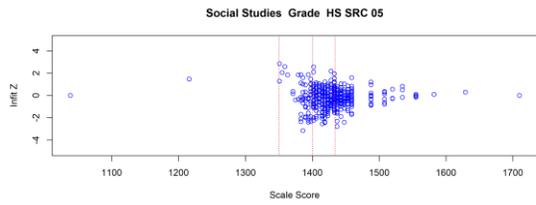
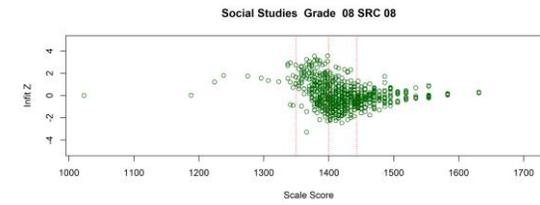
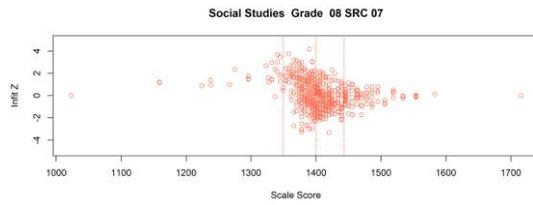
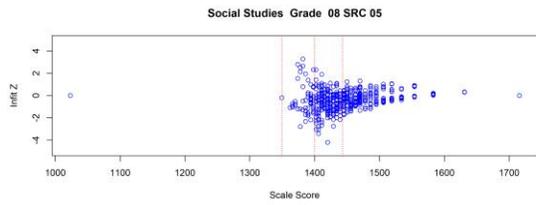
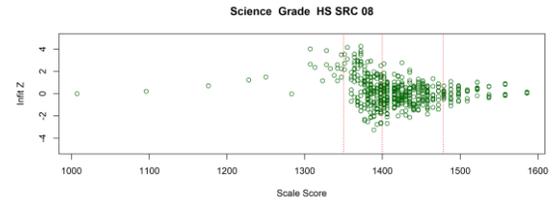
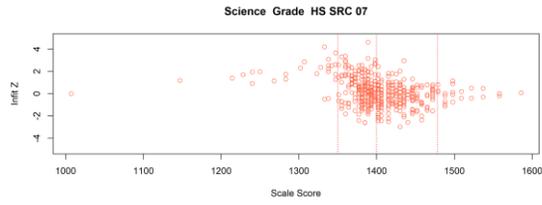
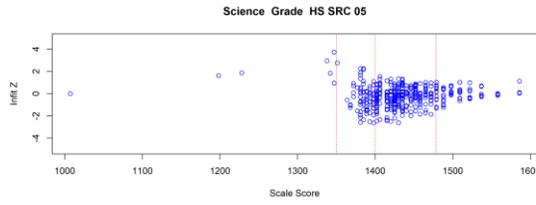
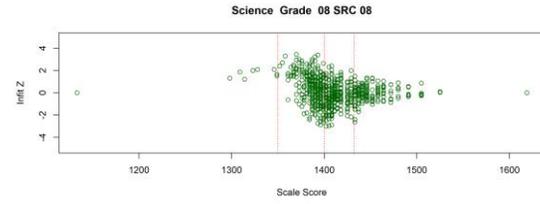
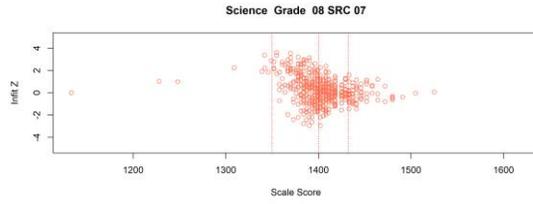
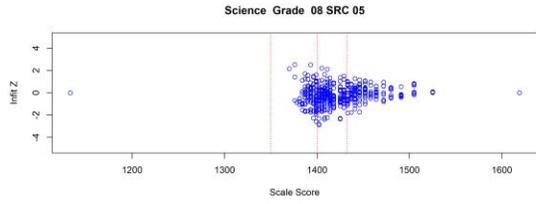


Figure 3. Person Infit Z and Scale Score by Disability Categories









Georgia Alternate Assessment 2.0 (GAA 2.0)

# Science Item Cognitive Validity Study

2023 Spring

Data Recognition Corporation Psychometrics Services  
11-1-2023

## **Table of Contents**

Purpose of the Report.....	3
Appropriateness for Grade Level.....	3
Test Design and Cognitive Complexity.....	5
Item Difficulty Modeling: Predictors .....	8
Regression Results .....	10
Conclusion.....	12

## **Purpose of the Report**

The purpose of this analysis is to investigate the cognitive processes used to complete GAA 2.0 science tasks as part of ongoing validity evaluations for this assessment. This line of research responds to peer review feedback which requested additional cognitive validity evidence, specifically for science assessments, because science was not included in the original cog labs study.

CE 3.2 Validity Based on Cognitive Processes - Documentation to show that the State's assessments tap the intended science cognitive processes appropriate for each grade level.

We examine the cognitive complexity for each grade level and investigate the test design and cognitive relationship through this study. Data used for this study is from the Spring 2023 administration and it is the same data as the psychometric spring analysis.

All tasks for the GAA 2.0 undergo extensive expert and educator review before and after field testing to determine whether the cognitive complexity of the knowledge, skills, and abilities required to successfully complete each task align to the cognitive complexity of Georgia's Extended Content standards. This expert judgement which takes into account ALD alignment, DOK, adherence to the item specifications, and more, is the primary source of evidence that the GAA 2.0 tasks tap the intended science cognitive processes appropriate for each grade level, at the appropriate level of complexity. This study serves as an empirical analysis by modeling the relationship between difficulty, ALD alignment, DOK, and more, as collective indicators of cognitive complexity. This post-administration evaluation supplements the validity evidence by demonstrating that the within-task complexity is functioning as intended, the tasks call upon grade-level appropriate science cognitive processes in increasing complexity as specified in the extended content standards, and the relationship to ALDs and task part level is functioning as intended in the test design.

## **Appropriateness for Grade Level**

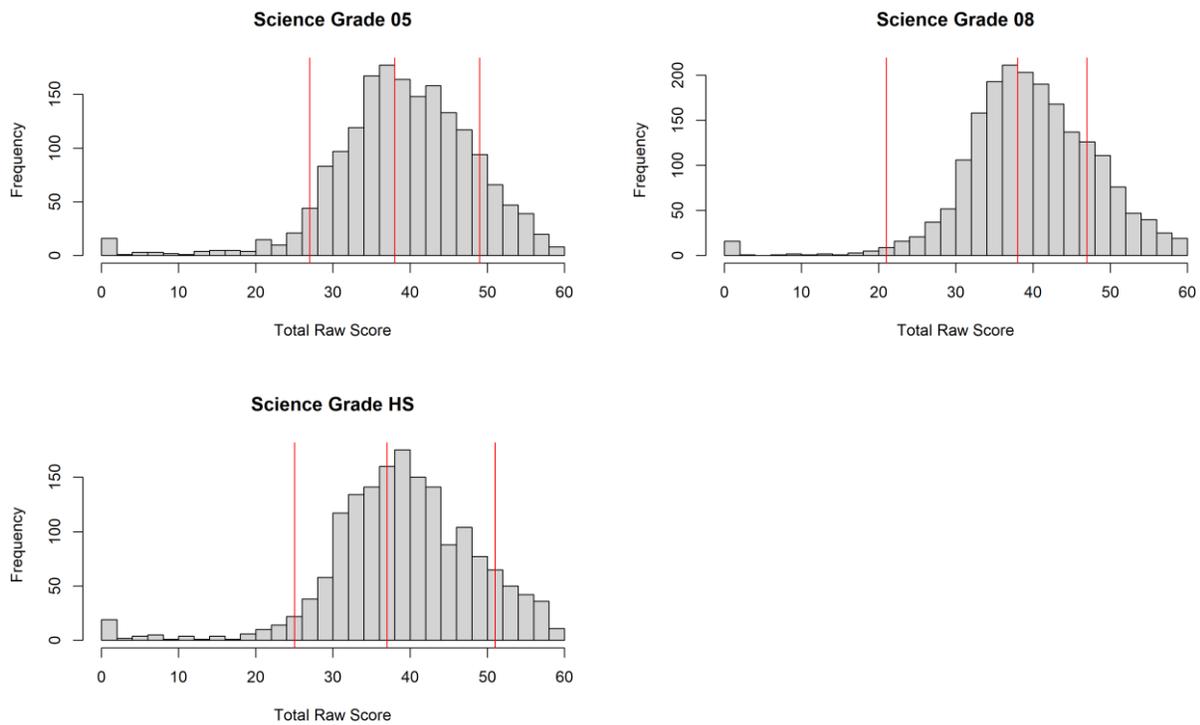
As a first step to evaluating cognitive complexity by grade level, we considered the distribution of the item p-value and test total score distribution. P-values range from 0.42 to 0.88 across grades, indicating that item p-values are not too low where most students cannot score points, or not too high where most students receive full points. The item difficulty ranges expressed in item p-value show that the items on the science assessments are appropriate for the grade level.

**Table 1. P-Value Summary Statistics with 2023 Administration**

Grade	N	Minimum	Maximum	Mean	SD
5	30	0.42	0.88	0.66	0.12
8	30	0.45	0.87	0.67	0.11
HS	30	0.49	0.84	0.66	0.11

The raw score distributions with cut scores show that the proportion of the students who score lower end and higher end are small. Most of the students are between 30 to 50 raw score points out of 60 maximum points on a form. The raw score distributions indicate that test forms are appropriate for grade levels.

**Figure 1. Total Raw Score Distribution**



## Test Design and Cognitive Complexity

Figure 2 shows the box plot of IRT item difficulty for each part. Item part and the IRT item difficulty relationship shows that Part A items consistently have lower item difficulty than Part B and Part C items. In grades 8 and HS, Part B and C items have less pronounced patterns of item difficulty. Item difficulty in all grades, however, is clearly progressing within tasks. The box plots in Figure 2 show the item difficulty for each part across all tasks. Figure 3 displays the progression of part item difficulty and student ability Wright Maps (for ELA and mathematics, the same plots are shown in the 2023 technical report appendix F). Not all, but in general, the item difficulties for Part A items were lower than those of Parts B and C items. The differences in item difficulty between Parts B and C are generally less pronounced.

**Figure 2. Item Difficulty and Item Part**

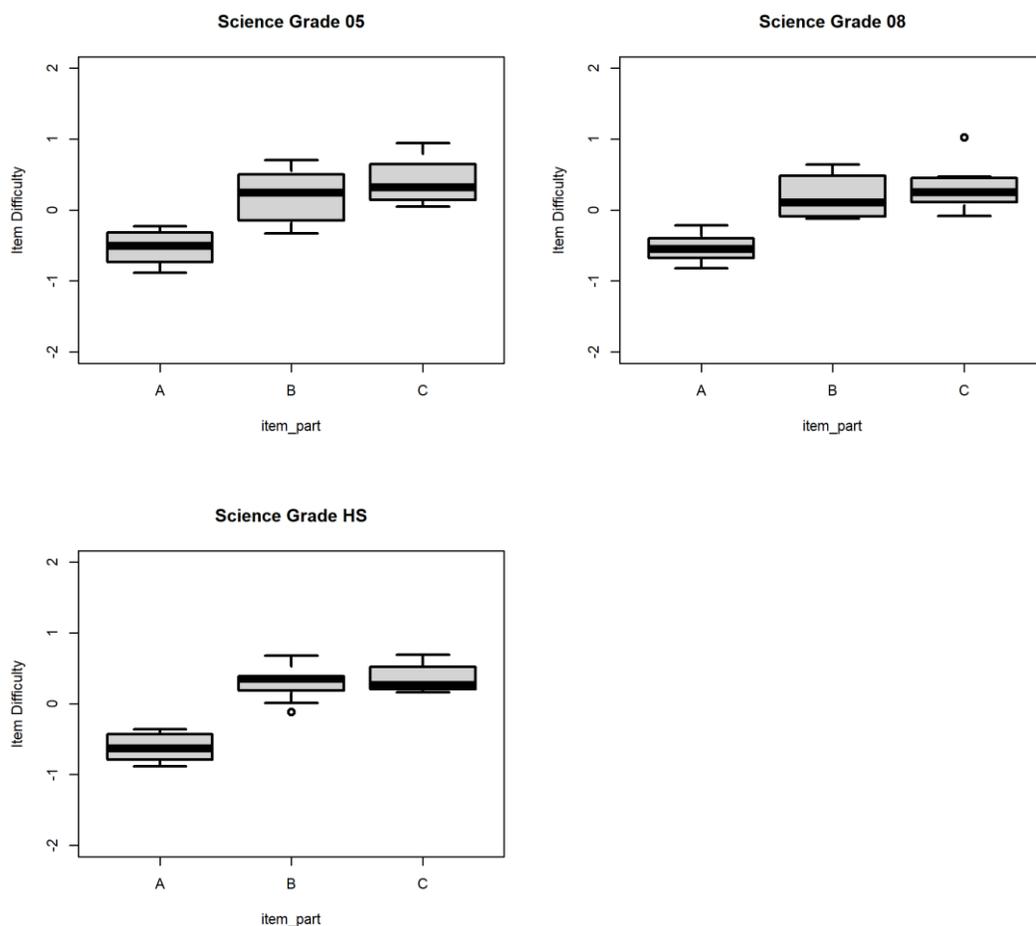
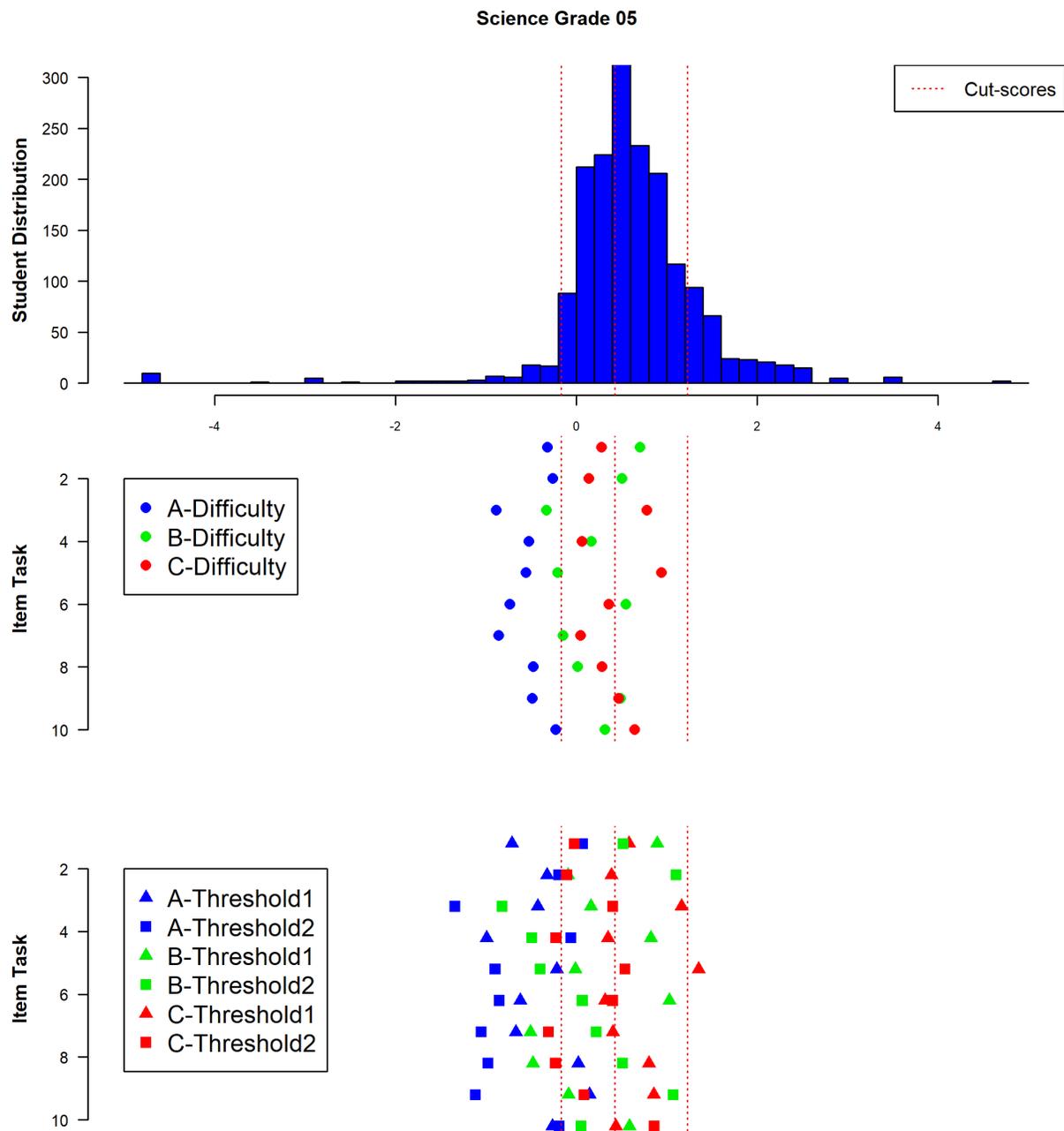
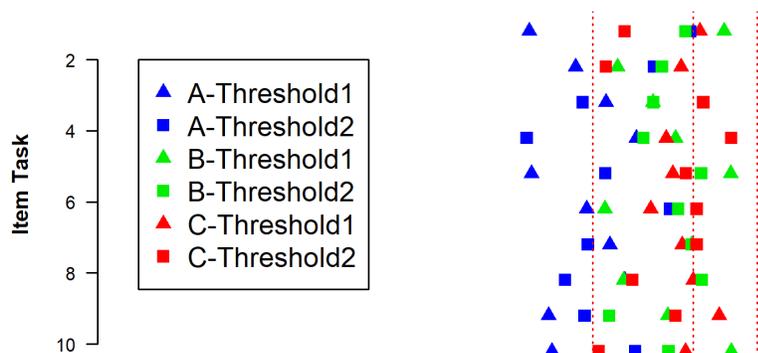
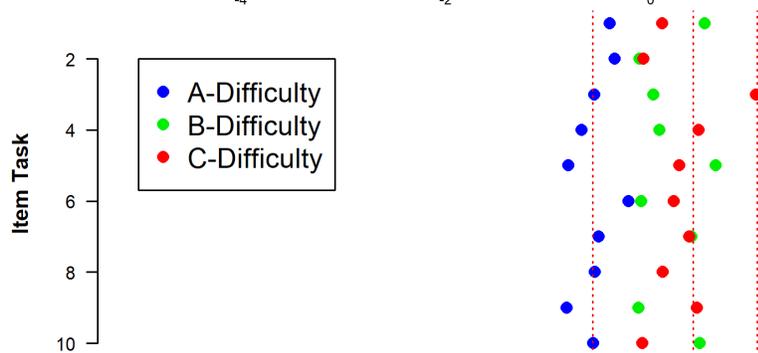
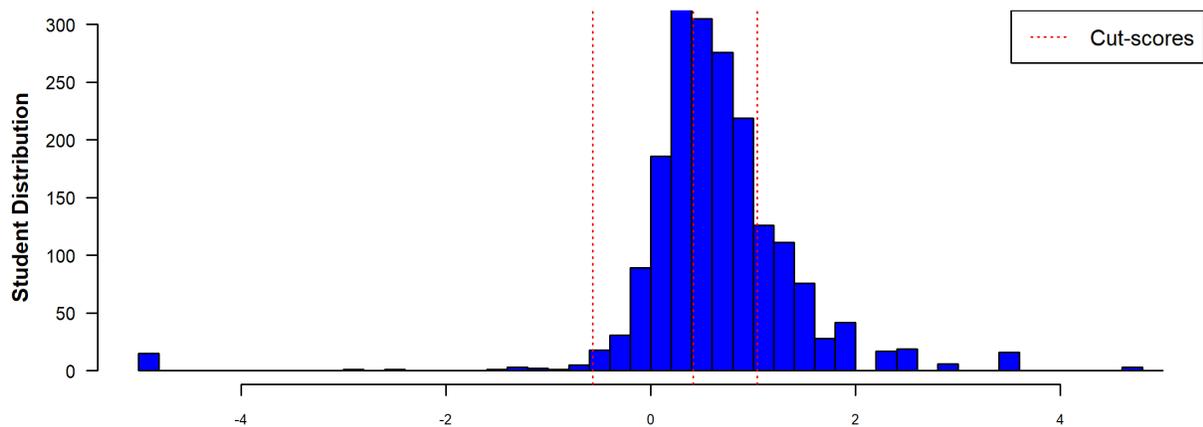
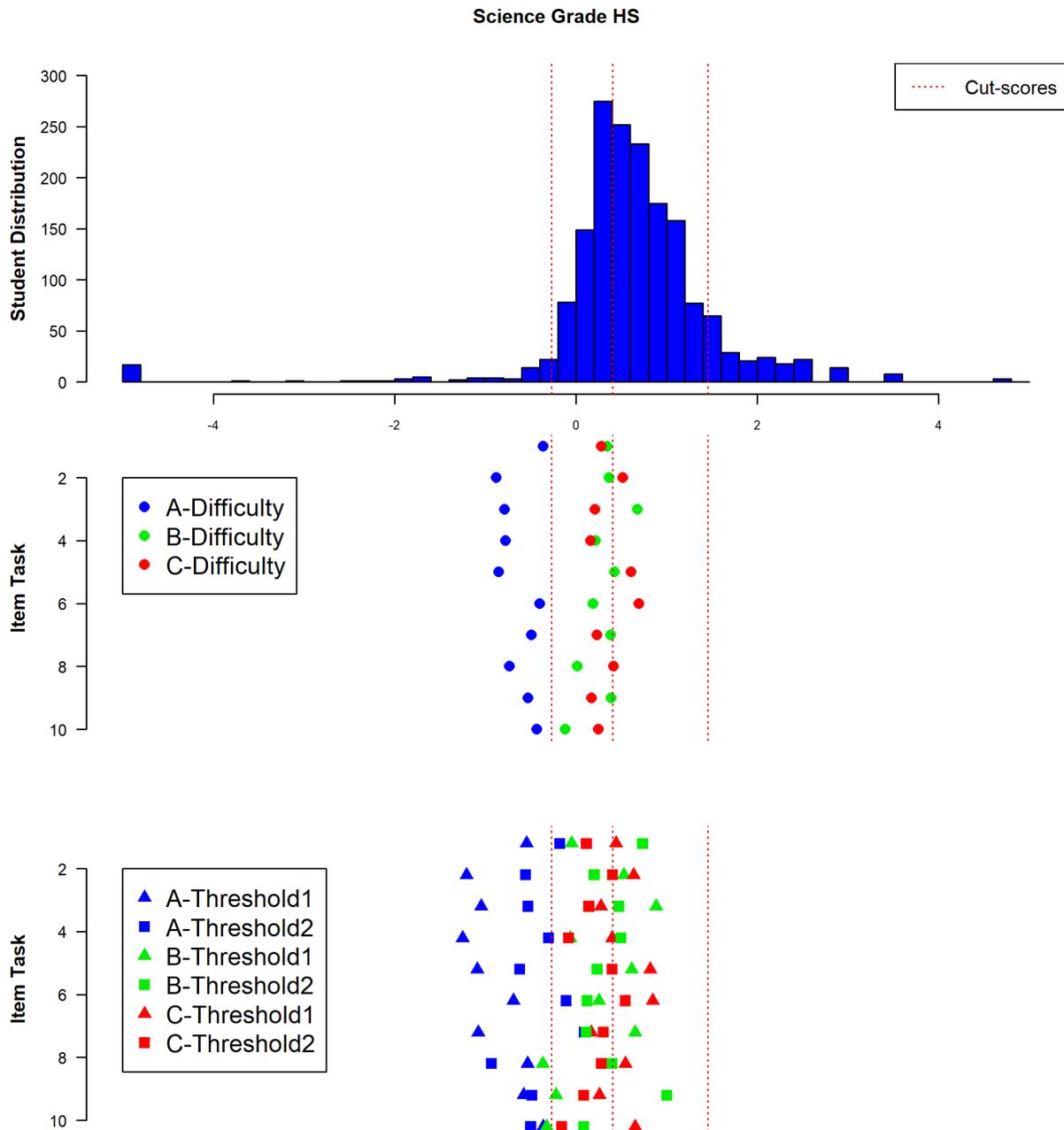


Figure 3. Science 2023 Item Difficulty-Student Ability Wright Maps



### Science Grade 08





### Item Difficulty Modeling: Predictors

To examine the item difficulty progression as the cognitive complexity of items increase, regression models were conducted. The predictors are the item parts, depth of knowledge (DOK), and item achievement level descriptors (ALD) which are thought to represent the item's cognitive complexity. The dependent variable is the item difficulty of the operational items. The sample size is 30 per grade as the model includes the operational items from the 2023 administration.

Item part are Parts A, B, and C with A aligned to the least complex skills inherent in the extended standard, and C aligned to the most complex skills inherent in the extended standard. Item part is included in the model as a categorical variable as the item part concept does not give a sense of interval variable in the first model.

Depth of knowledge (DOK) ranges from 1 through 3 with 1 being the lowest DOK. Most items are DOK 1 or 2. DOK is included as a continuous variable in the model as they are expressed in number and give a sense of interval variable.

Item achievement level descriptors (ALD) are included as a continuous variable in the model as they are expressed in number and give a sense of interval variable. Each level has following definitions:

- Level 1:
  - A **limited** understanding of the knowledge and skills
  - **May need substantial academic support** as they transition to the next grade/course, inclusive postsecondary education, or competitive integrated employment.
- Level 2:
  - A **partial** understanding of the knowledge and skills
  - **May need frequent academic support**
- Level 3:
  - An **adequate** understanding of the knowledge and skills
  - **May need occasional academic support**
- Level 4:
  - A **thorough** understanding of the knowledge and skills
  - **May need limited academic support**

Item parts, DOK and ALD are closely related. Most items on this assessment, by design, are either DOK 1 or DOK 2 (as seen in Table 3). With grades 5 and 8, there is only one item aligned to DOK 3. Grade HS has two items with DOK 3. Table 4 shows the relationship between item parts and item ALD. Most of the Part A items have ALD 1, most Part B items have ALD 2 or ALD 3, and the Part C items have ALD 3 or ALD 4. Given that DOK and ALD are closely related to item parts, DOK and ALD are also related to each other (Table 5).

**Table 2. Science Item Parts and DOK**

	DOK								
	Grade 5			Grade 8			Grade HS		
	1	2	3	1	2	3	1	2	3
Item Part A	10	0	0	10	0	0	10	0	0
Item Part B	0	10	0	0	10	0	0	10	0
Item Part C	0	9	1	0	9	1	0	8	2

**Table 3. Science Item Parts and ALD**

	ALD											
	Grade 5				Grade 8				Grade HS			
	1	2	3	4	1	2	3	4	1	2	3	4
Item Part A	7	3	0	0	5	4	1	0	3	7	0	0
Item Part B	0	8	2	0	0	0	9	1	0	4	6	0
Item Part C	0	0	7	3	0	0	1	9	0	0	3	7

**Table 4. Science Item Parts DOK and ALD**

	ALD											
	Grade 5				Grade 8				Grade HS			
	1	2	3	4	1	2	3	4	1	2	3	4
DOK 1	7	3	0	0	5	4	1	0	3	7	0	0
DOK 2	0	8	9	2	0	0	10	9	0	4	9	5
DOK 3	0	0	0	1	0	0	0	1	0	0	0	2

**Regression Results**

Table 6 through Table 8 show the regression results with all predictors. Item part predictor is a significant predictor (p-value less than 0.05) in all grades. Part A is the reference variable to Part B and Part C predictors. Therefore, the slopes for Part B and Part C are the difference in item difficulty compared to Part A. Both Part B and Part C slopes are statistically significant, and their item difficulties are higher than Part A. These results further support that the designation of part level represents categories of distinct complexity levels, as designed.

The predictors of DOK or ALD are not significant in all grades. With grade 5, the DOK predictor’s p-value was larger than the ALD predictor. With grades 8 and HS, the ALD predictor had a higher p-value than the DOK predictor. With grade 5, the model explained 67% of the total variance of item difficulty when the item part predictor is a categorical variable. The grades 8 and HS models explained 68% and 85% of the total variance of item difficulty when the item

part predictor is a categorical variable. Since the item parts, DOK and ALD are highly related, the multicollinearity presents complexity in interpreting the coefficient with models with/without DOK and ALD. When the goal of the model is to improve the model correlation it is reasonable to include all highly correlated predictors in the model. With this study, the model results with the item parts, DOK and ALD in predictors are shown. While a methodology limitation, the relationship between item part complexity, DOK, and ALD, is inherent and intentional within the test design.

The analysis included only 30 items for each grade, thus the significance may be difficult to determine with this design and the sample size. For this reason, we recommend this study's results be interpreted as one point of validity evidence that the item parts within the GAA 2.0 science assessments call on the intended science cognitive skills at the varying complexity levels inherent in the extended content standards. These results supplement, but do not replace, the role of expert judgement in the annual item development processes for these assessments.

**Table 5. Grade 5 Item Difficulty Modeling**

	<b>Coefficient</b>	<b>SE</b>	<b>T-statistics</b>	<b>P-value</b>
<b>Intercept</b>	-0.3465	0.3498	-0.9907	0.33
<b>Item Part B</b>	0.8821	0.3567	2.4727	0.02
<b>Item Part C</b>	1.2022	0.4168	2.8843	0.01
<b>DOK</b>	-0.0428	0.3446	-0.1241	0.90
<b>ALD</b>	-0.1104	0.1358	-0.8133	0.42

**Table 6. Grade 8 Item Difficulty Modeling**

	<b>Coefficient</b>	<b>SE</b>	<b>T-statistics</b>	<b>P-value</b>
<b>Intercept</b>	-0.3373	0.3528	-0.9561	0.35
<b>Item Part B</b>	0.9441	0.3588	2.6317	0.01
<b>Item Part C</b>	1.0659	0.4281	2.4898	0.02
<b>DOK</b>	-0.1919	0.2971	-0.6460	0.52
<b>ALD</b>	-0.0088	0.1132	-0.0775	0.94

**Table 7. Grade 5 Item Difficulty Modeling**

	<b>Coefficient</b>	<b>SE</b>	<b>T-statistics</b>	<b>P-value</b>
<b>Intercept</b>	-0.4336	0.2094	-2.0711	0.05
<b>Item Part B</b>	1.1008	0.1950	5.6441	0.00
<b>Item Part C</b>	1.2065	0.2544	4.7431	0.00
<b>DOK</b>	-0.1860	0.1684	-1.1047	0.28
<b>ALD</b>	-0.0030	0.0829	-0.0356	0.97

### **Conclusion**

The results of this line of research provide consistent evidence that the tasks developed and administered for the GAA 2.0 science assessments tap the appropriate cognitive processes in increasing within task complexity as the test design and extended content standards require. These empirical results support, but do not supplant, expert judgment points embedded in the GAA 2.0 development cycles as critical validity checks.