Exploratory Analysis of the Knox County Gifted and Talented Screening Data

Technical Report

Clint Sattler
Supervisor, Research and Evaluation
Knox County Schools
Department of Research, Evaluation, and Assessment

November 2019

Overview

The Knox County School's (KCS) Gifted and Talented (GT) Program experimented with using different data sources during the 2018-2019 school year (SY1819) in order to identify GT candidates among second grade students. Potential students were administered Reading- and Math-based GT assessments in six pilot schools. The GT department also provided the results from a skills-based inventory (survey) completed by the students' homeroom teachers for a subset of the assessed second grade students. The KCS GT department requested an analysis of the data by the KCS Department of Research, Evaluation, and Assessment (REA) to help inform the future GT screening process.

There is little evidence to suggest that the skills-based inventory provided new information to make GT placement decisions when compared to the results of an external skills-based intervention screener (Aimsweb spring benchmark percentiles). A logistic regression analysis provides evidence that student performance on the Reading-and Math-based assessments were the only variables that significantly predicted GT enrollment. In the future, it may still be desirable to collect the skills-based survey data for a targeted group of students who are performing on the lower limit of the GT expectations.

The results of the analysis suggest that diversifying access to GT programming could be accomplished by adjusting the content on the Reading- and Math-based assessments. The assessments could be redesigned to better capture performance on non-cognitive skills (creativity, persistence, etc.) that help identify gifted students.

These findings were generated by data collected at a very limited and non-representative sample of KCS students. Skills-based inventory data may be more predictive of GT enrollment in a different sample of students. Readers are cautioned from extrapolating the findings of this study to other populations of students.

Methodology

Assessment and survey data collection was coordinated by the KCS GT department. Assessment data were collected at six KCS elementary schools (A.L. Lotts, Brickey-McCloud, Cedar Bluff, Hardin Valley, Northshore, and Sequoyah). Skills-based survey data were collected at a subset of these schools (Cedar Bluff, Hardin Valley, Sequoyah, and Northshore). Students were linked to Pearson's Aimsweb 1.0 skills-based assessment data by the concatenation of student name and school. The Aimsweb data were collected during the KCS spring benchmark test window in SY1718. National percentiles were extracted for Reading-Curriculum Based Measures (oral reading fluency, R-CBM) and Mathematics-Computation (M-COMP) assessments. Percentiles were converted to normal curve equivalents (NCEs) for use in analyses. Additionally, the GT department provided the enrollment data to determine which students were enrolled in the GT program as third grade students during SY1920.

Methodology: Student Skills Inventory

Teachers rated students on a nine-item inventory covering the following domains: Inquiry, Innovative, Analytical, Conceptual, Creative, Motivated, Interest, Communication, and Persistence. All items were rated on a 1 to 4 scale. Each domain was qualified with a short descriptor, but there were no exemplar anchors to guide teacher ratings.

It is assumed from the data that all teachers did not engage with the inventory scales in the same way. Some teachers chose to use the scale in a continuous manner, while others seemed to constrain themselves to an ordinal scale. All responses used in the analysis rounded to the nearest integer, while limiting responses to the intended 1 to 4 scale.

Item response theory (IRT) modeling was used to validate the inventory scales and to generate latent student ability estimates (theta, as Z scores) from the teacher responses. IRT modeling used a general partial credit model and the expectation-maximization algorithm. Analysis of variance (ANOVA) modeling was used to determine if the student ability estimates could be used as a GT screening tool. Screening models indicated that there was low school-to-school variance in the student skills inventory (interclass correlation coefficient (ICC) = 0.023), so hierarchical modeling was not required. Similarly, the between-school ICCs for R-CBM and M-COMP NCEs indicated that non-nested modeling was sufficient (R-CBM ICC= 0.035, M-COMP ICC=0.039).

Methodology: Content-Aligned Assessments

The KCS GT department created standards-aligned assessments to gauge student performance against the state curriculum in both Reading/Language Arts (RLA) and Mathematics. The RLA assessment consisted of a narrative fiction passage and a non-fiction descriptive passage. After reading the passage, students were asked to respond to short-answer prompts with citation from the text, in addition to producing drawings to summarize the passages. The maximum score on the RLA assessment was 19 raw score points.

The Mathematics assessment was largely composed of word problems requiring the students to explain their rationale for their answers. The Mathematics assessment also included a small number of items associated with foundational math skills (multiplication tables, etc.). The maximum score on the Mathematics assessment was 28 raw score points.

The GT department also identified three items on the Mathematics exam as exemplars for the type of problems students would encounter in the GT curriculum. The responses to these items constituted a GT math sub-score on the assessment. Questions 4, 8, and 12 from the Mathematics assessment were used to calculate the GT math sub-score. Two parameter IRT modeling was used to calculate a latent GT math subtest ability score for each participating student. The IRT parameters associated with each item were examined to determine the discrimination ability of each item. Additionally, ANOVA was used to determine if performance on these math tasks could serve as a suitable GT screening tool.

Methodology: GT Enrollment Regression Model

The skills inventory, Aimsweb, and assessment results were linked to GT enrollment data. Students could be enrolled in a content-specific GT program (math or RLA) or in a cross-curricular program. Any student enrolled in any GT program was coded as a GT-enrolled student in subsequent modeling. Modeling used a logistic linking function. Screening models indicated that the between-school ICC associated with GT enrollment was 0.090, thus the final model nested students within schools.

The data were modeled with a multilevel logistic regression model in which students were clustered under schools. There were an inadequate number of students clustered under individual teachers to create a viable teacher-level model. The covariates included in the logistic regression included the theta scores estimated from the skills inventory, the theta estimates from the math assessment sub-score, the raw results from both the Reading and Mathematics GT screening assessments, and the R-CBM and M-COMP spring normal curve equivalents. The equation used for the model is contained below, for each student I in each school j.

$$\Pr(y_i = 1) = logit^{-1}\big(\alpha_{j[i]} + \beta_{iventory} * \theta_{inventory} + \beta_{subtest} * \theta_{subtest} + \beta_{RLA} * RLA\ Raw\ Score + \beta_{Math} * Math\ Raw\ Score + \beta_{rbcm} * RCBM\ NCE + \beta_{mcomp} * MCOMP\ NCE\big)$$

$$\alpha_j \sim N(\mu_\alpha, \sigma^2_{school}),$$

All calculations were done on R version 3.6.1 running on R Studio version 1.2.1335. IRT calculations used the "mirt" package version 1.31. Multi-level regression was accomplished using the "lme4" package version 1.1-21. Data visualizations used the package "ggplot2" version 3.2.0.

Results

The data being analyzed in this study originated in the four schools in which the teacher survey data were collected: Cedar Bluff, Hardin Valley, Sequoyah, and Northshore Elementary Schools. The demographic data for the SY1819 second grade cohort in the pilot schools and all other KCS schools are contained in Table 1. The demographic data were aggregated from the school rosters as reported on September 6, 2018. Table 1 includes the percentage of students that were members of a state-identified at-risk subgroup: Black, Hispanic, or Native American students (BHN), Economically Disadvantaged students (ED), English Language Learners (ELL), and Students with Disabilities (SWD).

Table 1: Demographic Comparison of SY1819 Second Grade Cohorts

	BHN	ED	ELL	SWD
Pilot Schools	15.2%	9.5%	6.9%	9.5%
All Other KCS Schools	31.2%	27.3%	6.9%	11.3%

As evident from Table 1, the data used in this analysis did not come from a demographically representative sample of KCS second grade students. As such, the results from this analysis are not generalizable to the other schools in the district.

Results: Student Skills Inventory

Student skills inventory data were collected from 228 students and was scored by 27 homeroom teachers at 4 schools. The descriptive statistics for each item on the inventory are contained in Table 2.

Table 2: Student Inventory Item-level Statistics

	Inquiry	Innovative	Analytical	Conceptual	Creative	Motivated	Interest	Communicator	Persistence
Mean	3.4	3.2	3.3	3.4	3.2	3.3	3.3	3.2	3.2
Std. Dev.	0.7	0.8	0.8	0.7	0.8	0.9	0.7	0.8	0.8
Minimum	1	1	1	1	1	1	1	1	1
1st Quartile	3	3	3	3	3	3	3	3	3
Median	4	3	3	4	3	4	3	3	3
3rd Quartile	4	4	4	4	4	4	4	4	4
Maximum	4	4	4	4	4	4	4	4	4

Examination of the factor loadings associated with the items suggests that the data from the skills inventory are suitable for IRT modeling. The domains exhibit invariance, and in this case, the assumption is made that all items load on the latent factor of interest. IRT

discrimination parameter estimates (a values) are all acceptable. Generally, the item fit statistics indicate that the response patterns in the skills inventory can adequately be explained by the partial credit IRT model. The only exception may be the "Innovative" item. The factor loadings, IRT parameter estimates, and item fit statistics are contained in Table 3.

Table 3: Skills Inventory IRT Factor Loadings, Parameter Estimates, and Fit Statistics

		IRT Parameters				Fit Statistics		
	Factor Loading	a	b1	b2	b4	Chi Sq.	DF	p
Inquiry	0.847	2.71	-1.39	-0.09	0.56	41.85	29	0.058
Innovative	0.913	3.80	-0.83	0.25	0.84	48.20	27	**0.007***
Analytical	0.927	4.22	-0.93	0.20	0.72	19.46	24	0.727
Conceptual	0.93	4.31	-1.05	0.03	0.67	21.96	23	0.523
Creative	0.9	3.52	-0.93	0.20	0.86	26.40	26	0.441
Motivated	0.834	2.57	-0.97	0.22	0.61	40.38	30	0.098
Interest	0.934	4.45	-1.00	0.10	0.77	30.09	22	0.116
Communicator	0.907	3.67	-0.90	0.25	0.86	27.14	27	0.456
Persistence	0.782	2.14	-1.01	0.19	0.65	36.44	34	0.356

Examination of the item characteristic curves indicates that all of the intervals of the rating scale were used. Additionally, the probability of scoring higher on the scale increases with student theta estimates (monotonically). These findings suggest that the data collected through the skills inventory framework are internally valid.

The distribution of theta scores is contained in Figure 1. Each item in the skills inventory was associated with a "difficulty" of agreeing with a specific item. The more "difficult" items would be more "difficult" for the teachers to score a student at the top of the scale. Accordingly, the theta estimates for the students represent the latent "ability" of the student in relation to the average item "difficulty". The mean theta estimate for the students in the sample was -0.0011

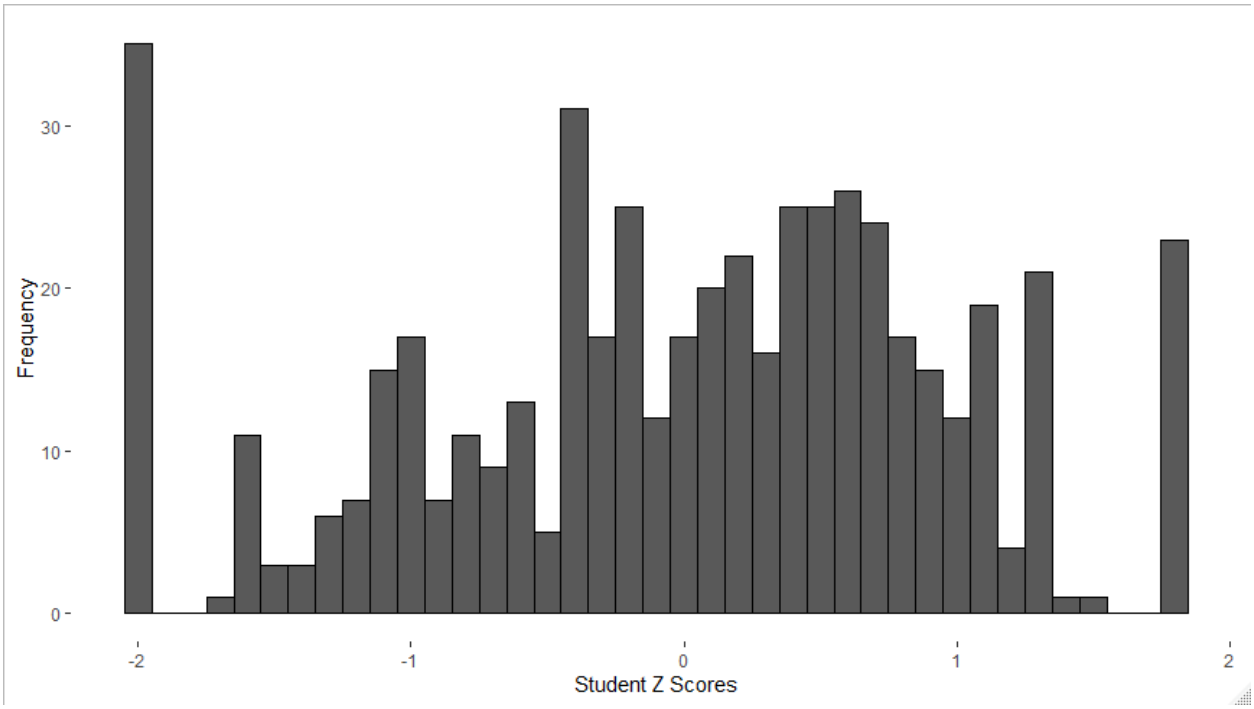

Figure 1:Histogram of Skill Inventory Theta Estimates

The test information curve (TIC) suggests that the response patterns in the skills inventory are best suited to discriminate students with average theta estimates (see Figure 2). If we assume that the objective of the skills inventory was to separate students with average theta estimates from students with high theta estimates, the peak of the TIC would need to shift towards higher theta values. In practical terms, this means that the inventory should likely include items that teachers were less inclined to score at the top of the rating range. For example, if the goal was to separate students with inventory scores in the top 10% from the other students in the sample, the ideal TIC would have a maximum value at a theta value of 1.28 (the Z score that corresponds to the 90th percentile).

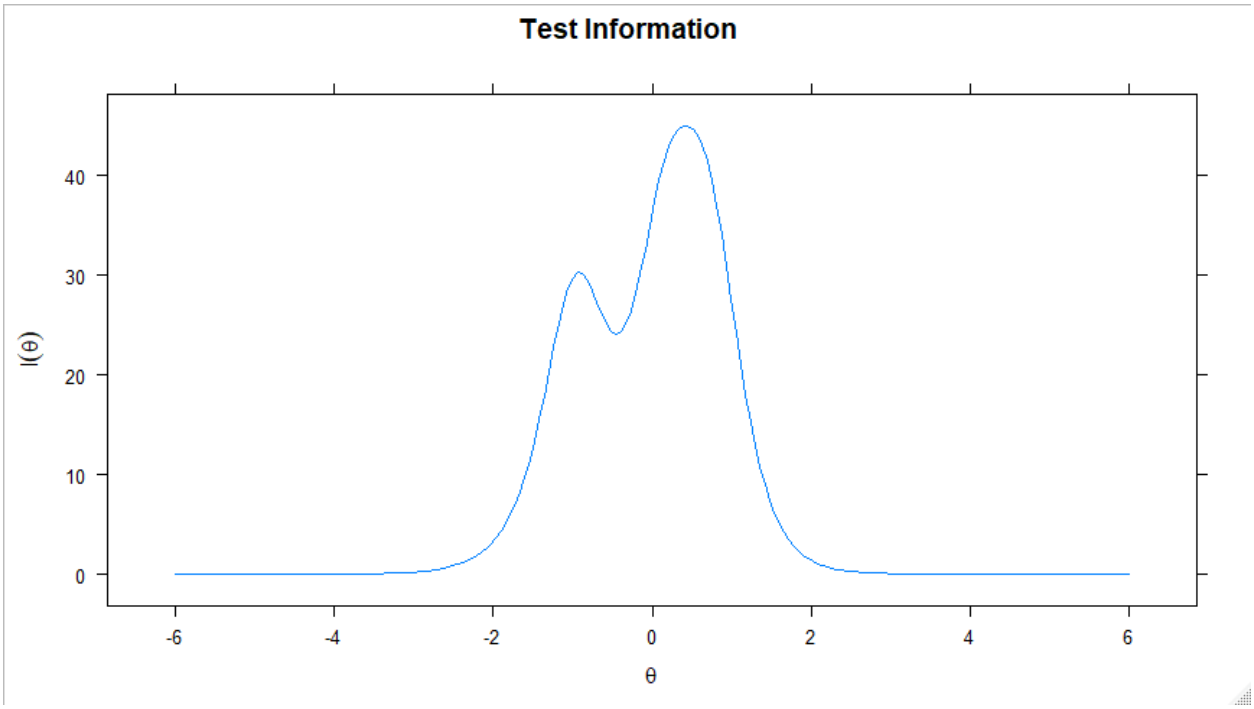

Figure 2: Skills Inventory Test Information Curve

Individual item information curves (IICs) were analyzed to determine which items contained the most information in the desired range. The area under the IIC was calculated between theta values of 0 and 10 to identify items best suited to discriminate between students with higher skills ability estimates those with average skills ability estimates. The "Analytical" and "Interest" items seem to contain the most information to make this distinction (Figures 3 and 4).

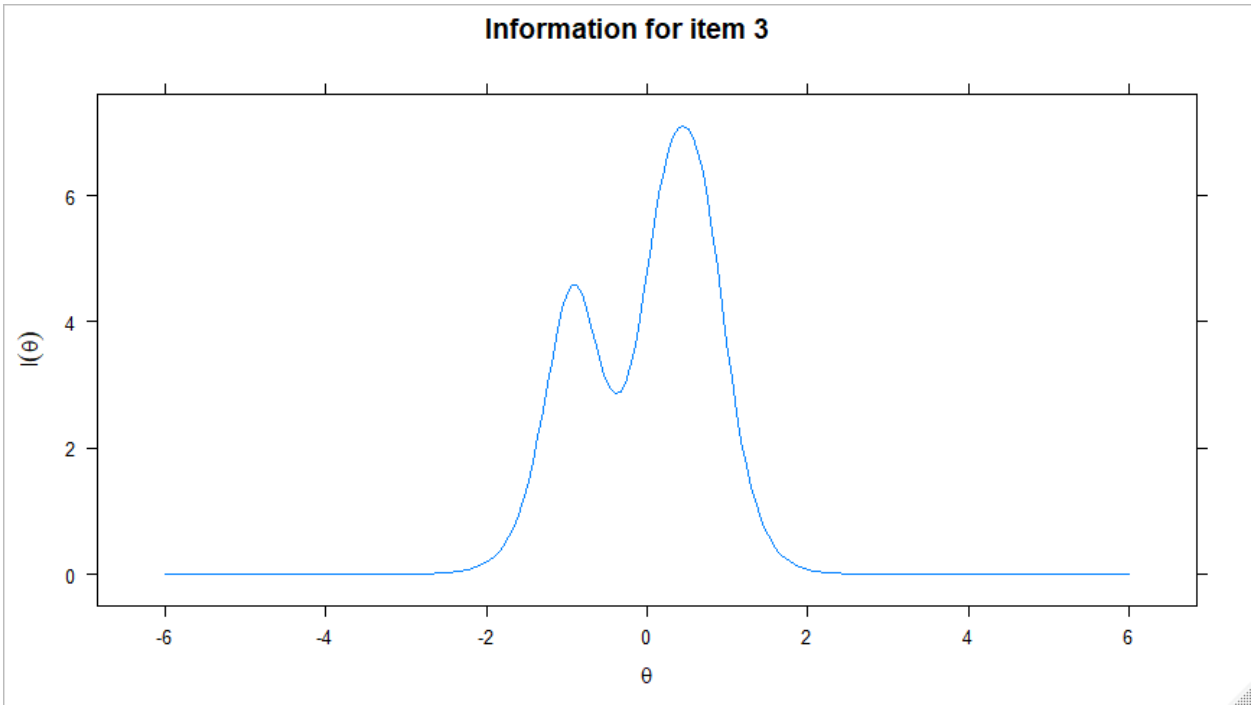

Figure 3: "Analytical" Item Information Curve

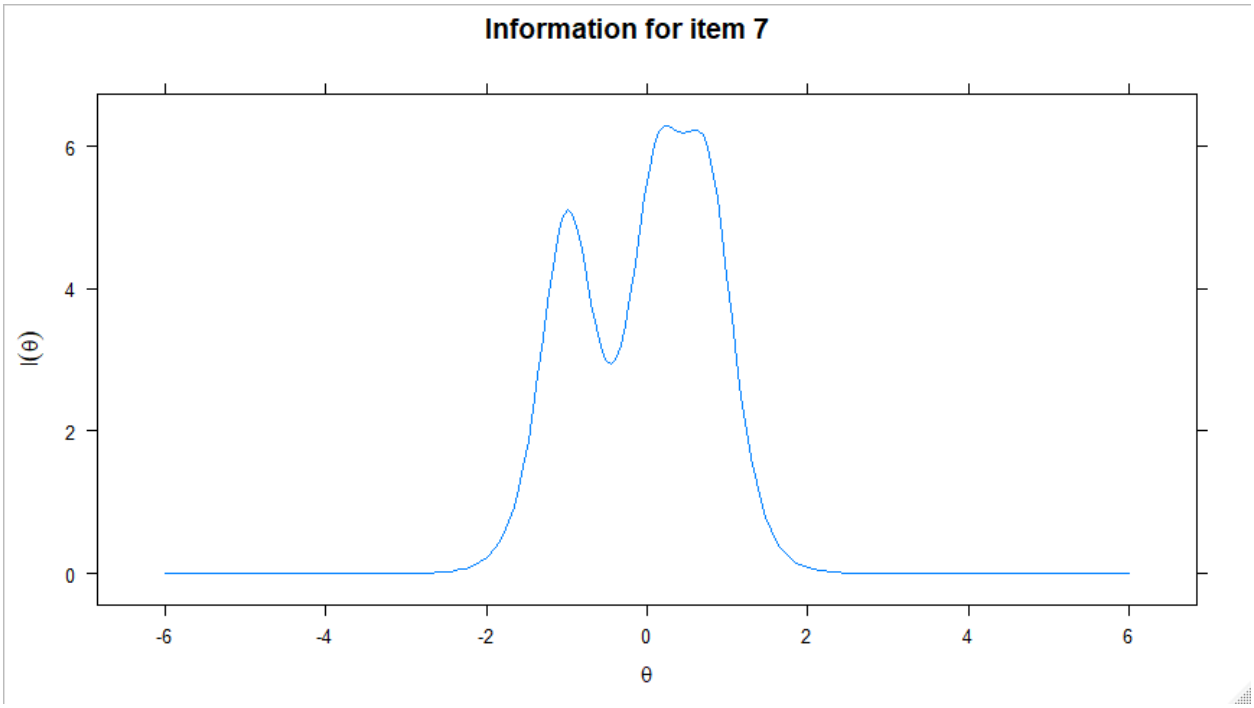

Figure 4: "Interest" Item Information Curve

Teachers were asked to rate students' performance as "above" grade level or "on" grade level in both Reading/Language Arts and Mathematics when they completed the skills inventory. Some teachers chose to write in additional categories, especially "below grade level", so all

students who were not labeled as "above" by their teachers were recoded as "not above grade level" for further analysis.

ANOVA testing was used to compare mean latent student ability estimates (measured by skills inventory) between students coded "above" and "not above" grade level for RLA. The ANOVA results indicate that we can reject the null hypothesis that there is no difference in mean latent student ability (as measured by the skills inventory) among students who were labeled as above grade level in RLA when compared to students who were recoded as not above grade level (N=194, F=48.84, p=4.38e-11). Readers should note that the sample suffered some attrition because some teachers did not complete the RLA grade-level ability portion of the skills inventory. Examination of Figure 5 indicates very little overlap in the distribution of skills inventory theta estimates among students classified as above grade level in RLA and students recoded as not above grade level in RLA.

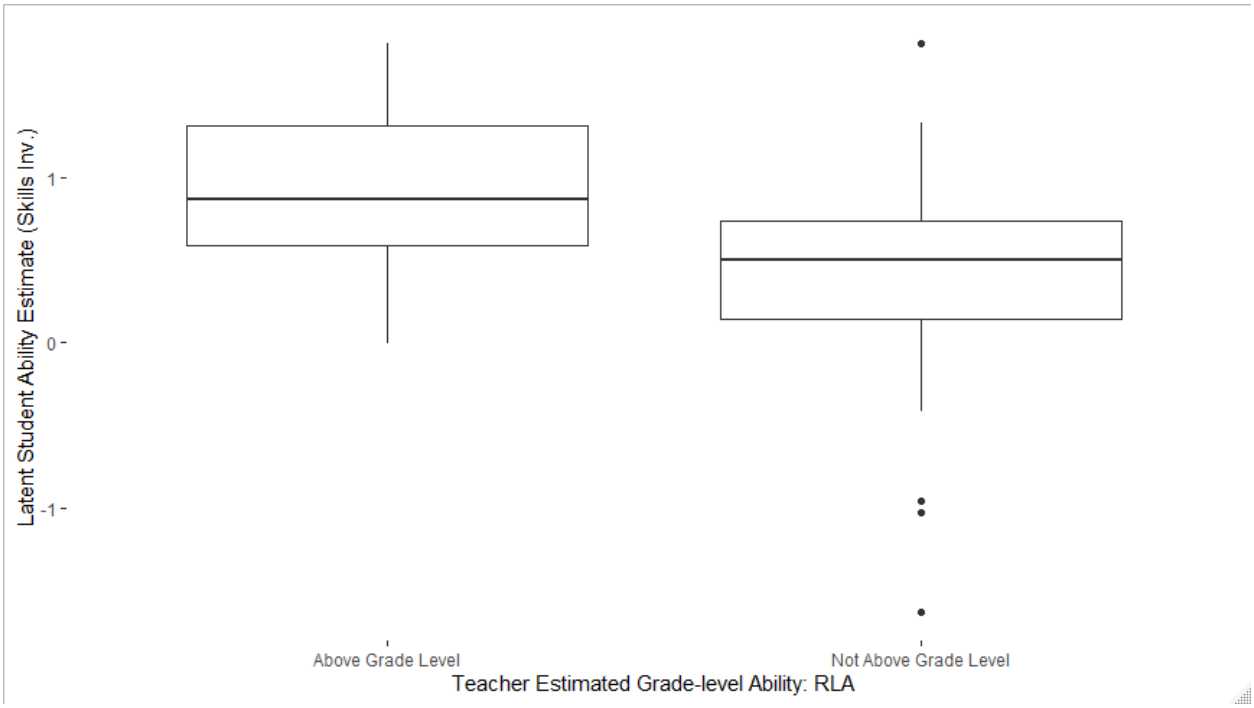

Figure 5: Boxplot of Skills Inventory Latent Ability by Teacher Grade Level Classification in RLA

ANOVA results indicate that we can reject the null hypothesis that there is no difference in mean latent student ability (as measured by the skills inventory) among students who were labeled as above grade level in math when compared to students who were recoded as not above grade level (N=194, F=32.76, p=3.93e-8). Examination of Figure 6 indicates very little overlap in the distribution of skills inventory theta estimates among students classified as above grade level in math and students recoded as not above grade level in math.

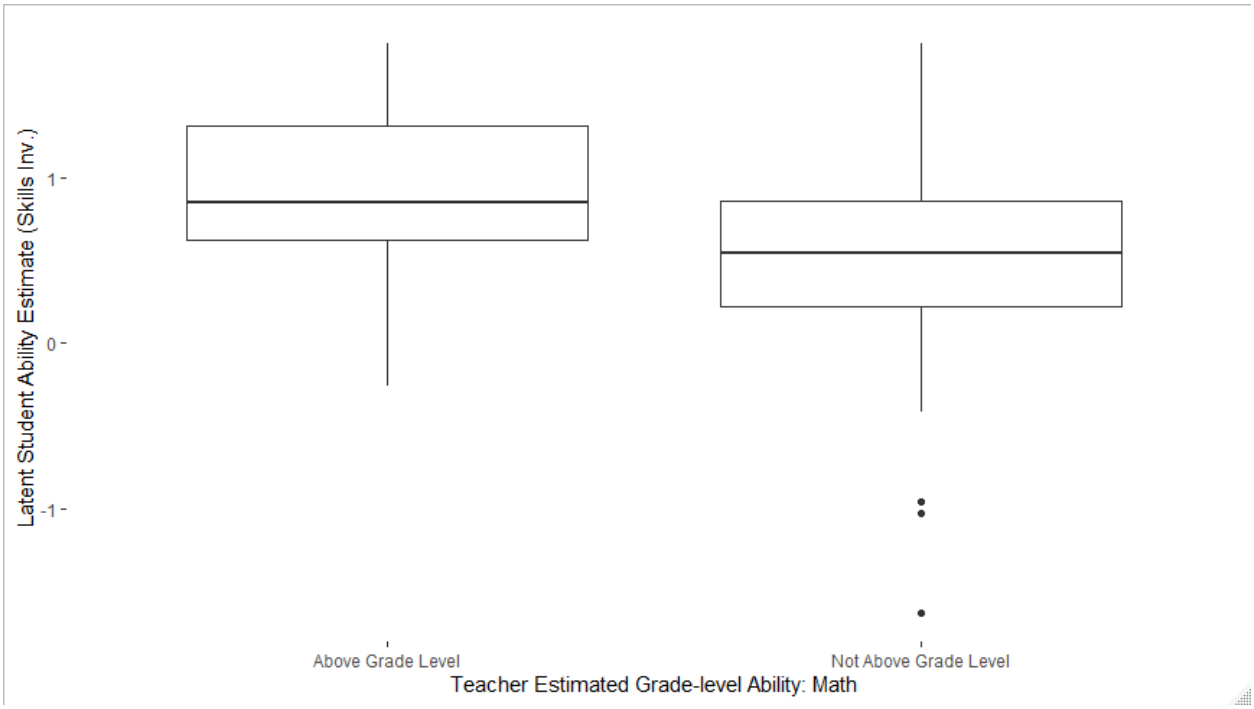

Figure 6: Boxplot of Skills Inventory Latent Ability by Teacher Grade Level Classification in Math

The spring benchmark Aimsweb data were used to check for concurrent validity between teacher grade-level ratings and Aimsweb performance (national NCE) in both RLA and math.

ANOVA results indicate that we can reject the null hypothesis that there is no difference in mean R-CBM NCE among students who were labeled as above grade level in RLA when compared to students who were recoded as not above grade level (N=194, F=57.67, p=1.3e-12). Examination of Figure 7 indicates more overlap between the distributions than was evident in Figure 5.

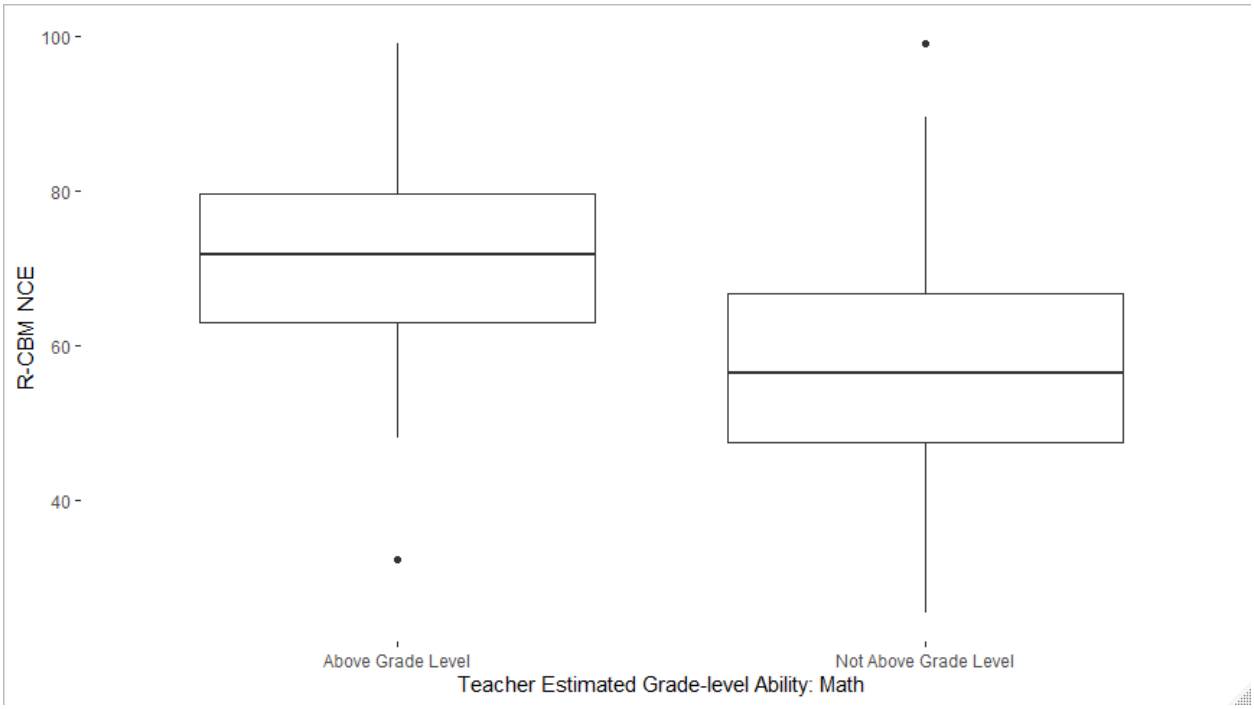

Figure 7: Boxplot of R-CBM NCE by Teacher Grade Level Classification in ELA

ANOVA results indicate that we can reject the null hypothesis that there is no difference in mean M-COMP NCE among students who were labeled as above grade level in math when compared to students who were recoded as not above grade level (N=194, F=24.97, p=1.3e-6). Examination of Figure 8 indicates more overlap between the distributions than was evident in Figure 6.

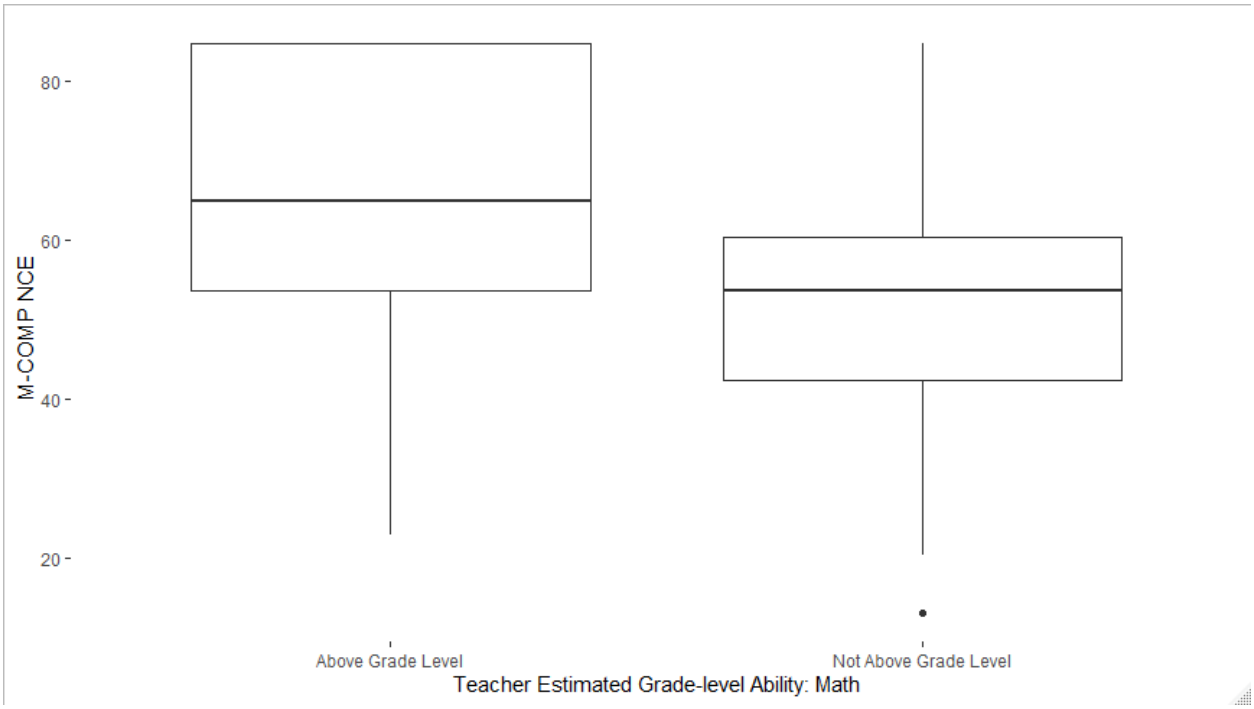

Figure 8: Boxplot of M-COMP NCE by Teacher Grade Level Classification in Math

Results: Content-Aligned Assessments

The distribution of the raw scores from the RLA and Math screening assessment is contained in Figures 9 and 10. The distribution of raw scores appears to follow a Gaussian distribution. The mean RLA raw score was 11.9 with a standard deviation of 3.1 (N=356). The mean math scaled score was 13.7 with a standard deviation of 5.7 (N=365).

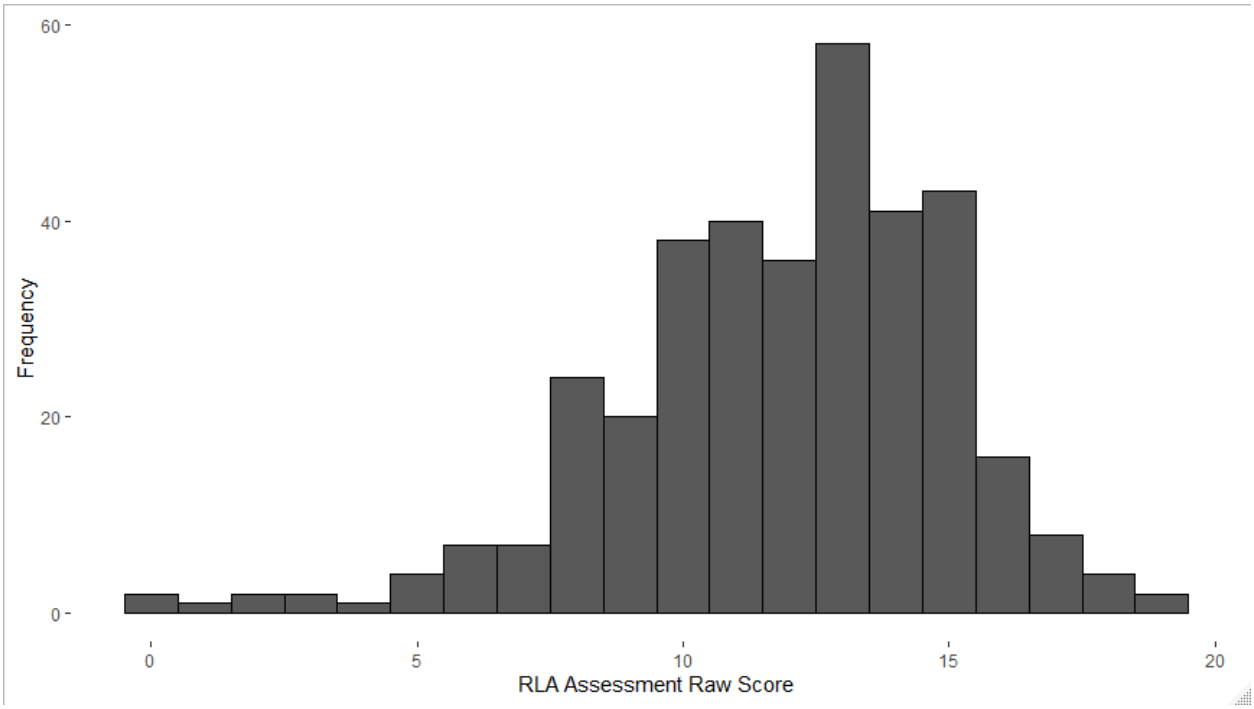

Figure 9: Distribution of RLA Assessment Raw Scores

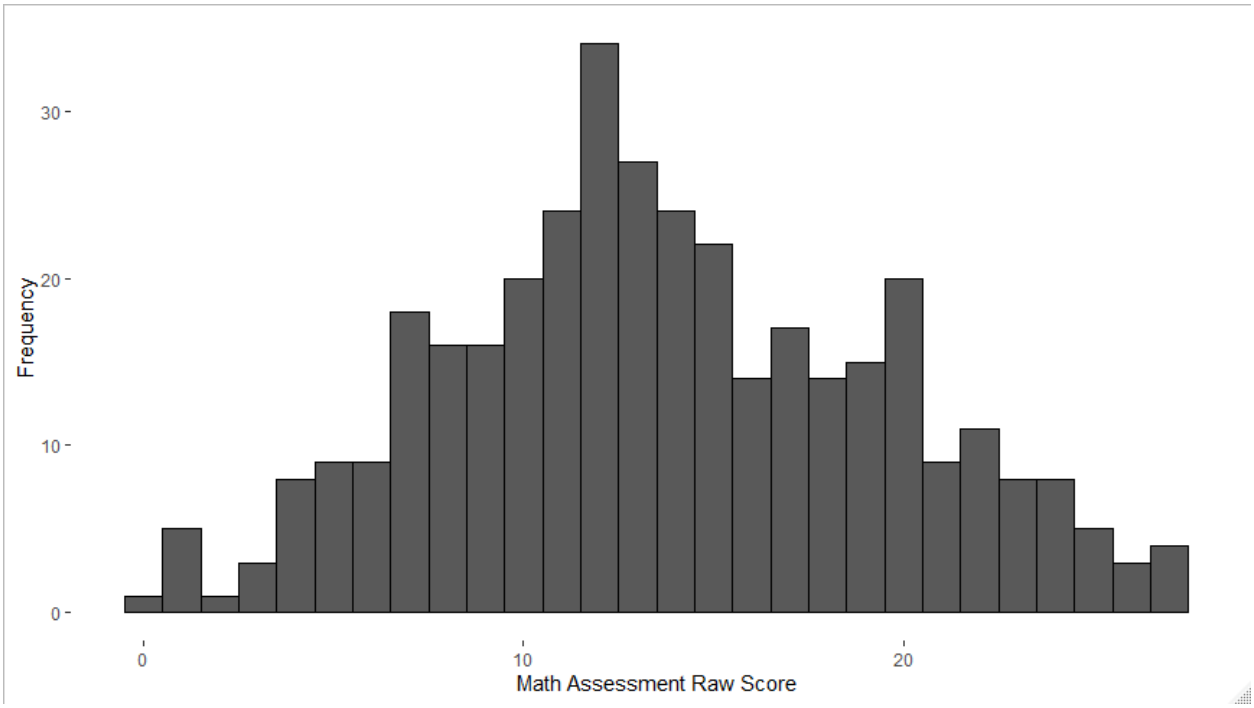

Figure 10: Distribution of Math Assessment Raw Scores

General linear modeling indicates that the raw RLA assessment score is significantly correlated with the Aimsweb R-CBM NCE (N=356, t=8.258, p =2.99e-15). By ANOVA testing, we reject the null hypothesis that there is no difference in the mean RLA assessment score by students identified as "above" grade level in RLA and students recoded as not above grade level (N=182, F=17.16, p=5.27e-5). General linear modeling indicates that we fail to reject the null hypothesis that the RLA assessment raw score is not correlated with the teacher survey ability estimate (N=213, t=1.505, p=0.134).

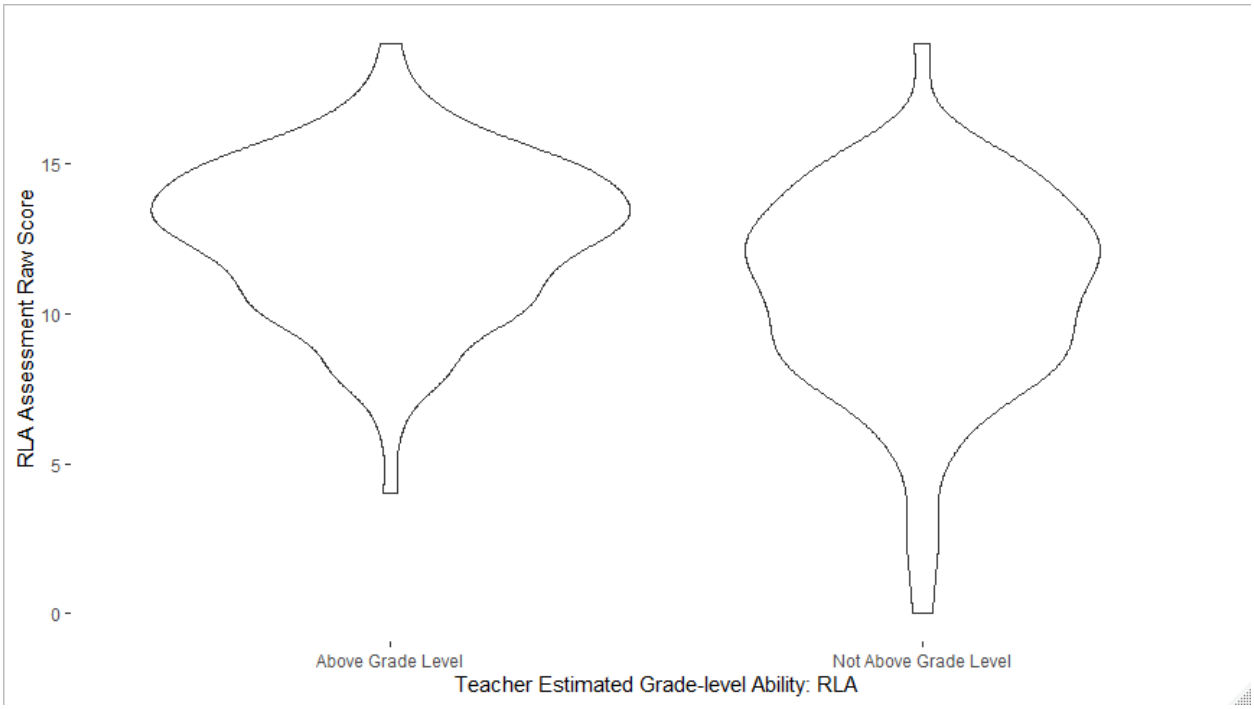

Figure 11: Violin Plot of RLA Assessment Raw Score by Teacher Grade Level Classification in RLA

General linear modeling indicates that the raw math assessment score is significantly correlated with the Aimsweb M-COMP NCE (N=365, t=6.348, p=6.48e-10). By ANOVA testing, we reject the null hypothesis that there is no difference in the mean math assessment score by students identified as "above" grade level in math and students recoded as not above grade level (N=189, F=12.18, p=6.04e-4). General linear modeling indicates that we reject the null hypothesis that the math assessment raw score is not correlated with the teacher survey ability estimate (N=220, t=4.441, p=1.42e-5).

Figure 12: Violin Plot of Math Assessment Raw Score by Teacher Grade Level Classification in Math

Results of an ANOVA test indicates that we fail to reject the null hypothesis that there is no relationship between the mean math assessment raw score and subtest score (N=365, F=11.17, p=2e-16). The Spearman correlation coefficient between the sum of subtest items marked correct and the total raw score on the math assessment was 0.50. This indicates that approximately 25% of the total variation in raw score can be explained by the total number of items correct on the subtest.

The results from the IRT analysis of the three-item math subtest are contained in Table 4. Question 4 exhibits low discrimination (a parameter) and seems poorly suited to differentiating student performance. Additionally, the factor loading of question 4 is low enough to raise concerns that scores related to question 4 do not reflect the same latent construct as scores on questions 8 and 12.

Table 4: Math Subtest Two-Parameter IRT model statistics

		IRT Parameters		Fit Statistics		
	Factor Loading	a	b	Chi Sq.	DF	p
Question 4	0.244	0.428	1.86	228.294	8	0.000
Question 8	0.625	1.362	0.214	139.726	8	0.000
Question 12	0.625	1.364	0.314	174.674	8	0.000

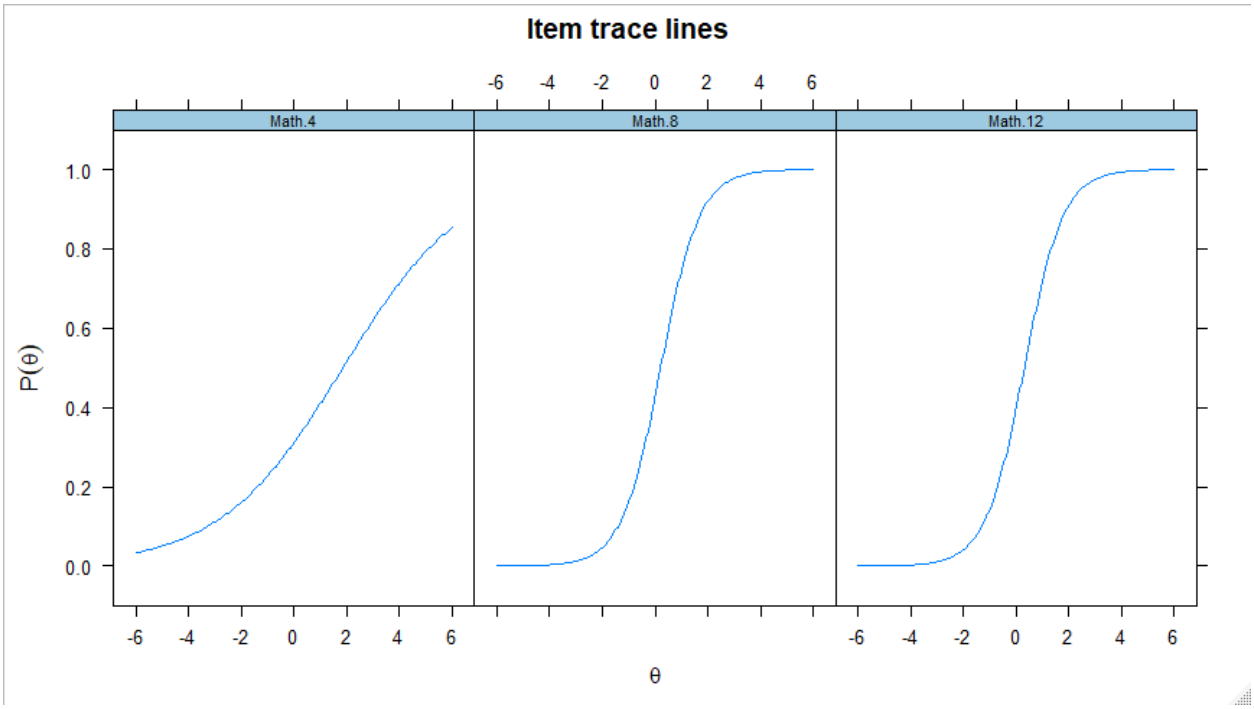

Figure 13: Math Subtest Item Characteristic Curves

The student level thetas (student ability as estimated by the three items on the math subtest) generated by the IRT analysis were saved for later regression modeling.

Results: GT Enrollment Regression Model

Comparison of the Akaike information criteria (AIC) between the null model and the proposed regression model suggests that the multi-level logistic regression model explains a significant amount of variance in the GT enrollment data (N=209, AIC_{null}=245.92, AIC_{model}=194.37, Chi Squared = 63.545, DF=6, p=8.54e-12). The fixed effects from the regression model are contained in Table 5. The by-group variance was 0.34.

Table 5: Logistic HLM Fixed Effects

Parameter	Estimate	Std. Error	z value	Pr(>|z|)
Intercept	-7.841	1.656	-4.734	**0.000***
$\beta_{inventory}$	0.308	0.369	0.836	0.403
$\beta_{subtest}$	0.239	0.430	0.556	0.578
β_{RLA}	0.172	0.082	2.109	**0.035***
β_{Math}	0.181	0.049	3.717	**0.000***
β_{rcbm}	0.022	0.015	1.481	0.139
β_{mcomp}	0.006	0.012	0.473	0.636

The parameter estimates associated with both the RLA and Math assessments are the only covariates that are significantly correlated to the probability of a student enrolling in the KCS GT program.

Conclusions & Considerations

There is evidence that the data collected to help inform GT enrollment decisions among the current cohort of third grade students in four KCS pilot schools are highly correlated and therefore unlikely to add substantially different information to the decision-making process. The aggregate "ability" of the students on the skills inventory, the estimation of grade-level ability, and Aimsweb 1.0 results are significantly correlated at the $\alpha=0.05$ level. This is important to note since the completion of the skills inventory requires more than a trivial amount of teacher time.

It is possible that the correlation among the variables is related to data bias. For example, teachers could have biased their grade-level ratings (above grade level or not above grade level) based on known Aimsweb results or assigned skills inventory scores to reflect their grade-level ratings based on their responses to the skills inventory. Without qualitative interviews, it is impossible to determine which data can be discarded as redundant information and which data are the key variable to classify student performance. Identification of the key variable would help inform future data collection policy.

The available data provides evidence that the GT placement decision correlates heavily with performance on the GT content-based assessments. None of the other variables investigated in this study correlate with GT enrollment at a statistically significant level. This analysis suggests that the performance on the math subset was not particularly predictive of GT enrollment.

The findings suggest that the district may want to adjust the GT screening assessments to better measure the cognitive and non-cognitive skills that differentiate a GT student from a non-GT student. The current assessment provides the most information to separate students who perform in the "average" range on the content-based assessments. It may be desirable to retool the assessment to provide more precise information to separate high performing students from average performing students. It is also possible that a criterion-referenced test would better serve as a GT screening assessment in which item difficulty is calibrated to the minimum expectation for a GT student.

The above conclusions may only apply to a pool of students with similar demographics as the pilot schools. It is possible that the results of the skills inventory would play a more prominent role in GT placements in a pool of schools with different demographics. Readers are cautioned not to generalize the findings and conclusions of this study to a group of students with different demographics.