2023–2024 STAAR Through-year Assessment Pilot (TTAP) Technical Report

Cambium Assessment, Inc.

Table of Contents

List of Tables

List of Figures

1. Introduction

The State of Texas Assessments of Academic Readiness (STAAR®) Through-year Assessment Pilot (TTAP) tests represent an innovative, through-year assessment model designed as a potential alternative to the STAAR summative tests. In the context of through-year assessments, this model serves as a progress monitoring system, offering students multiple opportunities throughout the academic year to demonstrate their mastery of standards. It also contributes to the prediction of their summative performance level reported at the end of the year.

TTAP was developed through close collaboration with Texas educators, administrators, students, and families. The progress monitoring system incorporates three distinct, short, testing opportunities held during the fall, winter, and spring. To ensure that all school districts can maintain their local curriculum, each TTAP progress monitoring opportunity covers the full scope of the curriculum. These opportunities use a multi-stage adaptive design, enabling shorter tests with enhanced accuracy to minimize disruptions to instructional time.

TTAP is a multi-year, fully online pilot program that was initiated in the 2022–2023 school year. The model is being piloted over several years to assess its benefits and to ensure that the design maintains the rigorous level of validity and reliability that STAAR currently meets. The ultimate goal is to establish a scoring methodology that is comparable to STAAR and suitable for state accountability. Participation in TTAP is optional and does not negate a campus's obligation to administer STAAR. For additional details about the STAAR TTAP assessments, please refer to the STAAR TTAP Assessments webpage[1].

A legislative report was produced in 2023 to summarize results from the first pilot in school year 2022–2023 (Year 1). This technical report provides comprehensive information about the 2023–2024 TTAP Assessments (Year 2), focusing on seven essential aspects. It covers the TTAP test design, administration, and participation; details student scores and performance level distributions; examines student growth across opportunities; assesses the reliability, validity, and fairness of the TTAP assessments, and introduces the special studies conducted in 2023–2024 that shape TTAP design and reporting decisions. Specifically, this report includes an overview of the following seven key aspects:

1) **Test Design, Administration, and Participation.** This section provides an overview of the intended use and purpose of the TTAP assessments, the assessment design, and details involved in the administration of the assessments, such as the testing windows and the number of administrations by test title and opportunity. This section also delves into the test participation data at the student, campus, and district levels and the demographics of the students involved.

[1] https://tea.texas.gov/student-assessment/assessment-initiatives/texas-through-year-assessment-pilot

2) **TTAP Scores from 2023–2024.** This section summarizes performance patterns in students' scale scores, performance levels, percentage correct scores by reporting category and item difficulty level, and their growth trends across multiple assessment opportunities.
3) **Reliability.** This section discusses the internal test reliability of the TTAP assessments.
4) **Validity.** This section provides criterion validity evidence that is reflected by the correlations between TTAP and STAAR summative scores.
5) **Fairness.** This section summarizes differential item functioning (DIF) analysis and item bias review procedures.
6) **Reporting.** This section provides an introduction about the TTAP reports at both the student level and the aggregated campus and district levels.
7) **Continuous Research and Improvement Plans.** This section summarizes TTAP special studies conducted in year 2023–2024. The objectives and key findings of each study will be reviewed to guide the future design and implementation of TTAP.

1.1 TTAP Intended Uses and Purposes

To guide the design and development of TTAP, the Texas Education Agency (TEA) and its vendors employ theories of action (TOAs) to establish connections between intended users and the fundamental challenges that assessment usage aim to address. The assessment stands as a critical component of this solution, with valid test score interpretation and utilization being critical outcomes.

The TEA's TOA envisions multiple short-term and long-term outcomes for the through-year testing program. It hypothesizes that TTAP will

- improve educator understanding of the relationship between instruction and assessment;
- improve student testing experience; and
- increase long-term learning of students.

These outcomes theoretically will result from the following actions:

- Students will take greater ownership of their learning.
- Educators will identify students in need of intervention.
- Administrators will provide better support to educators.

These outcomes may be made possible because the through-year assessments have been designed to be minimally disruptive to instruction (ranging from 50% to 75% of typical summative test length); they are 100% Texas Essential Knowledge and Skills (TEKS)-aligned; and they provide progress monitoring feedback. A cumulative scoring model in which each of the three shorter assessments contributes to a summative determination of student proficiency creates what may be considered three mid-stakes assessments. Consequently, TTAP has the

potential to furnish teachers with monitoring feedback for their instruction, enhance students' testing experiences, and promote long-term learning throughout the year.

1.2 Test Design and Item Development

TTAP tests follow a multistage test design. Multistage test design offers several advantages, including enhanced measurement precision through adaptive testing, efficient use of testing time by targeting areas of a test taker's ability, and reduced test anxiety by presenting appropriately challenging items. Such tests provide a customized assessment experience that matches individual abilities and ensure comprehensive coverage of content domains by strategically selecting items from a large item pool. Overall, multistage tests offer an accurate, efficient, and personalized assessment experience, leading to reliable and valid results with tests that are shorter than traditional fixed form assessments.

In a multistage test, forms within a stage are designed at varying difficulty levels (i.e., low, medium, or high) to adapt to students' abilities. This adaptive approach allows the test to measure students more accurately with a wide range of abilities. Test developers create these forms by calculating the average item difficulty within each form. For instance, in grade 6 mathematics, the average item difficulty for low, medium, and high forms is approximately -1.0, 0.0, and 1.0, respectively. These difficulty levels ensure that students encounter test items that are appropriately challenging based on their ability. This method helps in providing a more personalized assessment experience, improving the precision of the measurement across different ability levels.

Each TTAP test has three opportunities administered in the 2023–2024 school year. Each opportunity is a multistage assessment with two panels (stages). The multistage adaptive test is depicted in Figure 1. At Opportunity I (hereafter referred to as Opp. I), the students take a router form and then are routed to a form at the correct level of difficulty. In Opportunity II or III (hereafter referred to as Opp. II and Opp. III), if the opportunity is the first for the student, they will take the medium form as the router form. For a student who has tested in a prior opportunity, Opp. II and Opp. III start the student on the low, medium, or high form, based on their final ability from the most recent, previous opportunity completed and the routing rule to a specific form.

For the item development and review, Pearson takes on the major role for TTAP item development, with TEA personnel involved throughout the item development process. For a comprehensive overview of the item development process, readers can consult the Item Development and Review section of Chapter 2 in the STAAR Technical Digest[2].

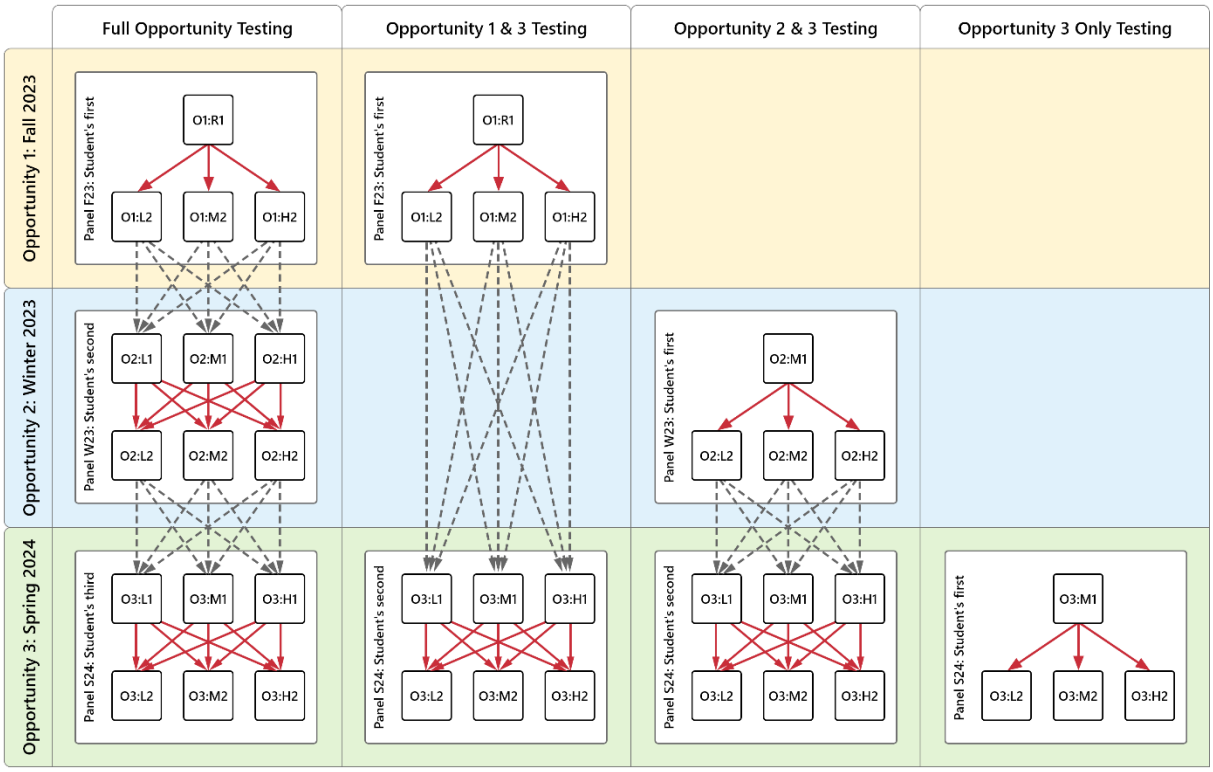
Items are classified into low, medium, and high difficulty levels based on item parameters. These difficulty categories are then used in student-, campus-, and district-level reports to provide

[2] https://tea.texas.gov/student-assessment/testing/student-assessment-overview/2023-tech-digest.pdf

detailed insights into performance. The classification is determined using item response theory (IRT) and is based on the ability to have at least a 2/3 (0.67) likelihood of success on items. For dichotomous items, this means a 2/3 likelihood of achieving a score of 1. For polytomous 2-point items, it refers to a 2/3 likelihood of achieving a score of 2. The classification thresholds are as follows:

- **Low Difficulty.** Items are classified as low difficulty if the ability required to achieve 67% correctness is lower than the meet performance level cut.
- **Medium Difficulty**. Items are classified as medium difficulty if the ability required is greater than or equal to the meet cut but lower than the master cut.
- **High Difficulty.** Items are classified as high difficulty if the ability required is higher than the master cut.

Figure 1: TTAP Design



O1: Opp. I, O2: Opp. II, O3: Opp. III
R1: Router Segment 1, L1: Low Segment 1, M1: Medium Segment 1, H1: High Segment 1
L2: Low Segment 2, M2: Medium Segment 2, H2: High Segment 2

Appendix A presents the test information function (TIF) curves of the test forms in each content-area and grade-level TTAP assessment in relationship to the corresponding STAAR Approaches, Meets, and Masters Grade Level performance cut scores.

1.3 Blueprints

TTAP test forms are constructed by Pearson based on criteria detailed in their Test Construction Specifications, and blueprints that represent proportionally shortened versions of the STAAR summative assessment. Table 1 compares the number of items on the TTAP and STAAR summative assessments (RC = reporting category), and Table 2 lists the names of the RCs.

Table 1: Comparison Between STAAR Summative and TTAP Blueprints

Assessment	Test	RC1	RC2	RC3	RC4	Total Items
Grade 5 Science	STAAR Redesign	4–6	6–8	8–10	10–12	32
	Through-year OP1	3	4	4	6	17
	Through-year OP2	3	4	4	6	17
	Through-year OP3	4	6	8	10	28
	Through-year Total Counts	10	14	16	22	62
Grade 6 Mathematics	STAAR Redesign	8–10	13–15	5–7	6–8	36
	Through-year OP1	5	7	3	5	20
	Through-year OP2	5	7	3	5	20
	Through-year OP3	7	11	6	6	30
	Through-year Total Counts	17	25	12	16	70
Grade 7 Mathematics	STAAR Redesign	4–6	14–16	11–13	5–7	38
	Through-year OP1	3	7	6	4	20
	Through-year OP2	3	7	6	4	20
	Through-year OP3	4	13	10	5	32
	Through-Year Total Counts	10	27	22	13	72
Grade 8 Social Studies	STAAR Redesign	15–17	8–10	8–10	5–7	40
	Through-year OP1	8	4	5	3	20
	Through-year OP2	8	4	5	3	20
	Through-year OP3	13	8	8	5	34
	Through-year Total Counts	29	16	18	11	74

Table 2: Summary of Reporting Categories

Assessment	RC1	RC2	RC3	RC4
Grade 5 Science	Matter and Energy	Force, Motion, and Energy	Earth and Space	Organisms and Environments
Grade 6 Math	Numerical Representations and Relationships	Computations and Algebraic Relationships	Geometry and Measurement	Data Analysis and Personal Financial Literacy
Grade 7 Math	Probability and Numerical Representations	Computations and Algebraic Relationships	Geometry and Measurement	Data Analysis and Personal Financial Literacy
Grade 8 Social Studies	History	Geography and Culture	Government and Citizenship	Economics, Science, Technology, and Society

1.4 2023–2024 TTAP Administration

The 2023–2024 TTAP assessments include three test opportunities. Table 3 represents TTAP assessment scopes and administration schedules.

Table 3: 2023–2024 STAAR TTAP Assessment Administration Schedules

Content	Language	Grade	Opp. I	Opp. II	Opp. III
Science	Spanish	5	November 6–13, 2023	January 29–February 2, 2024	March 25–29, 2024
Science	English	5			
Mathematics	English	6			
Mathematics	English	7			
Social Studies	English	8			

In the 2023–2024 school year, more than 5,3000 TTAP assessments were administered. Table 4 provides insight into the number of students who participated in each opportunity for each TTAP test. Additionally, the two rightmost columns present the count of students who completed all three opportunities of a TTAP test and those who took at least one opportunity of a TTAP test. The numbers in Table 4 reflect sample sizes following the application of exclusion rules, which help exclude test cases like off-grade test takers and students who did not meet attemptedness rules. A comprehensive list of these exclusion rules can be found in Appendix B. It is worth noting that the number of students who took grade 5 science Spanish version is relatively small, which could potentially limit the interpretability of results. In contrast, the other four tests all have sample sizes exceeding 6,800, ensuring that meaningful results can be derived from the data.

Table 4: TTAP Assessments Administered in the 2023–2024 School Year

Assessment	Opp I (N)	Opp II (N)	Opp III (N)	Total N Took All Three Opps	Total N Took at Least One Opp
Grade 5 Science (Spanish)	284	344	389	261	409
Grade 5 Science (English)	15,000	14,979	15,229	13,738	15,968
Grade 6 Mathematics (English)	7,890	8,011	8,046	7,278	8,492
Grade 7 Mathematics (English)	6,355	6,369	6,363	5,712	6,832
Grade 8 Social Studies (English)	20,074	19,774	20,072	17,951	21,421
Total	49,603	49,477	50,099	44,940	53,122

1.5 Test Participation

Table 5 provides additional insight into the counts of districts, campuses, and students who engaged in at least one TTAP assessment during the 2023–2024 school year. In this period, a total of 93 school districts, 315 campuses, and 53,122 students participated in TTAP administrations, which highlights the extensive reach of the TTAP assessments.

Table 5: TTAP District, Campus, and Unique Student Participation for Each TTAP Assessment in 2023–2024

Assessment	Number of Districts	Number of Campuses	Number of Unique Students
Grade 5 Science (Spanish)	28	66	409
Grade 5 Science (English)	72	200	15,968
Grade 6 Mathematics (English)	56	87	8,492
Grade 7 Mathematics (English)	55	73	6,832
Grade 8 Social Studies (English)	76	135	21,421
Total	93	315	53,122

In addition, the demographic characteristics of the 2023–2024 TTAP assessment participants have been compared with the State's student population in the same year to evaluate the sample representativeness of TTAP participants. Summarized demographic data for all students who took the STAAR summative tests in spring 2024 and those who participated in at least one TTAP assessment are presented in Table 6 through Table 10. For ease of reference in our analyses, the variable names and mapping can be found in Appendices C and D.

There are some notable demographic differences between the students who took TTAP grade 5 Spanish Science and those who took the STAAR grade 5 Spanish Science. However, it is important to acknowledge that the TTAP sample size for grade 5 science in Spanish is relatively small, which limits the significance of direct comparisons. For the other four assessments, in

most demographic comparisons, the percentages within each category exhibit striking similarities, with differences generally below 5%. However, there are a few exceptions to this trend. All percentage differences exceeding 5% were highlighted in bold within the tables. Note that the demographic characteristics are not exhaustive so the values may not add up to 100%. When analyzing the other tests in comparison to their respective state student populations, the following trends are noticed:

1. There is a slightly higher representation of white students and slightly lower representation of Black or African American students in the grade 5 Science TTAP sample compared to those who took STAAR.
2. There are slightly lower percentages of current limited English proficiency students (grade 5, 6, 7), economically disadvantaged students (grade 5), and at-risk students (grade 5) in the TTAP samples.
3. There is a slightly higher percentage of Title I, Part A Participants in the grade 7 and 8 TTAP samples.

These observations provide valuable insights into the demographic composition of TTAP assessment participants in relation to the broader student population, even though the differences are generally small. While we do observe a few variations, it is important to emphasize that these differences are generally minor and should not be overemphasized.

Table 6: TTAP Participating Student Demographic Characteristics (Spanish Grade 5 Science)

Demographic	STAAR Spring 2024	TTAP 2023–2024	Difference in Percentage
Number of Students	13,036	409	NA
Male	50.2	55.5	**5.3**
Female	49.7	44.5	**5.2**
Hispanic/Latino	96.4	97.6	1.2
American Indian or Alaska Native	0.4	0.2	0.2
Asian	0.0	0.0	0.0
Black or African American	0.1	0.0	0.1
Native Hawaiian or Pacific Islander	0.0	0.0	0.0
White	1.2	0.5	0.7
Two or More Races	0.2	0.0	0.2
Economically Disadvantaged	82.7	76.3	**6.4**
Title I, Part A Participants	91.6	95.6	4.0
Migrant	0.5	0.7	0.2
Current Limited English Proficient	98.2	99.3	1.1
Bilingual	78.6	81.2	2.6
ESL Participants	4.4	7.1	2.7
Special Education	7.4	7.6	0.2
Gifted/Talented Participants	4.2	0.2	4.0
At-Risk	88.5	93.6	**5.1**

Table 7: TTAP Participating Student Demographic Characteristics (English Grade 5 Science)

Demographic	STAAR Spring 2024	TTAP 2023–2024	Difference in Percentage
Number of Students	380,977	15,968	NA
Male	50.9	50.3	0.6
Female	49.1	49.7	0.6
Hispanic/Latino	51.0	47.1	3.9
American Indian or Alaska Native	0.3	0.2	0.1
Asian	6.0	6.9	0.9
Black or African American	12.8	7.8	**5.0**
Native Hawaiian or Pacific Islander	0.2	0.1	0.1
White	26.0	34.3	**8.3**
Two or More Races	3.2	3.2	0.0
Economically Disadvantaged	61.2	55.3	**5.9**
Title I, Part A Participants	71.9	72.7	0.8
Migrant	0.3	0.3	0.0
Current Limited English Proficient	24.6	18.6	**6.0**
Bilingual	11.8	7.6	4.2
ESL Participants	7.2	7.3	0.1
Special Education	16.7	17.7	1.0
Gifted/Talented Participants	11.7	12.5	0.8
At-Risk	49.2	44.2	**5.0**

Table 8: TTAP Participating Student Demographic Characteristics (English Grade 6 Mathematics)

Demographic	STAAR Spring 2024	TTAP 2023–2024	Difference in Percentage
Number of Students	384,178	8,492	NA
Male	50.8	50.1	0.7
Female	49.2	49.9	0.7
Hispanic/Latino	53.1	50.4	2.7
American Indian or Alaska Native	0.3	0.5	0.2
Asian	5.2	6.7	1.5
Black or African American	12.6	9.3	3.3
Native Hawaiian or Pacific Islander	0.2	0.1	0.1
White	25.0	29.7	4.7
Two or More Races	3.0	3.1	0.1
Economically Disadvantaged	62.4	58.2	4.2
Title I, Part A Participants	65.3	63.0	2.3
Migrant	0.3	0.6	0.3
Current Limited English Proficient	27.0	20.3	**6.7**
Bilingual	3.1	5.0	1.9
ESL Participants	17.2	12.7	4.5
Special Education	15.0	15.5	0.5
Gifted/Talented Participants	10.4	8.9	1.5
At-Risk	52.9	48.0	4.9

Table 9: TTAP Participating Student Demographic Characteristics (English Grade 7 Mathematics)

Demographic	STAAR Spring 2024	TTAP 2023–2024	Difference in Percentage
Number of Students	317,638	6,832	NA
Male	50.7	50.3	0.4
Female	49.3	49.7	0.4
Hispanic/Latino	54.9	57.4	2.5
American Indian or Alaska Native	0.3	0.6	0.3
Asian	4.1	1.9	2.2
Black or African American	13.2	10.5	2.7
Native Hawaiian or Pacific Islander	0.2	0.1	0.1
White	23.7	26.3	2.6
Two or More Races	2.8	2.9	0.1
Economically Disadvantaged	64.9	65.7	0.8
Title I, Part A Participants	63.4	70.9	**7.5**
Migrant	0.3	0.6	0.3
Current Limited English Proficient	27.4	21.9	**5.5**
Bilingual	0.9	0.1	0.8
ESL Participants	19.7	18.1	1.6
Special Education	15.5	16.2	0.7
Gifted/Talented Participants	6.9	6.4	0.5
At-Risk	58.6	55.7	2.9

Table 10: TTAP Participating Student Demographic Characteristics (English Grade 8 Social Studies)

Demographic	STAAR Spring 2024	TTAP 2023–2024	Difference in Percentage
Number of Students	405,749	21,421	NA
Male	51.2	51.0	0.2
Female	48.8	49.0	0.2
Hispanic/Latino	52.9	57.2	4.3
American Indian or Alaska Native	0.3	0.3	0.0
Asian	5.4	4.7	0.7
Black or African American	12.4	8.4	4.0
Native Hawaiian or Pacific Islander	0.2	0.1	0.1
White	25.2	26.3	1.1
Two or More Races	2.9	2.7	0.2
Economically Disadvantaged	60.3	58.8	1.5
Title I, Part A Participants	60.9	67.3	**6.4**
Migrant	0.3	0.5	0.2
Current Limited English Proficient	24.9	23.0	1.9
Bilingual	0.7	0.5	0.2
ESL Participants	18.5	15.6	2.9
Special Education	12.0	11.8	0.2
Gifted/Talented Participants	10.7	10.8	0.1
At-Risk	53.5	53.6	0.1

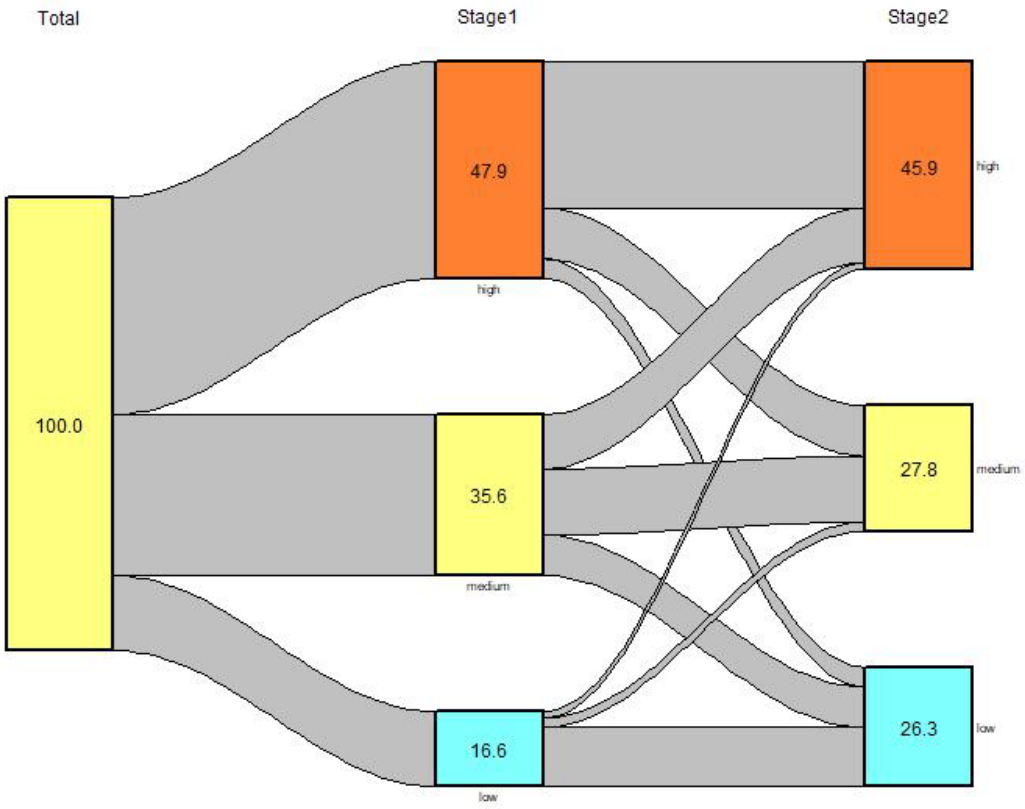
1.6 Percentage of Students Taking Different Test Forms

Table 11 lists the percentages of students who were routed to each of the Stage 1 and Stage 2 panels during the 2023–2024 test administration. To illustrate the number of students routed to different panels during Stage 1 and 2, Figure 2 visually represents the percentage of students routed to various paths for the Grade 5 science test (English) Opp. II, serving as an example. Appendix E includes the visual representations for all tests and opportunities. Based on the numbers and percentages in Table 11, it is evident that a certain percentage of students switched difficulty levels between stages (e.g., low-medium, medium-high, or high-low). For example, in the grade 6 Mathematics test Opp. II, 15.2% of the students began with a medium module but were routed to a low module, and 23.1% of the students started with a medium module but were routed to a high module.

Table 11: Percentages of Students Taking Different Test Forms

Assessment	Route	Opp. I		Opp. II		Opp. III	
		N	%	N	%	N	%
Grade 5 Science (Spanish)	Low-Low	N/A	N/A	78	22.7	125	32.1
	Low-Medium	N/A	N/A	14	4.1	54	13.9
	Low-High	N/A	N/A	2	0.6	15	3.9
	Medium-Low	87	30.6	98	28.5	43	11.1
	Medium-Medium	129	45.4	74	21.5	67	17.2
	Medium-High	68	23.9	21	6.1	7	1.8
	High-Low	N/A	N/A	11	3.2	5	1.3
	High-Medium	N/A	N/A	25	7.3	37	9.5
	High-High	N/A	N/A	21	6.1	36	9.3
Grade 5 Science (English)	Low-Low	N/A	N/A	1,953	13.0	2,082	13.7
	Low-Medium	N/A	N/A	304	2.0	1,114	7.3
	Low-High	N/A	N/A	223	1.5	490	3.2
	Medium-Low	3,501	23.3	1,344	9.0	538	3.5
	Medium-Medium	5,112	34.1	2,183	14.6	1,460	9.6
	Medium-High	6,387	42.6	1,803	12.0	713	4.7
	High-Low	N/A	N/A	648	4.3	151	1.0
	High-Medium	N/A	N/A	1,677	11.2	2,203	14.5
	High-High	N/A	N/A	4,844	32.3	6,478	42.5
Grade 6 Mathematics (English)	Low-Low	N/A	N/A	1,989	24.8	1,921	23.9
	Low-Medium	N/A	N/A	325	4.1	979	12.2
	Low-High	N/A	N/A	240	3.0	163	2.0
	Medium-Low	2,631	33.3	1,216	15.2	472	5.9
	Medium-Medium	4,556	57.7	1,414	17.7	1,760	21.9
	Medium-High	703	8.9	1,854	23.1	1,189	14.8
	High-Low	N/A	N/A	24	0.3	45	0.6
	High-Medium	N/A	N/A	136	1.7	241	3.0
	High-High	N/A	N/A	813	10.1	1,276	15.9
Grade 7 Mathematics (English)	Low-Low	N/A	N/A	1,508	23.7	2,232	35.1
	Low-Medium	N/A	N/A	2,296	36.0	375	5.9
	Low-High	N/A	N/A	249	3.9	79	1.2
	Medium-Low	2,618	41.2	221	3.5	1,329	20.9
	Medium-Medium	2,681	42.2	938	14.7	619	9.7
	Medium-High	1,056	16.6	321	5.0	524	8.2
	High-Low	N/A	N/A	37	0.6	221	3.5
	High-Medium	N/A	N/A	294	4.6	287	4.5
	High-High	N/A	N/A	505	7.9	697	11.0
Grade 8 Social Studies (English)	Low-Low	N/A	N/A	5,508	27.9	6,659	33.2
	Low-Medium	N/A	N/A	2,652	13.4	1,076	5.4
	Low-High	N/A	N/A	288	1.5	481	2.4
	Medium-Low	7,307	36.4	2,092	10.6	2,363	11.8
	Medium-Medium	8,243	41.1	4,365	22.1	2,259	11.3
	Medium-High	4,524	22.5	1,748	8.8	1,865	9.3
	High-Low	N/A	N/A	220	1.1	316	1.6
	High-Medium	N/A	N/A	1,417	7.2	1,052	5.2
	High-High	N/A	N/A	1,484	7.5	4,001	19.9

Figure 2: Percentage of Students Routed to Different Paths (Science, English, Opp. II as an example)



2. TTAP Scores from 2023–2024

At the individual student level, the reported scores included item scores (i.e., whether a student answered each item correctly), scale scores, score gain/loss/no change between opportunities, percentage of correct responses categorized by reporting category and item difficulty level, and current performance levels, which categorize students into the following four levels: 1) Currently Does Not Meet Grade Level, 2) Currently Approaches Grade Level, 3) Currently Meets Grade Level, and 4) Currently Masters Grade Level.

In this section, we provide a detailed overview of the results from each of these reported scores. Additionally, a comprehensive comparison of these reported scores across multiple opportunities is offered to uncover valuable insights into the trends and patterns of student growth as they progress through the year.

2.1 Scaling and Equating

Scaling and equating are statistical procedures that account for the differences in difficulty across test forms and administrations. These procedures place scores on a common scale for meaningful comparison. Similar to the STAAR summative assessments, the TTAP assessments use the Rasch partial-credit model (RPCM; Masters & Wright, 1997), calibrated with Winsteps version 5.6.0.0 (Linacre, 2023). All TTAP assessments are pre-equated prior to test administration. Detailed information on the scaling and equating method can be found in the STAAR Technical Digest, specifically in Chapter 3, Standard Technical Processes3. This method links newly developed items to the existing item bank scale through a set of items that have previously appeared on one or more test forms. This approach enables the determination of the difficulty of newly developed items even before their administration.

With pre-equated item parameters, students' theta scores and the conditional standard error of measurement (CSEM) for each theta score are estimated. Theta scores represent a student's ability level on a standardized scale. To make these scores more interpretable and comparable across different test forms and administrations, the theta scores are converted to scaled scores through a linear transformation. This transformation ensures that the scores are presented in a format easier for interpretation and comparison of student performance.

2.2 Scale Score Gain/Loss Between Opportunities

One of the reported scores is the scale score, which allows comparisons across different opportunities and test forms. Students' growth in terms of their scale scores across three opportunities is analyzed. Descriptive statistics of scale scores from each opportunity are presented in Table 12. In general, students' average scale scores exhibit an increase across opportunities, except for grade 5 Spanish science, where the observed anomaly may be attributed to the relatively small sample size. The mean score from Grade 7 mathematics Opp. III is slightly lower than the mean score from Opp. II, which might be influenced by outliers. The median score (50^{th} P), which is more robust against outliers, is higher in Opp. III compared to Opp. II.

[3] https://tea.texas.gov/student-assessment/testing/student-assessment-overview/2023-tech-digest.pdf

Table 12: Student TTAP Score Growth Across Opportunities

Test	Opportunity	N	Mean	SD	Min	25th P	50th P	75th P	Max
Grade 5 Science (Spanish)	Opp. I	284	805.014	41.611	646	784	806	832	917
	Opp. II	344	799.311	46.747	633	768	802	828	914
	Opp. III	389	801.468	50.650	677	765	801	836	941
Grade 5 Science (English)	Opp. I	15,000	833.622	49.361	646	801	838	868	1,112
	Opp. II	14,979	844.370	54.226	622	809	848	881	1,112
	Opp. III	15,229	855.424	58.347	637	818	860	897	1,112
Grade 6 Mathematics (English)	Opp. I	7,890	637.010	116.978	64	571	632	703	1,364
	Opp. II	8,011	672.546	142.722	64	574	672	763	1,364
	Opp. III	8,046	692.018	156.520	101	582	689	792	1,364
Grade 7 Mathematics (English)	Opp. I	6,355	686.640	122.532	119	617	691	754	1,419
	Opp. II	6,369	710.833	131.185	199	623	700	796	1,419
	Opp. III	6,363	708.465	137.347	270	612	708	799	1,289
Grade 8 Social Studies (English)	Opp. I	20,074	900.338	51.473	626	864	901	937	1,176
	Opp. II	19,774	906.457	51.279	626	866	908	943	1,122
	Opp. III	20,072	913.090	54.076	697	871	912	950	1,150

Note. The notations 25th P, 50th P, and 75th P correspond to the 25th, 50th, and 75th percentiles, respectively.

To evaluate the magnitude of scale score growth across opportunities, the effect size of scale score gain between opportunities is calculated and presented in Table 13. The effect size is determined using Cohen's *d*, a widely used statistical measure that quantifies the effect size of the difference between two groups or conditions and assesses the magnitude of an effect. For reference, Cohen's *d* values are typically interpreted as follows: approximately 0.2 signifies a small effect size; 0.5 represents a medium effect size; and values around 0.8 or higher indicate a large effect size.

In addition to scale scores, students receive a gain, loss, or no change score that reflect their scale score changes across opportunities. Table 14 presents the percentage of students who experienced gains, losses, or no changes in their scaled scores across opportunities.

The effect sizes in Table 13 are around 0.2 or lower consistently, implying that the observed growths in scale scores are relatively small. For grade 5 Spanish science, the effect size is close to 0 or around -0.1 across opportunities. For grade 7 mathematics, the effect size is close to 0 for growth between Opp. II and Opp. III, indicating minimal change between winter and spring. Overall, the effect sizes for Opp. II vs. Opp. I tend to be larger than those for Opp. III vs. Opp. II, suggesting that students showed more progress from fall to winter than they did from winter to spring. The effect sizes reflecting annual growth, specifically between Opp. III vs. Opp. I, range from small to medium.

These trends are similarly reflected in the percentages of students who gained, lost, or experienced no change in their scale scores from Opp. II vs. I, Opp. III vs. II, and Opp. III vs. Opp. I presented in Table 14.

Table 13: Effect Size of Student TTAP Scale Score Growth Across Opportunities

Assessment	Opp. II vs. I	Opp. III vs. II	Opp. III vs. I
Grade 5 Science (Spanish)	-0.129	0.044	-0.077
Grade 5 Science (English)	0.207	0.196	0.403
Grade 6 Mathematics (English)	0.272	0.130	0.398
Grade 7 Mathematics (English)	0.191	-0.018	0.168
Grade 8 Social Studies (English)	0.119	0.126	0.242

Table 14: Percentage of Student with Gain, Loss, or No Change TTAP Scale Scores Across Opportunities

Assessment	Opp. II vs. I Percentage of Gain/Loss/No Change			Opp. III vs. II Percentage of Gain/Loss/No Change			Opp. III vs. I Percentage of Gain/Loss/No Change		
	Loss %	Gain %	No Change %	Loss %	Gain %	No Change %	Loss %	Gain %	No Change %
Grade 5 Science (Spanish)	47.9	49.7	2.4	45.8	52.7	1.4	47.7	51.7	0.7
Grade 5 Science (English)	36.1	62.4	1.5	37.7	61.2	1.1	26.3	72.8	0.9
Grade 6 Mathematics (English)	34.9	65.0	0.1	41.4	58.4	0.3	26.8	73.0	0.2
Grade 7 Mathematics (English)	37.1	62.5	0.4	50.3	49.1	0.5	37.3	62.3	0.4
Grade 8 Social Studies (English)	43.1	56.2	0.7	41.4	58.0	0.6	33.5	65.3	1.2

2.3 TTAP Performance Level

Student performance on the TTAP assessments is categorized into four performance levels presented here. The distribution of students among these performance levels is summarized in Table 15 for each TTAP opportunity, as well as the distribution of performance levels in STAAR. Overall, students exhibit a trend of advancing to higher achievement levels across the opportunities. When comparing the distribution of students' performance levels between Opp. III and STAAR, it is notable that STAAR reports slightly higher percentages of students at the "Masters" or "Meets" levels than TTAP. In general, the percentages at each performance level between TTAP Opp. III and STAAR show similar trends.

- Currently Does Not Meet grade level
- Currently Approaches grade level
- Currently Meets grade level
- Currently Masters grade level

Table 15: Student Performance Level Distribution Across Opportunities

Assessment	PL	Opp. I (N)	Opp. II (N)	Opp. III (N)	STAAR (N)	Opp. I (%)	Opp. II (%)	Opp. III (%)	STAAR (%)
Grade 5 Science (Spanish)	1	246	296	318	319	86.6	86.0	81.7	78.0
	2	35	41	57	65	12.3	11.9	14.7	15.9
	3	3	7	9	21	1.1	2.0	2.3	5.1
	4	0	0	5	4	0.0	0.0	1.3	1.0
	Total	284	344	389	409	100.0	100.0	100.0	100.0
5 Science (English)	1	9,450	7,871	6,597	6,715	63.0	52.5	43.3	42.1
	2	4,204	4,495	4,590	4,737	28.0	30.0	30.1	29.7
	3	934	1,664	2,330	2,457	6.2	11.1	15.3	15.4
	4	412	949	1,712	2,059	2.7	6.3	11.2	12.9
	Total	15,000	14,979	15,229	15,968	100.0	100.0	100.0	100.0
Grade 6 Mathematics (English)	1	3,226	2,810	2,446	2,159	40.9	35.1	30.4	25.4
	2	3,302	2,892	2,766	2,847	41.9	36.1	34.4	33.5
	3	1,210	1,784	1,970	2,318	15.3	22.3	24.5	27.3
	4	152	525	864	1,168	1.9	6.6	10.7	13.8
	Total	7,890	8,011	8,046	8,492	100.0	100.0	100.0	100.0
Grade 7 Mathematics (English)	1	3,761	3,226	3,009	3,123	59.2	50.7	47.3	45.7
	2	1,579	1,519	1,664	1,565	24.8	23.8	26.2	22.9
	3	880	1,398	1,474	1,590	13.8	22.0	23.2	23.3
	4	135	226	216	554	2.1	3.5	3.4	8.1
	Total	6,355	6,369	6,363	6,832	100.0	100.0	100.0	100.0
Grade 8 Social Studies (English)	1	11,042	10,424	9,719	8,935	55.0	52.7	48.4	41.7
	2	5,759	5,417	5,245	5,611	28.7	27.4	26.1	26.2
	3	2,029	2,439	2,781	3,153	10.1	12.3	13.9	14.7
	4	1,244	1,494	2,327	3,722	6.2	7.6	11.6	17.4
	Total	20,074	19,774	20,072	21,421	100.0	100.0	100.0	100.0

Note. Level 1 is Currently does not meet grade level, Level 2 is Currently approaches grade level, Level 3 is Currently meets grade level, and Level 4 is Currently masters grade level.

3. Reliability

3.1 Marginal Reliability

The marginal reliability coefficient (Samejima, 1977, 1994) is used to evaluate the internal test reliability on adaptive assessments. This measure evaluates how well the items on a test that reflect the same construct yield similar results. Marginal reliability is the result of combining measurement errors estimated at different points on the achievement scale into a single index. The formula used to calculate marginal reliability is:

$$\rho_\theta = \frac{\sigma_\theta^2 - M_{S_\theta^2}}{\sigma_\theta^2}$$

where σ_θ^2 is the observed variance of the ability estimates, θ, and $M_{S_\theta^2}$ is the observed mean of the score's conditional error variances at each value of θ. Tests are considered reliable when their marginal reliability coefficients range from 0.80 and above.

Table 16 provides a comparison of the marginal reliability coefficients for TTAP and STAAR during the 2023–2024 school year. The table also includes reliability at the subgroup level for gender and ethnicity, but only for subgroups with sample sizes equal to or larger than 200. Reliabilities for smaller subgroups are omitted to prevent potentially misleading conclusions based on limited data.

When assessing the three opportunities within TTAP, Opp. I exhibits lower reliabilities, while Opp. III demonstrates higher reliabilities. The longer test length of Opp. III contributes to the expected increase in reliability. Comparing the reliability of TTAP Opp. III with STAAR, Opp. III demonstrates higher reliabilities for grade 5 Spanish and English science, as well as grade 6 mathematics. However, STAAR reports higher reliabilities for grade 7 mathematics and grade 8 social studies, both at the overall level and in most subgroup analyses.

Upon examining reliabilities at the subgroup level, there is a general pattern of comparability across subgroups, with a few exceptions. For some subgroups, such as Black or African American, Hispanic or Latino, two races, and female students, the reliabilities for these subgroups fall below 0.7 (indicated by bolded values in Table 16), which are relatively lower when compared to other ethnicity or gender subgroups.

Table 16: Test Reliabilities of TTAP and STAAR

Assessment	Group	N	Opp. I	Opp. II	Opp. III	STAAR
Grade 5 Science (Spanish)	All	409	**0.671**	0.735	0.854	0.732
	Ethnic: H	399	**0.672**	0.735	0.855	0.729
	Sex: M	227			0.867	0.784
Grade 5 Science (English)	All	15,968	0.752	0.798	0.884	0.858
	Ethnic: A	1,097	0.701	0.763	0.861	0.837
	Ethnic: B	1,250	0.749	0.788	0.877	0.806
	Ethnic: H	7,517	0.724	0.78	0.874	0.831
	Ethnic: T	510	0.746	0.797	0.881	0.864
	Ethnic: W	5,478	0.722	0.768	0.864	0.852
	Sex: F	7,931	0.749	0.783	0.882	0.851
	Sex: M	8,037	0.754	0.808	0.886	0.863
Grade 6 Mathematics (English)	All	8,492	0.736	0.824	0.898	0.888
	Ethnic: A	567	0.763	0.811	0.881	0.823
	Ethnic: B	788	0.719	0.807	0.871	0.854
	Ethnic: H	4,279	0.705	0.804	0.885	0.876
	Ethnic: T	261	**0.690**	0.818	0.903	0.887
	Ethnic: W	2,522	0.705	0.798	0.886	0.879
	Sex: F	4,240	0.727	0.819	0.893	0.884
	Sex: M	4,252	0.744	0.828	0.902	0.892
Grade 7 Mathematics (English)	All	6,832	0.742	0.786	0.879	0.890
	Ethnic: B	715	**0.643**	0.733	0.838	0.833
	Ethnic: H	3,921	0.737	0.779	0.876	0.888
	Ethnic: T	200				0.892
	Ethnic: W	1,800	0.733	0.768	0.867	0.884
	Sex: F	3,398	**0.699**	0.768	0.864	0.881
	Sex: M	3,434	0.773	0.800	0.890	0.897
Grade 8 Social Studies (English)	All	21,421	0.782	0.808	0.892	0.895
	Ethnic: A	1,015	0.750	0.784	0.879	0.850
	Ethnic: B	1,793	0.730	0.784	0.867	0.859
	Ethnic: H	12,251	0.755	0.780	0.873	0.874
	Ethnic: T	579	0.803	0.837	0.902	0.905
	Ethnic: W	5,636	0.766	0.791	0.886	0.892
	Sex: F	10,490	0.764	0.793	0.886	0.890
	Sex: M	10,930	0.796	0.819	0.898	0.898

Note. Reliability is only reported for subgroups with sample sizes equal to or greater than 200.
Sex: F – Female, Sex: M – Male
Ethnic: A – Asian, Ethnic: B – Black or African American, Ethnic: H – Hispanic/Latino, Ethnic: T – Two races, Ethnic: W – White

3.2 Classification Consistency and Accuracy

Information regarding classification consistency and accuracy has been derived from actual test outcomes from the 2023–2024 test administration. Because all test scores have inherent measurement error, these classifications are also prone to errors. Two metrics are often used to assess the quality of these classifications: consistency and accuracy. Consistency measures the percentage of students who are placed in the same performance levels if they take two parallel forms of a test. Accuracy measures the percentage of students correctly classified into their true performance levels based on their observed test scores. Although related, classification consistency and accuracy are distinct concepts; high consistency does not always equate to high accuracy, and vice versa. To gain a better understanding of classification quality, we analyzed both consistency and accuracy of students' performance level classifications, using results from tests with established performance standards.

We applied the same methods outlined in the STAAR Technical Digest to compute classification consistency and accuracy. Estimates of marginal classification accuracy and consistency are calculated using Rudner's (2000, 2005) method and its extensions by Li (2006). Table 17 presents the classification consistency and accuracy for each opportunity of TTAP tests, along with these statistics from the corresponding STAAR tests documented in the latest STAAR Technical Digest from Spring 2024. The classification consistency and accuracy values for TTAP are comparable to those observed in the STAAR assessments. For all TTAP tests, except for Spanish Science, which has higher classification consistency and accuracy, the classification consistency ranges from 0.636 to 0.717, while the classification accuracy falls between 0.726 and 0.791.

Table 17: Classification Consistency and Accuracy

Assessment	Opps	N	Classification Consistency	Classification Accuracy
Grade 5 Science (Spanish)	Opp. I	284	0.835	0.882
	Opp. II	344	0.842	0.883
	Opp. III	389	0.851	0.890
	STAAR	397,753	0.819	0.869
Grade 5 Science (English)	Opp. I	15,000	0.688	0.763
	Opp. II	14,979	0.647	0.726
	Opp. III	15,229	0.671	0.753
	STAAR	380,984	0.677	0.759
Grade 6 Mathematics (English)	Opp. I	7,890	0.636	0.732
	Opp. II	8,011	0.642	0.737
	Opp. III	8,046	0.699	0.783
	STAAR	387,455	0.699	0.783
Grade 7 Mathematics (English)	Opp. I	6,355	0.659	0.744
	Opp. II	6,369	0.645	0.733
	Opp. III	6,363	0.717	0.791
	STAAR	324,109	0.739	0.810
Grade 8 Social Studies (English)	Opp. I	20,074	0.667	0.742
	Opp. II	19,774	0.661	0.739
	Opp. III	20,072	0.715	0.788
	STAAR	405,802	0.717	0.792

Notes. 1. Consistency indicates the proportion of students that would be classified into the same performance levels if they were administered a parallel test form. The proportions are converted to a 0%–100% scale. 2. Accuracy indicates the proportion of students that are accurately classified. The proportions are converted to a 0%–100% scale.

4. Validity

4.1 TTAP and STAAR Correlations

The Pearson correlations between the TTAP and STAAR summative scale scores are calculated as criterion validity evidence of the TTAP scores. Pearson correlation is a statistical measure that quantifies the strength and direction of the linear relationship between two continuous variables. It provides a value between -1 and 1, where -1 indicates a perfect negative linear relationship, 1 indicates a perfect positive linear relationship, and 0 suggests no linear relationship between the variables. Table 18 shows the Pearson correlations between TTAP and STAAR scores by opportunity, subject, and grade.

Table 18 also showcases patterns of associations across different opportunities (Opp. I, Opp. II, Opp. III) and STAAR. The correlation values from Spanish Science are in the range between 0.466-0.706, indicating a relationship that varies from moderate to strong.

Across the various other values in the table, the correlations between Opp. I, Opp. II, and Opp. III are moderately strong, generally ranging between 0.706 to 0.809 This suggests a consistent

positive relationship in scores across these opportunities. The correlations between Opp. I, Opp. II, Opp. III and STAAR are also moderately strong, falling between 0.722 and 0.852. It is notable that the correlation values between Opp. III and STAAR tend to be higher than those between Opp. I or Opp. II with STAAR, showing that Opp. III is a better predictor for STAAR scores. Overall, the results indicate moderate to strong positive relationships between the various opportunities and STAAR, with a more pronounced relationship in the latter opportunities. The correlations, considered criterion validity evidence of the TTAP scores, are moderately high, except for Spanish Science, where the sample size is relatively small.

Table 18: Pearson Correlation Coefficients Between the TTAP and Summative Assessment Scale Scores

Assessment	Opp.	Opp. I	Opp. II	Opp. III	STAAR
Grade 5 Science (Spanish)	Opp. I	1.000	0.572	0.538	0.466
	Opp. II		1.000	0.688	0.632
	Opp. III			1.000	0.706
	STAAR				1.000
Grade 5 Science (English)	Opp. I	1.000	0.716	0.741	0.729
	Opp. II		1.000	0.766	0.738
	Opp. III			1.000	0.814
	STAAR				1.000
Grade 6 Mathematics (English)	Opp. I	1.000	0.733	0.747	0.722
	Opp. II		1.000	0.809	0.790
	Opp. III			1.000	0.852
	STAAR				1.000
Grade 7 Mathematics (English)	Opp. I	1.000	0.706	0.746	0.733
	Opp. II		1.000	0.779	0.760
	Opp. III			1.000	0.819
	STAAR				1.000
Grade 8 Social Studies (English)	Opp. I	1.000	0.761	0.774	0.766
	Opp. II		1.000	0.806	0.796
	Opp. III			1.000	0.851
	STAAR				1.000

4.2 Prediction Agreement

Beginning with the 2023–2024 TTAP assessments, students' scaled scores on the TTAP are used to predict their performance levels on STAAR assessments categorized into four levels with three cut scores. These four performance levels are

- Predicted to be Masters Grade Level;
- Predicted to be Meets Grade Level;
- Predicted to be Approaches Grade Level; and
- Predicted to be Did Not Meet Grade Level.

Receiver operating characteristic (ROC) curves were employed to predict students' STAAR performance level based on their TTAP scale score. These curves were employed to find the TTAP scale score that optimizes the accuracy of predicting STAAR performance levels, while balancing true positives and true negatives. In essence, ROC curve analyses help identify the threshold TTAP score that strikes the best balance in accurately predicting students' performance on the STAAR assessment. Table 19–Table 21 are the prediction summaries by TTAP assessments and assessment opportunities. These summaries include prediction accuracy, specificity (true negative rate), sensitivity (true positive rate), and Area under the ROC Curve (AUC). The AUC measures the overall ability of the classifier to discriminate between positive and negative instances.

Table 19, Table 20, and Table 21 present the prediction results by opportunities. In each of the tables, the optimally derived TTAP cut scores using the Youden Index (Youden, 1950) for Approaches, Meets, and Masters are presented in the Cut column. The other columns present values based on the evaluation metrics. The values in Table 19 - Table 21 that are highlighted in green show cells with convincing evidence (>=0.8), yellow denote acceptable evidence (>=0.7 and <0.8) according to the National Center on Intensive Intervention criteria.

Results show that all the AUC observed were at or above 0.85. The specificity and sensitivity values are either above 0.80 or close to 0.80. Among the three opportunities, the specificity, sensitivity, and AUC values are lowest in Opp. 1 and highest in Opp. 3. This pattern aligns with expectations, given that Opp. 3, administered closest to the STAAR assessment, is anticipated to yield superior predictions of STAAR performance levels in comparison to the other two opportunities. Within the same opportunity and test, predictions of the "Approach" performance level often marginally lag behind predictions for the other two performance levels.

Table 19: Prediction Accuracy Summary (Opp. I)

Assessment	Cut	Performance Level	Accuracy	Specificity	Sensitivity	AUC
Grade 5 Science (both English and Spanish)	829	Approaches	0.79	0.78	0.79	0.87
	854	Meets	0.80	0.81	0.80	0.88
	865	Masters	0.81	0.8	0.84	0.90
Grade 6 Mathematics (English)	622	Approaches	0.74	0.82	0.72	0.85
	674	Meets	0.80	0.85	0.72	0.87
	702	Masters	0.80	0.79	0.84	0.9
Grade 7 Mathematics (English)	691	Approaches	0.76	0.78	0.75	0.85
	716	Meets	0.79	0.78	0.81	0.89
	765	Masters	0.83	0.82	0.92	0.94
Grade 8 Social Studies (English)	896	Approaches	0.81	0.82	0.8	0.89
	915	Meets	0.83	0.83	0.83	0.91
	927	Masters	0.81	0.8	0.89	0.92

Table 20: Prediction Accuracy Summary (Opp. II)

Assessment	Cut	Performance Level	Accuracy	Specificity	Sensitivity	AUC
Grade 5 Science (both English and Spanish)	848	Approaches	0.79	0.84	0.75	0.87
	870	Meets	0.81	0.83	0.78	0.88
	881	Masters	0.81	0.81	0.84	0.90
Grade 6 Mathematics (English)	618	Approaches	0.81	0.82	0.80	0.89
	696	Meets	0.82	0.82	0.81	0.90
	767	Masters	0.85	0.85	0.86	0.93
Grade 7 Mathematics (English)	718	Approaches	0.79	0.83	0.75	0.87
	748	Meets	0.83	0.84	0.82	0.90
	799	Masters	0.85	0.84	0.90	0.94
Grade 8 Social Studies (English)	898	Approaches	0.83	0.80	0.84	0.90
	925	Meets	0.83	0.84	0.82	0.91
	934	Masters	0.83	0.81	0.88	0.93

Table 21: Prediction Accuracy Summary (Opp. III)

Assessment	Cut	Performance Level	Accuracy	Specificity	Sensitivity	AUC
Grade 5 Science (both English and Spanish)	847	Approaches	0.83	0.79	0.86	0.91
	883	Meets	0.84	0.85	0.83	0.92
	897	Masters	0.84	0.84	0.89	0.94
Grade 6 Mathematics (English)	622	Approaches	0.84	0.84	0.84	0.92
	692	Meets	0.85	0.83	0.88	0.94
	804	Masters	0.87	0.86	0.91	0.95
Grade 7 Mathematics (English)	705	Approaches	0.83	0.85	0.82	0.91
	741	Meets	0.84	0.82	0.88	0.93
	831	Masters	0.89	0.89	0.91	0.96
Grade 8 Social Studies (English)	903	Approaches	0.85	0.86	0.85	0.93
	927	Meets	0.85	0.84	0.89	0.94
	948	Masters	0.88	0.87	0.90	0.95

Other validity evidence for the TTAP assessments comes from a variety of sources in relation to the STAAR assessments, including test content, response processes, internal structure, and analysis of the consequences of testing. Refer to Technical Digest[4] Chapter 3, Standard Technical Processes and Chapter 4, State of Texas Assessments of Academic Readiness (STAAR) for additional information about validity.

5. Fairness

The fairness of the TTAP assessments can be examined by a statistical evaluation using DIF and a bias review by content specialists. For the statistical evaluation, the Mantel-Haenszel (MH) method (1959) has been applied to the TTAP assessments to assess DIF of the items. DIF refers to items that appear to function differently across identifiable groups, typically across different demographic groups. DIF is officially collected on this program using field-test data. The MH method is the most cited and studied method for detecting DIF. DIF analysis has been conducted for all items regarding gender and ethnicity bias. All field-tested items are carefully evaluated for DIF prior to being placed on an operational form. The following focal and reference groups are used:

Focal Group		Reference Group
Females (F)	vs.	Males (M)
African Americans (AA)	vs.	Whites (W)
Hispanics (H)	vs.	Whites (W)

[4] https://tea.texas.gov/student-assessment/testing/student-assessment-overview/2023-tech-digest.pdf

A generalized MH procedure is applied to calculate DIF. The generalizations include (1) adaptation to polytomous items and (2) improved variance estimators to render the test statistics valid under complex sample designs. With this procedure, each student's ability estimate on the operational items (e.g., raw score) on a given test is used as the ability-matching variable. The corresponding scores are typically divided into 10 intervals to compute the Mantel-Haenszel Chi-Square ($MH\chi^2$) DIF statistics for balancing the stability and sensitivity of the DIF scoring category selection, population permitting. The analysis program computes the $MH\chi^2$ value, the conditional odds ratio, and the MH-delta for dichotomous items; the generalized Mantel-Haenszel Chi-Square ($GMH\chi^2$) and the standardized mean difference (SMD) are computed for polytomous items.

Items are classified into three categories (A, B, or C), ranging from no evidence of DIF to severe DIF according to the Educational Testing Service (ETS) classification convention for dichotomous items (Dorans & Holland, 1993) and the ETS/National Assessment of Educational Progress (NAEP) classification generalization for polytomous items (as cited in Michaelides, 2008), which is illustrated in Table 22. Table 22 presents the criteria for each level of classification. Items are also categorized as positive DIF (+A, +B, or +C), signifying that the item favors the focal group (e.g., African American/Black, Hispanic, female), or negative DIF (–A, –B, or –C), signifying that the item favors the reference group (e.g., White, male). Items are flagged if their DIF statistics fall into the "C" category for any group. A DIF classification of "C" indicates that the item shows significant DIF and should be reviewed for potential content bias, differential validity, or other issues that may reduce item fairness. These items are flagged regardless of whether the DIF statistic favors the focal or reference group.

It should be noted that DIF analyses serve merely to identify test items that have unusual statistical characteristics related to student group performance. The DIF analyses alone do not prove that specific items are biased. Such judgments are made by item reviewers who are knowledgeable about the State's content standards, instructional methodology, and student testing behavior.

Table 22: DIF Classification Rules for Items

DELTA Metric	
Category	Rule
C	$GMH\chi^2$ is significant at .05 and $\lvert\Delta_{MH}\rvert > 1.5$
B	$GMH\chi^2$ is significant at .05 and $1 < \lvert\Delta_{MH}\rvert \leq 1.5$
A	$GMH\chi^2$ is not significant at .05 or $\lvert\Delta_{MH}\rvert \leq 1$
SMD Metric	
Category	Rule
C	$GMH\chi^2$ is significant at .05 and $\frac{\lvert SMD\rvert}{\sigma} > .25$
B	$GMH\chi^2$ is significant at .05 and $.17 < \frac{\lvert SMD\rvert}{\sigma} \leq .25$
A	$GMH\chi^2$ is not significant at .05 or $\frac{\lvert SMD\rvert}{\sigma} \leq .17$

6. Reporting

Reporting occurs at various levels, including the student, campus, and district levels. More detailed information is accessible at the individual student level compared to the aggregated levels. Figure 3, Figure 4, and Figure 5 provide visual representations of the reports available at the individual student level, offering detailed insights into each student's performance. On the other hand, Figure 6 and Figure 7 depict the reports available at the campus and district levels, providing a broader overview of performance trends and patterns across groups of students.

6.1 Student-Level Reports

Student reports provide valuable insights for educators, parents, and students themselves to monitor academic progress throughout the school year. At the individual student level, Figure 3 outlines the comprehensive set of scores and indicators that students receive.

- **Scale Score.** Students are provided with a scale score, which varies depending on the subject area.
  - Vertical scale score (for mathematics assessments)
  - Horizontal scale score (for science and social studies assessments)
- **Opportunity Performance Level.** This classification categorizes a student's performance into one of four levels:
  - Currently Did Not Meet grade level.
  - Currently Approaches grade level.
  - Currently Meets grade level.
  - Currently Masters grade level.
- **Score Difference Between Opportunities.** Students' progress is tracked by comparing their performance across different opportunities:

  - Opp. I: No gain or loss score is reported since students do not have scores from previous opportunities.
  - Opp. II
    - Difference in scale scores between Opp. II and Opp. I.
  - Opp. III
    - Difference in scale scores between Opp. III and Opp. I.
    - Difference in scale scores between Opp. III and Opp. II.
- **Predicted STAAR Performance Level.** Students are provided with a predicted performance level on the STAAR assessment, categorized into four levels:
  - Predicted to be Masters Grade Level.
  - Predicted to be Meets Grade Level.
  - Predicted to be Approaches Grade Level.
  - Predicted to be Did Not Meet Grade Level.
- **Forms Received (Difficulty Level of Stage 1 and Stage 2 Forms).** Students' assessments are further detailed by indicating the difficulty level of the forms received:
  - Opp. I: Low/Medium/High.
  - Opp. II/III:
    - Stage 1 Form: Low, Medium, or High.
    - Stage 2 Form: Low, Medium, or High.

Figure 3: Individual Student Report (Overall Scores)

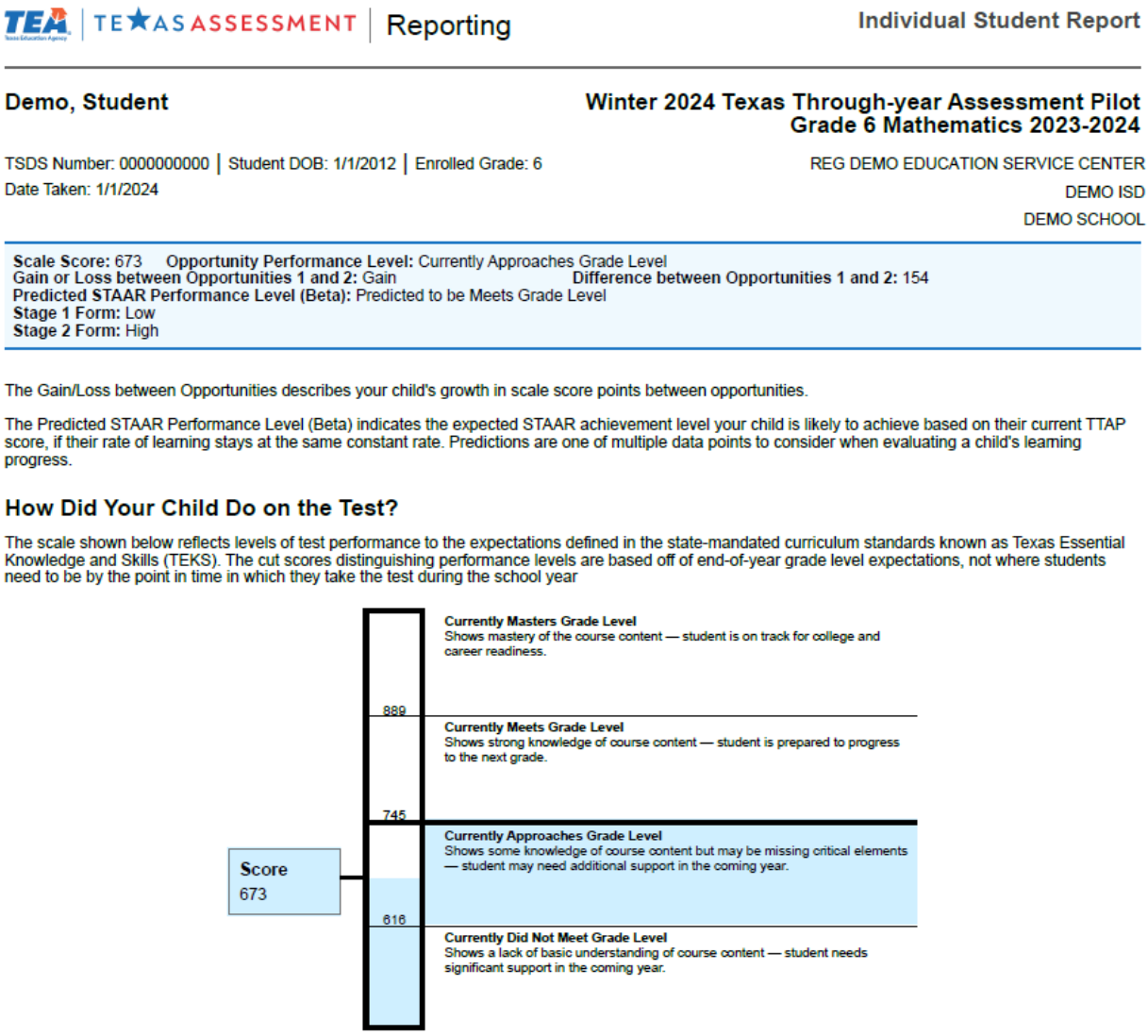


Figure 4 illustrates that in addition to the previously mentioned scores, students also receive detailed information within each reporting category. This includes the following:

- **Total Points Possible by Item Difficulty (Low/Medium/High).** Students are informed of the total number of points that could have been earned on items categorized by difficulty level.
- **Percent of Points the Student Earned by Item Difficulty (Low/Medium/High).** This metric indicates the percentage of points the student actually earned out of the total possible points, categorized by item difficulty level (low, medium, high). It offers a nuanced understanding of students' performance relative to the difficulty of the items attempted.

- **Number of Points Earned and the Item Difficulty Category (Low/Medium/High) at the Item Level.** Students are provided with a breakdown for each item attempted, specifying the number of points earned alongside the corresponding item difficulty category (low, medium, high). This granular level of detail allows students, educators, and parents to identify specific areas of strength and weakness within each reporting category.

Figure 4: Individual Student Report (Reporting Category Level Scores)

How Did Your Child Perform on Different Areas of the Test?

Reporting categories are groupings of related skills.

Reporting Category	(1) Low Difficulty-Total Points Possible	(1) Low Difficulty Percent Correct	(2) Medium Difficulty-Total Points Possible	(2) Medium Difficulty Percent Correct	(3) High Difficulty-Total Points Possible	(3) High Difficulty Percent Correct
1. Numerical Representations and Relationships	1	100%	5	20%	1	0%
2. Computations and Algebraic Relationships	3	100%	1	0%	3	33%
3. Geometry and Measurement	2	100%	0	N/A	1	0%
4. Data Analysis and Personal Financial Literacy	2	100%	0	N/A	3	0%

How Did Your Child Perform on Each Item?

The tables below are organized by reporting category and show how your student scored on each question on the assessment.

1. Numerical Representations and Relationships

Item #	Item Difficulty	Standard Key	Student Expectation	Points
4	Medium	6.1.4.E	Represent ratios and percents with concrete models, fractions, and decimals.	0/1
7	Low	6.1.2.D	Order a set of rational numbers arising from mathematical and real-world contexts.	1/1
10	Medium	6.1.7.A	Generate equivalent numerical expressions using order of operations, including whole number exponents and prime factorization.	1/2
13	Medium	6.1.7.C	Determine if two expressions are equivalent using concrete models, pictorial models, and algebraic representations.	0/2
19	High	6.1.2.D	Order a set of rational numbers arising from mathematical and real-world contexts.	0/1

2. Computations and Algebraic Relationships

Item #	Item Difficulty	Standard Key	Student Expectation	Points
2	Low	6.2.3.D	Add, subtract, multiply, and divide integers fluently.	1/1
5	Low	6.2.3.D	Add, subtract, multiply, and divide integers fluently.	1/1
9	Low	6.2.3.C	Represent integer operations with concrete models and connect the actions with the models to standardized algorithms.	1/1
12	High	6.2.6.B	Write an equation that represents the relationship between independent and dependent quantities from a table.	0/1
15	High	6.2.4.B	Apply qualitative and quantitative reasoning to solve prediction and comparison of real-world problems involving ratios and rates.	0/1
17	High	6.2.3.D	Add, subtract, multiply, and divide integers fluently.	1/1
20	Medium	6.2.5.B	Solve real-world problems to find the whole given a part and the percent, to find the part given the whole and the percent, and to find the percent given the part and the whole, including the use of concrete and pictorial models.	0/1

3. Geometry and Measurement

Item #	Item Difficulty	Standard Key	Student Expectation	Points
3	Low	6.3.8.A	Extend previous knowledge of triangles and their properties to include the sum of angles of a triangle, the relationship between the lengths of sides and measures of angles in a triangle, and determining when three lengths form a triangle.	1/1
8	Low	6.3.11.A	Graph points in all four quadrants using ordered pairs of rational numbers.	1/1
16	High	6.3.8.B	Model area formulas for parallelograms, trapezoids, and triangles by decomposing and rearranging parts of these shapes.	0/1

4. Data Analysis and Personal Financial Literacy

Item #	Item Difficulty	Standard Key	Student Expectation	Points
6	Low	6.4.12.A	Represent numeric data graphically, including dot plots, stem-and-leaf plots, histograms, and box plots.	1/1
11	Low	6.4.13.A	Interpret numeric data summarized in dot plots, stem-and-leaf plots, histograms, and box plots.	1/1
14	High	6.4.13.A	Interpret numeric data summarized in dot plots, stem-and-leaf plots, histograms, and box plots.	0/1
18	High	6.4.14.C	Balance a check register that includes deposits, withdrawals, and transfers.	0/1
21	High	6.4.12.D	Summarize categorical data with numerical and graphical summaries, including the mode, the percent of values in each category (relative frequency table), and the percent bar graph, and use these summaries to describe the data distribution.	0/1

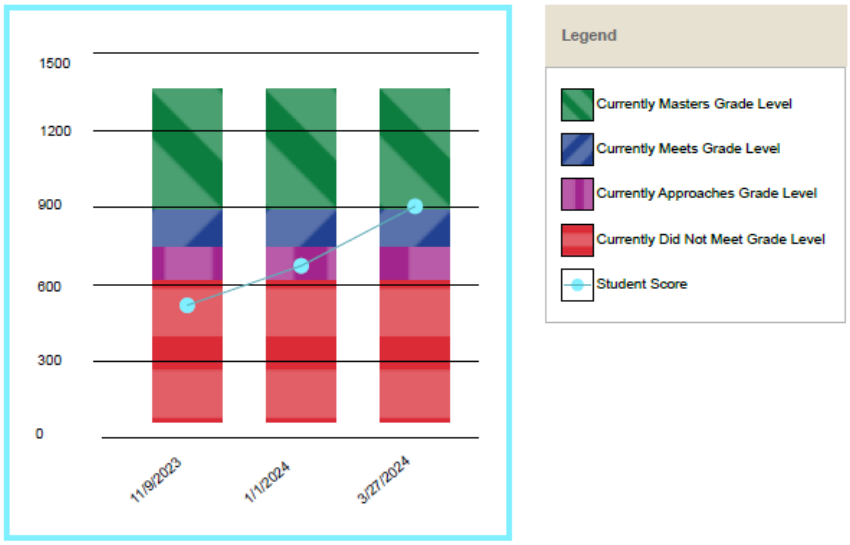
Finally, a student's performance over three opportunities is tracked. Figure 5 serves to illustrate the tracking of a student's performance across three distinct opportunities. Student performance levels and scale scores are displayed in both chart and table formats in the student-level report. This allows for monitoring progress over time, facilitating identification of trends and areas for improvement.

Figure 5: Individual Student Report (Progress Monitoring)

Your Child's Progress

Longitudinal Trend Chart Information

The chart below reports your child's performance over time. The shaded areas in multiple colors indicate the scale score range in each achievement level. Each mark on the graph represents your child's score and indicates whether he or she met the standards that year.



Your Child's Progress

Date	Test Administration	Assessment Name	Scale Score	Opportunity Performance Level
11/9/2023	Opportunity 1	Fall 2023 Texas Through-year Assessment Pilot Grade 6 Mathematics	519	Currently Did Not Meet Grade Level
1/1/2024	Opportunity 2	Winter 2024 Texas Through-year Assessment Pilot Grade 6 Mathematics	673	Currently Approaches Grade Level
3/27/2024	Opportunity 3	Spring 2024 Texas Through-year Assessment Pilot Grade 6 Mathematics	905	Currently Masters Grade Level

6.2 Campus-/District-Level Reports

As depicted in Figure 6 and Figure 7, the following scores are presented in the campus- or district-level reports.

- **Aggregated Mean Scale Score Across District or Campus (Average Score).** This score represents the average scale score attained by students within the district or campus, offering a measure of academic achievement by students within the aggregated unit overall.
- **Distribution of Students Among Performance Levels (Opportunity Performance Distribution).** This highlights how students are distributed across different performance levels (e.g., Did Not Meet grade level, Approaches grade level, Meets grade level, Masters grade level), providing insights into the overall proficiency levels of the students within the district or campus.
- **Average Percent of Items Answered Correctly (Average Percent Correct), by Item Difficulty (Low/Medium/High), at the Reporting Category Level.** This metric reveals the average percentage of items answered correctly by students, categorized by item

difficulty levels (low, medium, high) within each reporting category. It offers a detailed assessment of students' performance across different reporting categories and item difficulty levels.

- **Mean Raw Score by Item.** This denotes the average raw score attained by students for each individual item, providing a nuanced understanding of performance at the granular level. It aids in identifying specific areas of strength and weakness within the curriculum, guiding instructional decisions.

Figure 6: District/Campus Report (Scale Score and Performance Level)

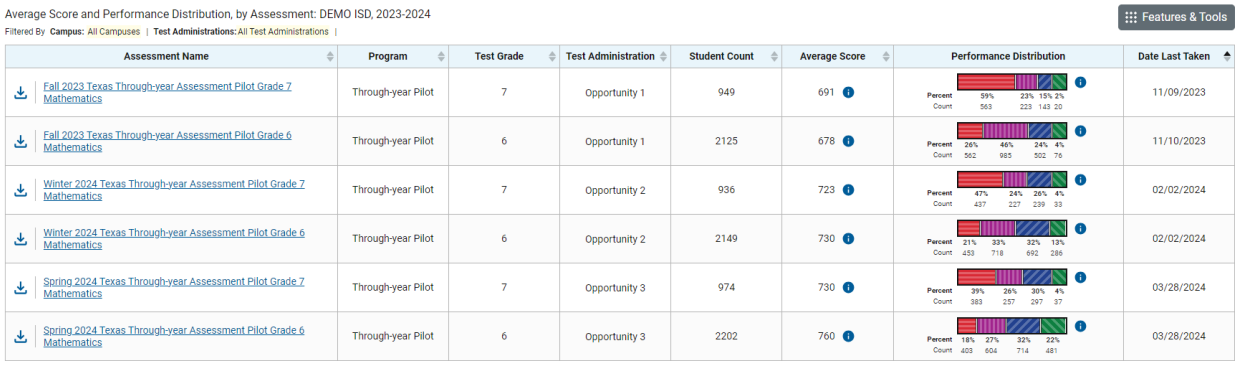
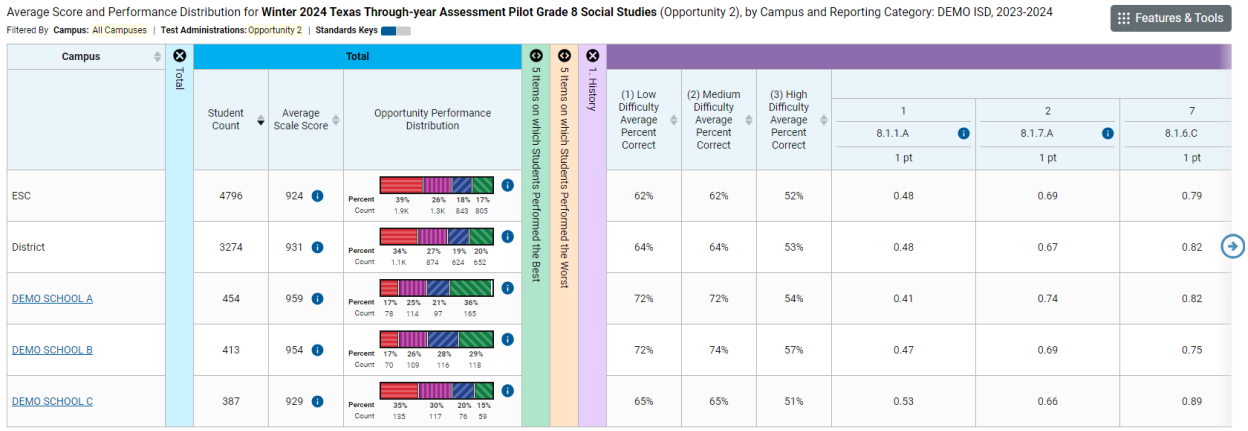


Figure 7: District/Campus Report (Percentage Correct)



7. Continuous Research and Improvement Plans

In the 2023–2024 school year, six special studies were conducted to evaluate the reliability and validity of TTAP assessments and its comparability with STAAR. The findings from these studies are detailed in this section to guide the future design and implementation of TTAP. Continuous research and improvement plans are essential for ensuring that TTAP aligns with its intended purpose of bridging interim and summative assessment systems into a single, coherent assessment system.

In the coming years, we plan to undertake comprehensive reevaluations of TTAP assessments. This will include examining the test design to ensure it effectively measures the intended skills and knowledge. We will also explore various score reporting options to provide actionable insights for teachers and students. Additionally, the psychometric properties of the assessments will be assessed to confirm their reliability, validity, and fairness. Through continuous research, we aim to enhance the overall quality and design of TTAP assessments.

Special Study 1: Comparing the Psychometric Properties of TTAP and STAAR

Test administration year 2023–2024 was the second pilot year of TTAP. Continuing the comparison study carried out for the first pilot year, this study aimed to evaluate and compare the psychometric properties of TTAP and STAAR. Simulations were performed for each of the three opportunities of TTAP and STAAR using a shared set of true ability values. Results showed that TTAP Opp. III had similar psychometric properties with STAAR, suggesting that TTAP Opp. III was more efficient compared to STAAR due to its shorter test lengths. Using the Spearman-Brown method (Spearman, 1910), the findings also revealed that the psychometric properties of TTAP Opp. I and Opp. II showed small reliability gains compared to STAAR if STAAR's test length was shortened to match TTAP Opp. I and II. Among the four test titles of TTAP, grade 5 science had slightly degraded properties compared to other titles, which can be attributed to its items being easier for the simulated students reflecting the ability distribution of Texas's student population. Overall, this study shows preliminary positive evidence that TTAP and STAAR would likely provide similar interpretations of student ability if administered within the same testing window for Opp. III, consistent with the findings from the previous year. In addition, the psychometric properties of TTAP Opp. I and II are comparable to those of STAAR after accounting for the effects of differential test lengths.

Special Study 2: Investigating Approaches for Establishing the Final Summative Score of Record on TTAP

TTAP is a multiyear pilot that is currently investigating creating a linear composite final score for summative determinations and accountability for test events that follow particular guidelines. This study is an extension of previous work on this topic (Gianopulos et al., 2024) using data from Pilot Year 2 with method modifications based on feedback from the Texas Technical Advisory Committee. We compared the seven original methods for producing linear composite scores and included multiple regression as a new method. Mean equating (Kolen & Brennan, 2014) was used to adjust the STAAR scores to align to the third TTAP test opportunity (Opp. III) to improve outcome interpretability. Finally, we investigated the extent and nature of missing data.

Linear composites are essentially weighted averages of the three TTAP opportunity scores (Opp. I, Opp. II, Opp. III). Findings showed the linear composite score consistently reduced measurement noise across tests, across criteria, and across years in comparison to stand-alone

test scores. Reducing measurement noise meant the linear composite scores were more stable measures of student performance. Linear composites tended to underestimate STAAR, while the help-not-hurt rule, which took the higher of Opp. III or the linear composite, tended to overestimate STAAR. The linear composite scores tended to show the highest correlations with STAAR of the methods studied.

Findings showed that missing TTAP scores were not missing at random and were non-ignorable. Students with one or more missing TTAP score scored lower on STAAR. These findings were robust to any of the studied conditions. We found multiple regression was the best current method for increasing agreement to the aligned STAAR score. However, if comparability to STAAR were not the success criteria, using the maximum score across three opportunities with the same number of points would have sound psychometric qualities and offer many logistical and motivational advantages.

Special Study 3: A TTAP-to-STAAR Prediction Study

Through-year assessments are multiple-administration assessments designed to integrate interim and summative assessment systems into a single, coherent system. One purpose of through-year assessments is to predict whether a student will achieve a performance level of importance on a large-scale summative assessment (Perie, Marion, & Gong, 2009). These predictions provide teachers and students valuable information in terms of whether a student is likely or unlikely to meet performance standards.

Previous work (Schneider, Liu, & Robinson, 2022) found that the ROC curves functioned effectively to identify cuts and provide predictions that give teachers information regarding whether a student is likely or unlikely to meet performance standards. Using this ROC method, cuts were identified based on data from 2022–2023. These cuts were then implemented in the 2023–2024 TTAP tests to provide students with their predicted STAAR performance levels at each opportunity. The findings from the previous work (Schneider et al.) also underscores the recommendation to annually recalibrate the cuts using the ROC method.

In this study, CAI recalibrated the cuts using the ROC method for each of the three TTAP opportunities using data from the 2023–2024 school year. The predictions were evaluated in terms of sensitivity, specificity, and AUC. The predictions will provide teachers and students with helpful information in terms of whether a student is likely or unlikely to meet performance standards.

Special Study 4: Mathematics Item Difficulty Modeling Study

An item difficulty modeling (IDM) study can illuminate item features that predict difficulty and should be embedded into Range Performance Level Descriptors (RPLDs). An IDM study was conducted to explore the connections between item features (e.g., cognitive, content, and stimulus demands) and item difficulty using 2,718 mathematics items from grades 3, 6, and 7

TTAP and STAAR item pools. Embedding item features that drive difficulty into RPLDs ensures their validity. Understanding how these features influence difficulty is essential for guiding item developers in creating items with desired difficulty levels.

The study evaluated the predictive accuracy of linear regression and random forest models in determining item difficulties, achieving R-squared values between 0.42–0.56 and 0.38–0.55, respectively. These results suggest that more complex models do not necessarily provide better predictions than the simpler Ordinary Least Square linear regression model. The analysis highlighted significant predictors of item difficulty, such as average scores, median response times, equation item types, and items featuring tables or equations. Notably, longer median response times were associated with higher item difficulty, likely due to the complexity of the tasks requiring more processing time. Additionally, equation items and items with equations generally posed greater challenges. The study found that items with higher average scores and those featuring tables were less difficult, and multiple-select items were easier for grades 6 and 7, contradicting previous findings by Schneider, Chen, and Nichols (2021). These insights may be helpful to item writers aiming to craft items that align with specific difficulty levels. Understanding the factors influencing item difficulty, such as equation item types and median response times, allows for strategic adjustments to meet intended assessment objectives and proficiency levels.

Special Study 5: Routing Study

As a multistage assessment comprised of two stages, the TTAP makes routing decisions for placing students in three possible difficulty tiers in each stage. This study examined psychometric properties related to routing decisions in TTAP as well as related to its performance-level classifications. A series of research questions were investigated. A main research question was investigating the impact of incorrect routing decision on ability estimate. A simulation was performed using simulated students generated based on real data. Students that should have been routed to extreme difficulty categories (i.e., Low/High) were intentionally routed to opposite difficulty categories (i.e., High/Low). Several performance measures were calculated including ability estimation performance measure (RMSE), reliability when a routing decision is made, and performance-level classification measures (e.g., accuracy). Comparisons between simulated data and live data were made to evaluate the extent that the routing behavior in the simulated data resembles the routing behavior of live test administration, so that conclusions from the current study can be applicable to real data. The study found that if routing errors occurred at stage 1, the student theta was generally recovered in stage 2.

Special Study 6: Investigating Reporting Options for Student Within-Year Growth

This study reviewed the literature on growth models and conducted a number of investigations to identify growth models that may be useful for TTAP. The literature review uncovered prior findings that suggested that growth measures are too unreliable to be used for individual student

reporting. Growth curves were fit to TTAP test scores revealing the average rate of change for all tests combined was about 0.26 standard deviations per 40 days of instruction. Science and Mathematics at Grade 6 showed relatively higher within-year growth than Mathematics Grade 7 or Social Studies Grade 8. Achievement was positively associated with learning gains for all tests, especially so for Mathematics, implying that prior knowledge plays a bigger role in learning Mathematics than the other subject tests. To examine the sensitivity of student growth scores to real growth, simulations were conducted using the fitted growth curve model to define true growth. Three growth models were compared in terms of bias and correlations with the simulated growth scores: simple Gain Score, Gains in Performance Levels (Gain PL), and Student Growth Percentiles (SGP).

All three methods showed similar results, although Gain PL showed lower correlations likely due to restriction of range. Simulations revealed that 75% of large growth effects ($\geq$ .80) could be detected used a simple Gain Score, and 62% of high growth simulees ($\geq$80) could be detected using SGPs. These rates of detection indicate that growth measures are not sensitive to individual growth of typical magnitude. The standard error of measurements (SEMs) of the growth measures in this study were so large relative to the growth in each metric, that most of the modestly sized but real change was obscured by measurement noise. These findings replicate prior research findings. Among the three methods examined, there was not a clear winner. It is recommended that TTAP reporting utilize confidence intervals when reporting individual student simple Gain Scores to avoid confusing measurement noise with true change.

Special Study 7: Evaluating the Texas Through-Year Assessment Pilot Participation Efficacy

This study aimed to investigate whether student participation in TTAP impacted corresponding STAAR assessment performances, and the extent to which TTAP may have contributed to student growth and academic achievement.

The study used Coarsened Exact Matching (CEM) to match TTAP participants with non-TTAP examinees and applied weighting to generalize findings. Regression-based analyses assessed TTAP's impact on STAAR outcomes, with a focus on both efficacy and residual gain scores.

The findings suggest that TTAP participation was statistically significant in grade 6 mathematics and grade 8 social studies, with TTAP examinees generally outperforming expectations, while matched non-TTAP examinees underperformed expectations. When controlling for students' prior grade STAAR Reading Language Arts (RLA) and mathematics performance, grade 6 and 8 TTAP participants scored an average of 7 and 26 scale score points higher than expected on Spring 2024 STAAR, respectively. However, the practical effect size of TTAP participation, measured by ω^2, across all subjects accounted for less than 1% of the total scale score variability in each STAAR subject.

Grade 7 mathematics and grade 5 science showed mixed results, suggesting that TTAP participation in these subjects may not be as uniformly efficacious.

Currently, TTAP functions as a cross-sectional pilot program without full contextual implementation, such as standardized intervention processes. The study found significant effects in specific contexts, particularly in grades 6 and 8, which are critical transitional years. However, the findings in other subjects, like grade 5 science, were inconsistent, suggesting the need for further research. It is not known the extent to which matched non-TTAP examinees were administered other interim assessments.

References

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland, & H. Wainer (Eds.), *Differential item functioning* (pp. 35–66). Lawrence Erlbaum Associates. https://onlinelibrary.wiley.com/doi/pdf/10.1002/j.2333-8504.1992.tb01440.x#:~:text=The%20Mantel%2DHaenszel(MH),related%20procedures%20is%20then%20presented

Gianopulos, G., Rozunick, C., & Schneider, C. (2024, April). *A Comparison of Through-year Cumulative Scoring Methods*. NCSA 2024 Convention, Seattle, Wa., United States.

Kolen, M.J., & Brennan, R.L. (2014). *Test equating, scaling, and linking* (3rd ed.). Springer.

Lee, W.-C. (2010). Classification consistency and accuracy for complex assessments using item response theory. *Journal of Educational Measurement*, *47*(1), 1–17.

Li, S. (2006). Evaluating the consistency and accuracy of proficiency classifications using item response theory. *Doctoral Dissertations 1896 – February 2014.* 5761. Available online: https://scholarworks.umass.edu/dissertations_1/5761

Linacre, J. M. (2023). WINSTEPS® Rasch Measurement [Computer software]. Retrieved from winsteps.comMasters, G. N., & Wright, B. D. (1997). The partial credit model. In W. J. van der Linden, & R. K. Hambleton (Eds.), *Handbook of modern item response theory*. Springer.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, *22*(4), 719–748. https://doi.org/10.1093/jnci/22.4.719

Masters, G.N., & Wright, B.D. (1997). The Partial Credit Model. In van der Linden, W.J., & Hambleton, R.K. (Eds.) *Handbook of Modern Item Response Theory.* Springer, New York, NY. https://doi.org/10.1007/978-1-4757-2691-6_6

Michaelides, M. P. (2008). An illustration of a Mantel-Haenszel procedure to flag misbehaving common items in test equating. *Practical Assessment Research and Evaluation*, *13*(7). https://doi.org/10.7275/n04d-8767

Perie, M., Marion, S., & Gong, B. (2009). Moving toward a comprehensive assessment system: A framework for considering interim assessments. *Educational Measurement: Issues and Practice*, *28*(3), 5–13.

Rudner, L. M. (2000). Computing the expected proportions of misclassified examinees. Practical Assessment, Research & Evaluation, 7(14). Available online: https://scholarworks.umass.edu/cgi/viewcontent.cgi?article=1097 &context=pare

Rudner, L. M. (2005). Expected classification accuracy. Practical Assessment, Research & Evaluation, 10(13). Available online: https://scholarworks.umass.edu/pare/vol10/iss1/13/

Samejima, F. (1977). A use of the information function in tailored testing. *Applied Psychological Measurement*, *1*(2), 233–247.

Samejima, F. (1994). Estimation of reliability coefficients using the test information function and its modifications. *Applied Psychological Measurement*, *18*(3), 229–244.

Schneider, M. C., Chen, J., & Nichols, P. D. (2021, July). Using principled assessment design and item difficulty modeling to connect hybrid adaptive instructional and assessment systems: Proof of concept. In *International conference on human-computer interaction* (pp. 149–166). Springer.

Schneider, C., Liu, Y., & Robinson, J. (2022). *STAAR grades 3–8 RLA and mathematics: An interim-to-summative ROC study*. [Unpublished manuscript].

Spearman, C. (1910). Correlation calculated with faulty data. *British Journal of Psychology*, *3*(3), 271–295. https://doi.org/10.1111/j.2044-8295.1910.tb00206.x

Wells, C. S., & Sireci, S. G. (2020). Evaluating random and systematic error in student growth percentiles. *Applied Measurement in Education*, *33*(4), 349–361.

Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer*, *3*(1), 32–35. https://doi.org/10.1002/1097-0142

Appendix A: 2023–2024 TTAP Administration Test Information Functions

Figure A.1: TTAP 2023–2024 Test Information Function (Grade 5 Science)

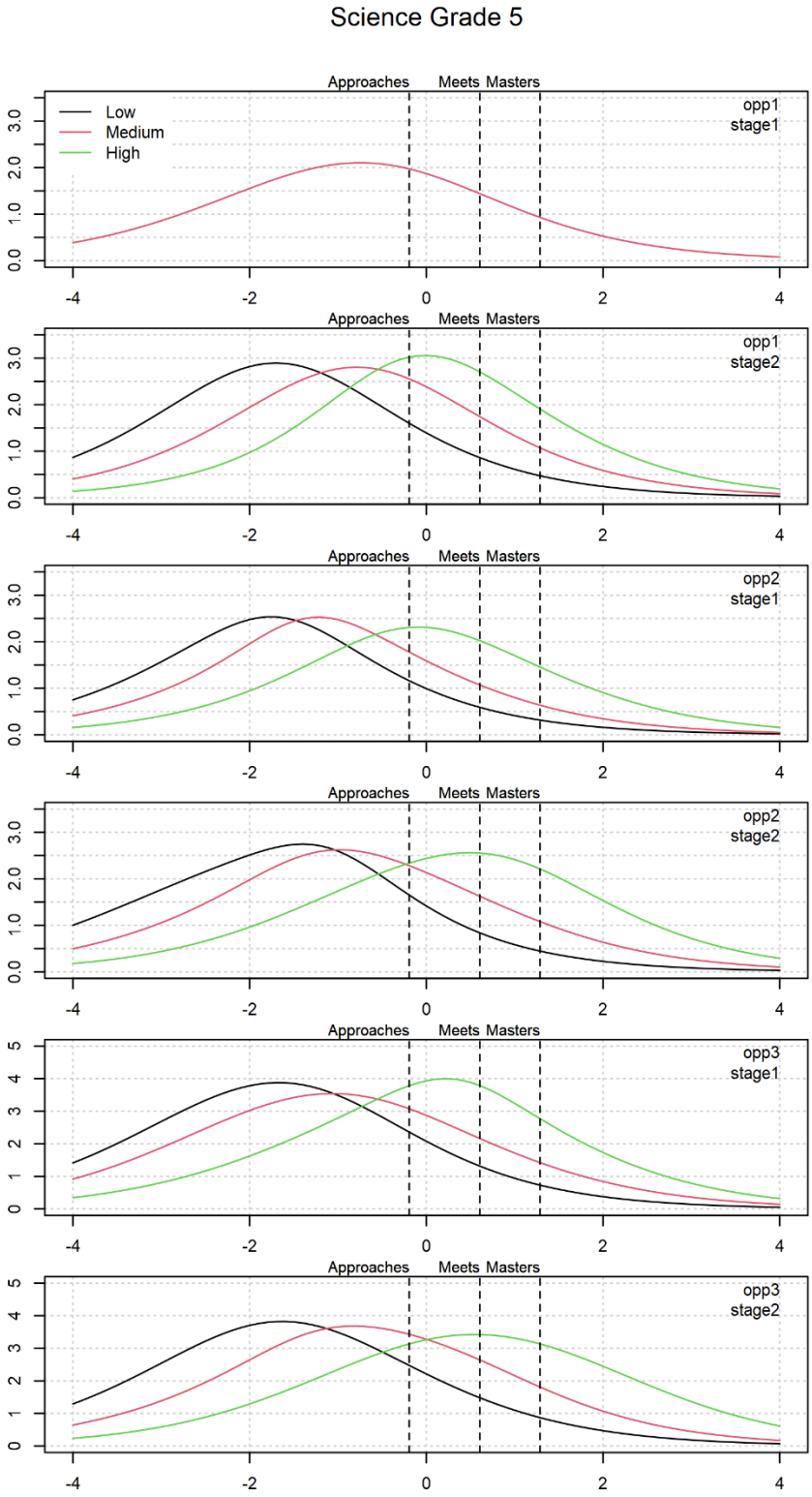


Figure A.2: TTAP 2023–2024 Test Information Function (Grade 6 Mathematics)

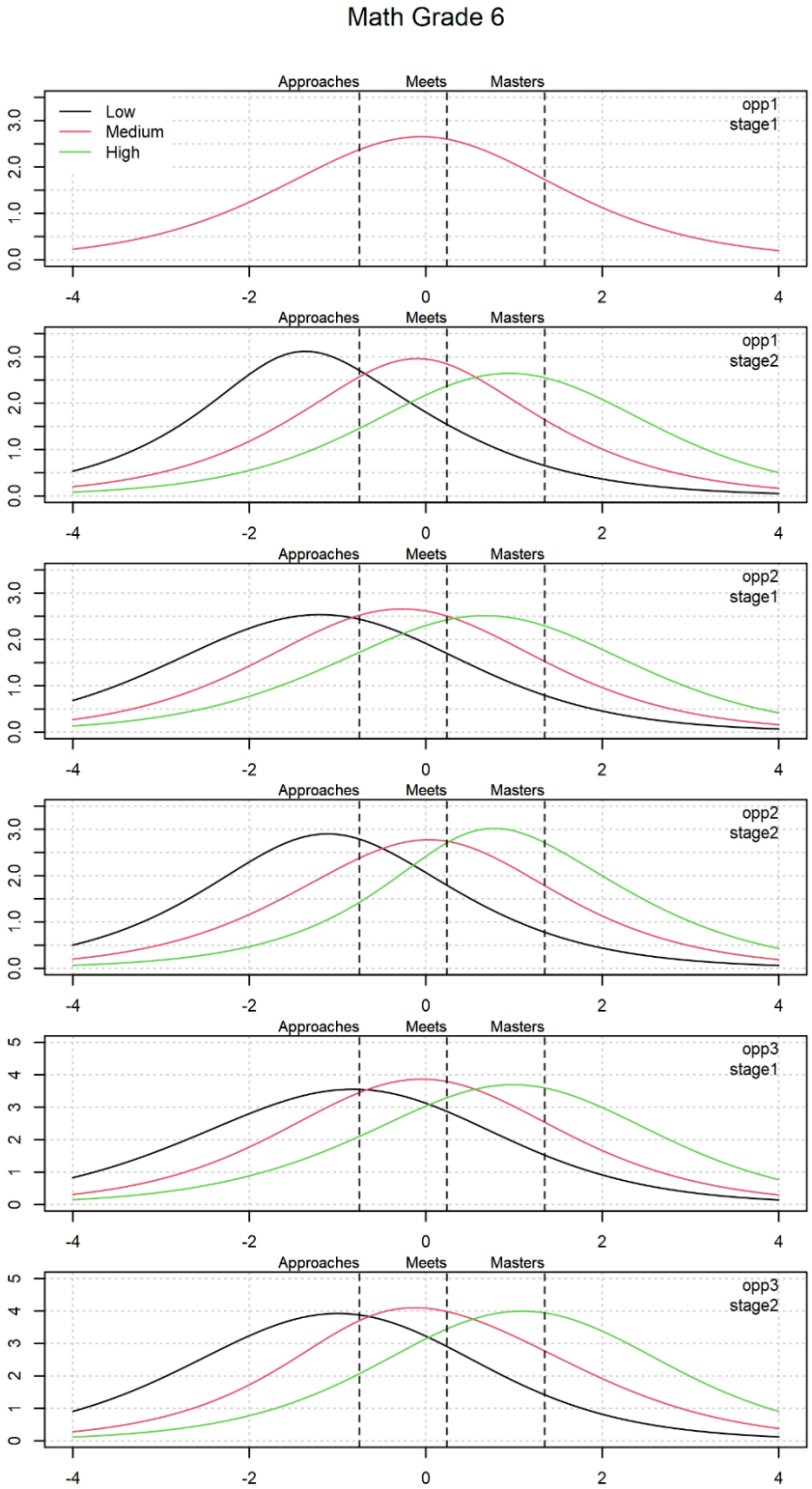


Figure A.3: TTAP 2023–2024 Test Information Function (Grade 7 Mathematics)

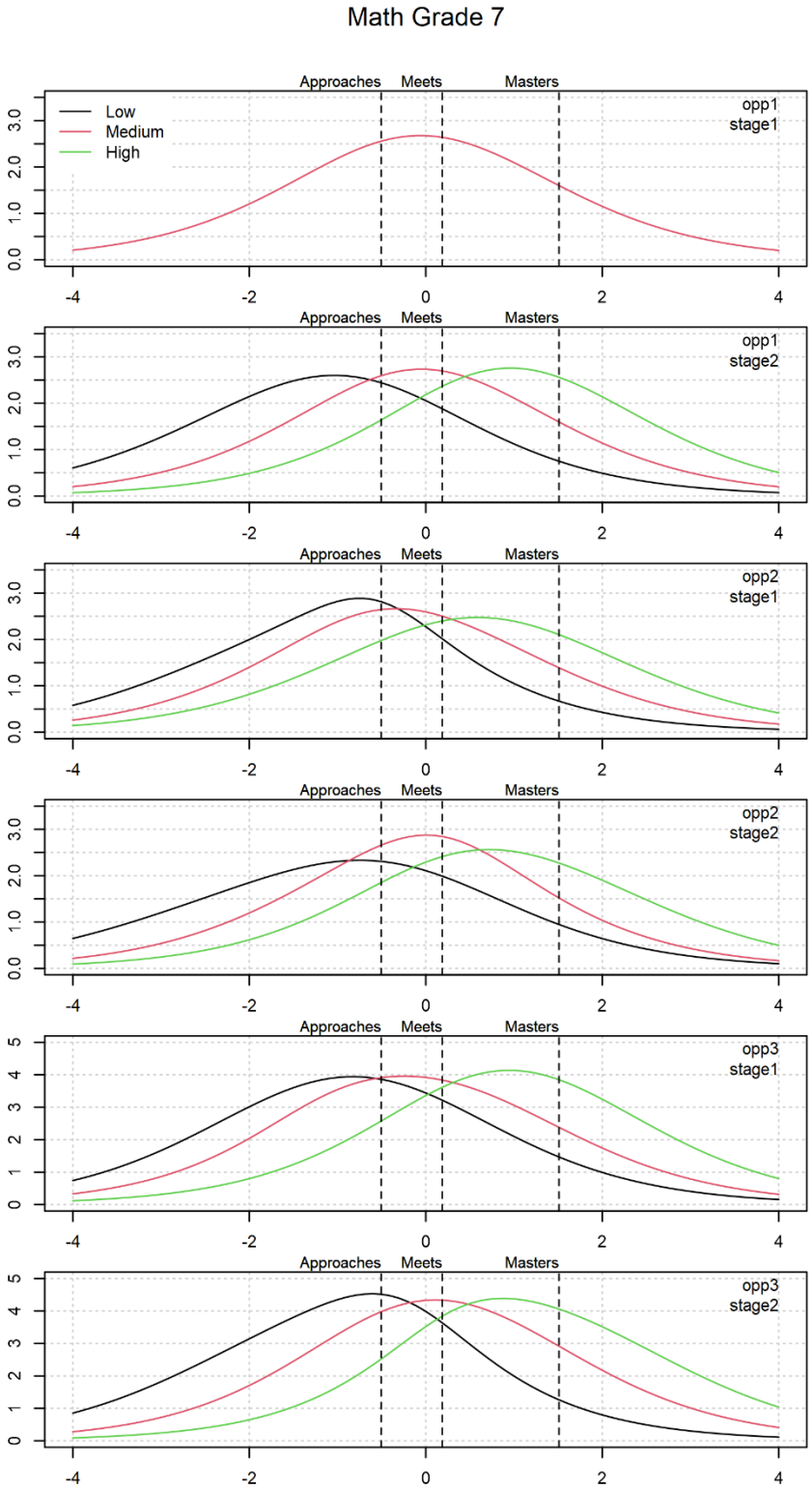
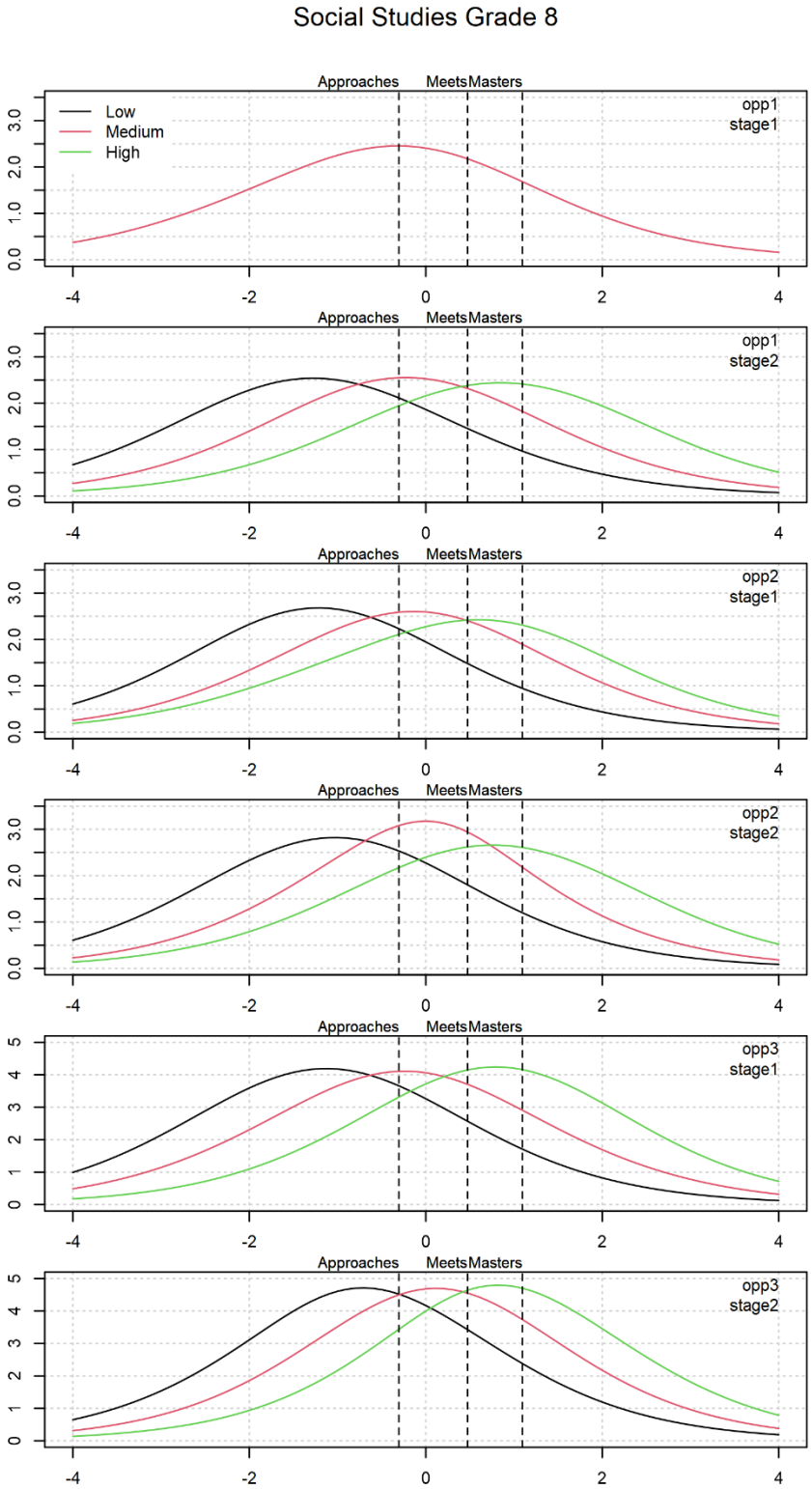


Figure A.4: TTAP 2023–2024 Test Information Function (Grade 8 Social Studies)



Appendix B: Data Cleaning and Merging

a) TTAP Data Files

The following cleaning rules were applied for the TTAP Database of Record (DOR) data files within each opportunity. Appendix B includes a data dictionary to explain each exclusion variable, possible values, and rules applied for inclusion or exclusion.

- Keep students with appropriate test status values
  - Using the variable "*status*," include values of "*scored*" and "*completed*"
- Remove students who have not attempted the test
  - Using the variable "*Overall_Attempted*," keep values of "*Y*"
- Remove private schools
  - Using "*RTS_REGION_EXTERNALID*," keep values between 1 and 20
    - Private schools are denoted under a region identifier with a value of 21
    - Demo schools are listed under region 99
- Remove students who tested off-grade
  - For example, for grade 6 mathematics summaries, keep only students with an "*RTS_EnrlGrdCd*" = 6
- Remove demo students
  - Using the variable "IsDemo," keep values of 0
- Separate English and Spanish for grade 5 science
  - For grade 5 science, use the variable "*segment_2_formID*" to determine if the student took an English or Spanish version of the TTAP assessments
- Within a given grade and subject, if a duplicate "*RTS_EXTERNALID*" occurs, keep the first observation.

b) Summative Data Files

The following cleaning rules were applied for the summative assessment data files:

- Remove private schools
  - Using "ESCREGIONNUMBER," keep values between 1 and 20
    - Private schools are denoted under a region identifier with a value of 21
- For grades 3–8, remove students who tested off-grade
  - Use "ENROLLEDGRADE" to select valid grade(s)
- Select language
  - Use "SCIENCELANGUAGEVERSION" to select "E" for English and "S" for Spanish versions for grade 5 science
- Only keep records with a score code of S
  - For grades 5–8
    - Use "SCORECODE-MATHEMATICS" of "S" for valid mathematics records
    - Use "SCORECODE-SOCIALSTUDIES" of "S" for valid social study records

        - Use “SCORECODE-SCIENCE” of “S” for valid science records
- Keep only records with respective DISCREPANCYINDICATOR value of 0
    - Use “DISCREPANCYINDICATORMATHEMATICS” for mathematics
    - Use “DISCREPANCYINDICATORSCIENCE” for science
    - Use “DISCREPANCYINDICATORSOCIALSTUDIES” for social studies
- Remove duplicated records by subject, grade, and student ID number, and keep the first observation.

Once the summative and TTAP data files are cleaned separately, they were merged by student ID (TSDS). CAI used the merged data files to generate the statistics for the TTAP technical report.

Appendix C: Demographic Variable Recode

The following table indicates the values for each demographic variable used in the summaries and how they were recoded for analyses.

Summative Data Variables	Values/Definitions	Recode for Analysis
SEX-CODE	M = Male, F = Female	M = Male, F = Female
ETHNICITY/RACEREPORTINGCATEGORY	H = Hispanic/Latino I = American Indian or Alaska Native A = Asian B = Black or African American P = Native Hawaiian or Other Pacific Islander W = White T = Two or More Races N = No Information Provided	H = Hispanic/Latino I = American Indian or Alaska Native A = Asian B = Black or African American P = Native Hawaiian or Other Pacific Islander W = White T = Two or More Races N = No Information Provided
ECONOMIC-DISADVANTAGE-CODE	1 = Eligible for free meals under the National School Lunch and Child Nutrition Program, 2 = Eligible for reduced-price meals under the National School Lunch and Child Nutrition Program, 9 = Other economic disadvantage, 0 = Not identified as economic disadvantaged	1, 2, 9 = Economically Disadvantaged 0 = Otherwise

Summative Data Variables	Values/Definitions	Recode for Analysis
TITLE-I-PART-A-INDICATOR-CODE	6 = Student attends campus with schoolwide program, 7 = Student participates in program at targeted assistance school, 8 = Student is a previous participant in the program at a targeted assistance school (not a current participant), 9 = Student does not attend a Title I, Part A school but receives Title I, Part A services because the student is homeless, 0 = Student does not currently participate in and has not previously participated in the program at current campus	6, 7, 9 = Title-I Part A 0, 8 = Otherwise
MIGRANT-INDICATOR-CODE	1 = Yes 0 = No	1 = Migrant 0 = Otherwise
EMERGENTBILINGUALINDICATORCODE	C - Identified as Emergent Bilingual (EB)/English learner (EL) F - Monitored 1st Year (M1), reclassified from EB/EL S - Monitored 2nd Year (M2), reclassified from EB/EL T - Monitored 3rd Year (M3), reclassified from EB/EL R - Monitored 4th Year (M4), reclassified from EB/EL E - Former EB/EL (Post Monitoring) 0 - Non-Emergent Bilingual (Non-EB)/Non-English learner (Non-EL)	C = Emergent Bilingual 0, E, F, S, T, R = Otherwise
BILINGUAL-INDICATOR-CODE	2 = Transitional bilingual/early exit, 3 = Transitional bilingual/late exit, 4 = Dual language immersion/two-way, 5 = Dual language immersion/one-way, 0 = Student is not participating in a state-	2, 3, 4, 5 = Bilingual 0 = Otherwise

Summative Data Variables	Values/Definitions	Recode for Analysis
	approved full bilingual program	
ESL-INDICATOR-CODE	2 = ESL/content-based, 3 = ESL/pull-out, 0 = Student is not participating in a state-approved ESL program	2, 3 = ESL 0 = Otherwise
SPECIAL-ED-INDICATOR-CODE	1 = Student is participating in a special education program, 0 = Student is not participating in a special education program	1 = Special Ed 0 = Otherwise
GIFTED-TALENTED-INDICATOR-CODE	1 = Yes 0 = No	1 = Gifted and Talented 0 = Otherwise
AT-RISK-INDICATOR-CODE	1 = Yes 0 = No	1 = At Risk 0 = Otherwise

Appendix D: DOR Extract Variable Dictionary

DOR Extract Variables	Values/Definitions	Rules for Inclusion/Exclusion
Status	Status of the opportunity. Possible values are completed, submitted, scored, reported, expired, invalidated, and reset.	Keep values of *scored* and *completed.*
Overall_Attempted	Attempted indicates if the student met the attemptedness criteria for the given assessment. Possible values are Y and N (some blanks may occur with certain status values).	Keep values of *Y.*
RTS_REGION_EXTERNALID	Numeric identifier (external ID) for the region to which the student belongs. Private schools are denoted with a region identifier of 21 and demo schools are listed under a region identifier of 99.	Keep values between *1* and *20.*
RTS_EnrlGrdCd	The grade in which a student is registered in TIDE. Possible values are EE, PK, KG, 01, 02, 03, 04, 05, 06, 07, 08, 09, 10, 11, 12, and OS.	For grades 3–8, remove off-grade testers. For end-of-course (EOC), remove *'OS'.*
isDemo	The demo variable indicates if the record is for a demo student or actual student.	Keep values of *0.*

Appendix E: Percentage of Students Routed to Different Paths

Figure E.1: Opp. I, Grade 5 Science, Spanish

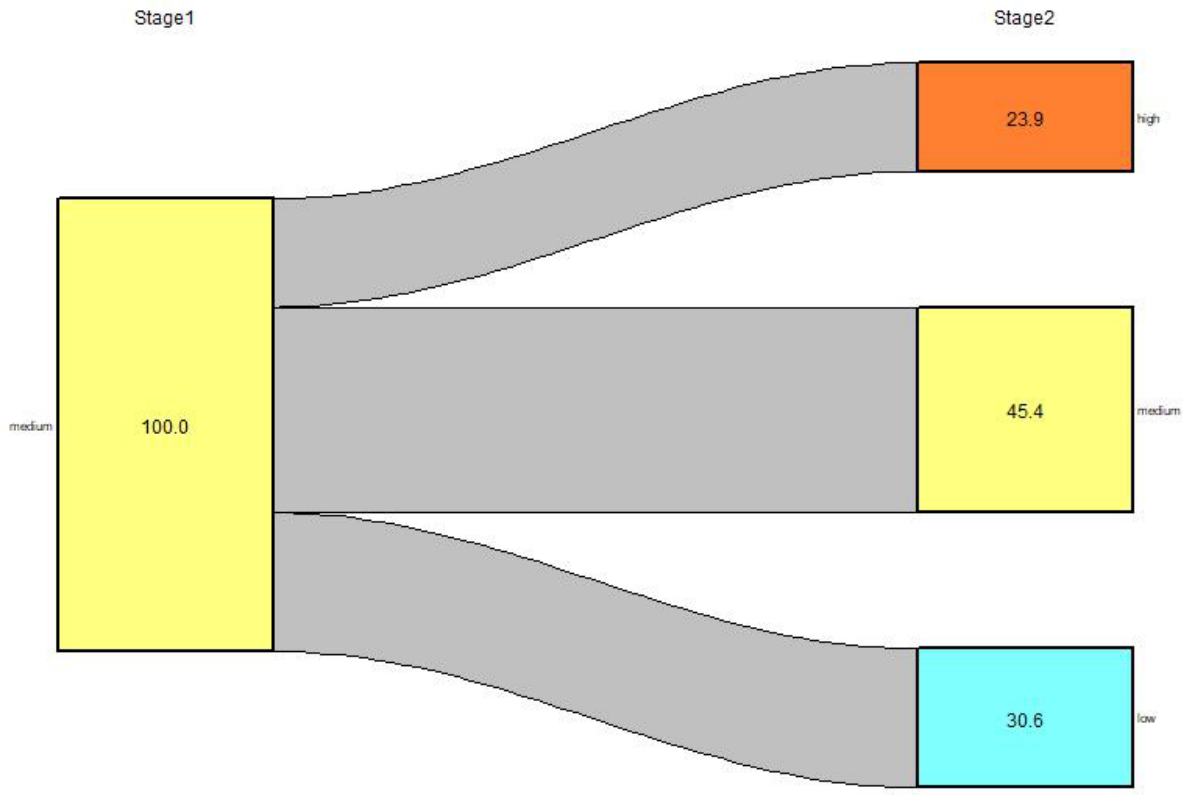


Figure E.2: Opp. I, Grade 5 Science, English

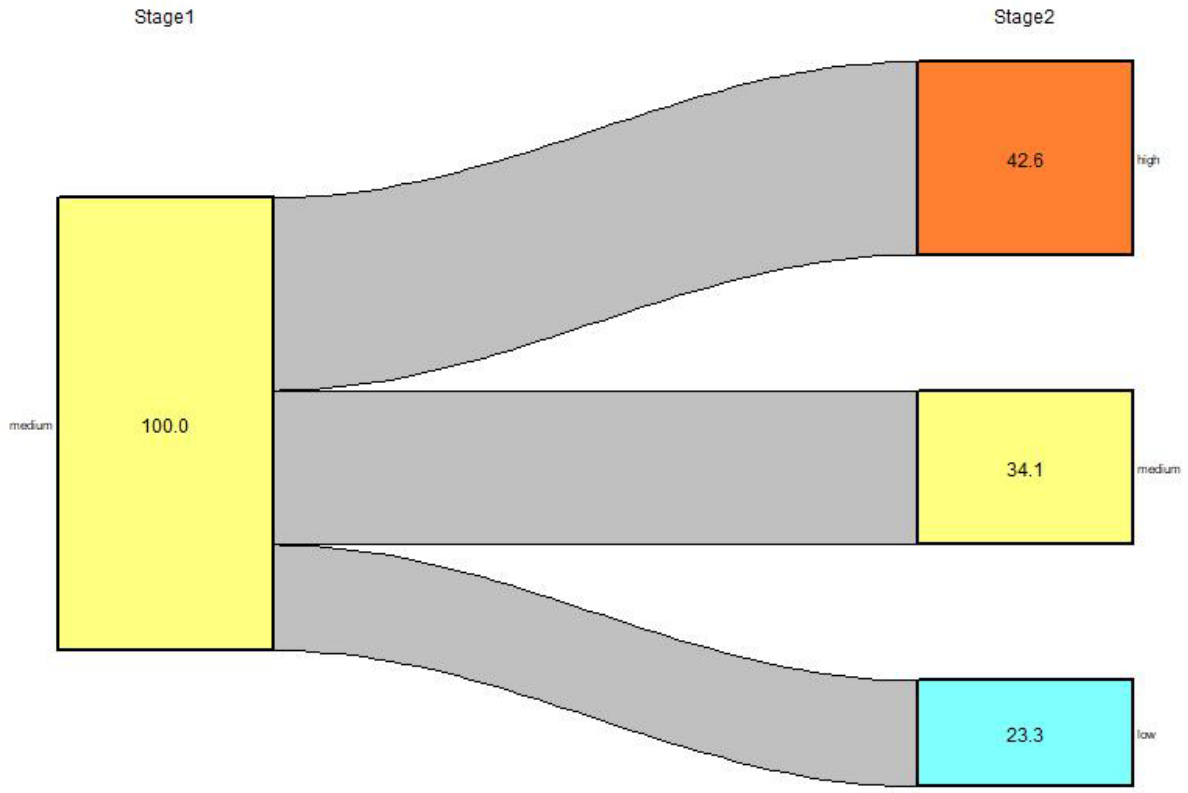


Figure E.3: Opp. I, Grade 6 Mathematics, English

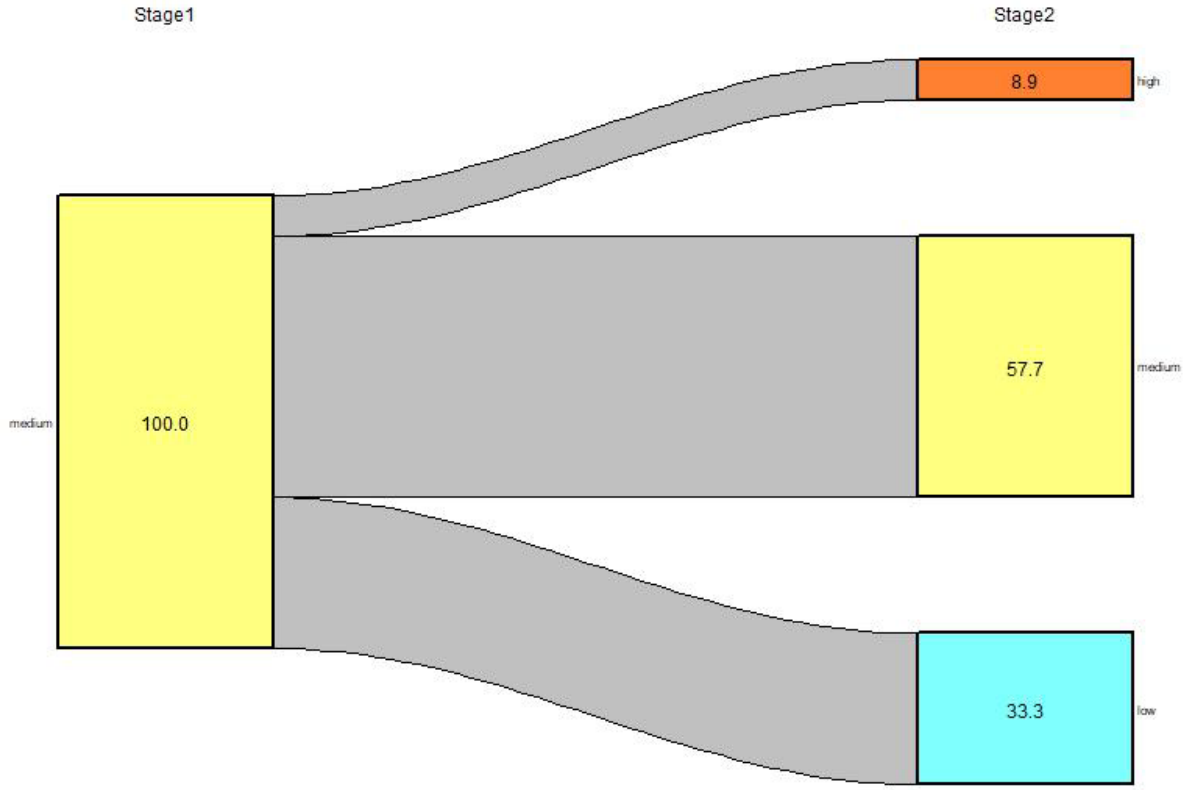


Figure E.4: Opp. I, Grade 7 Mathematics, English

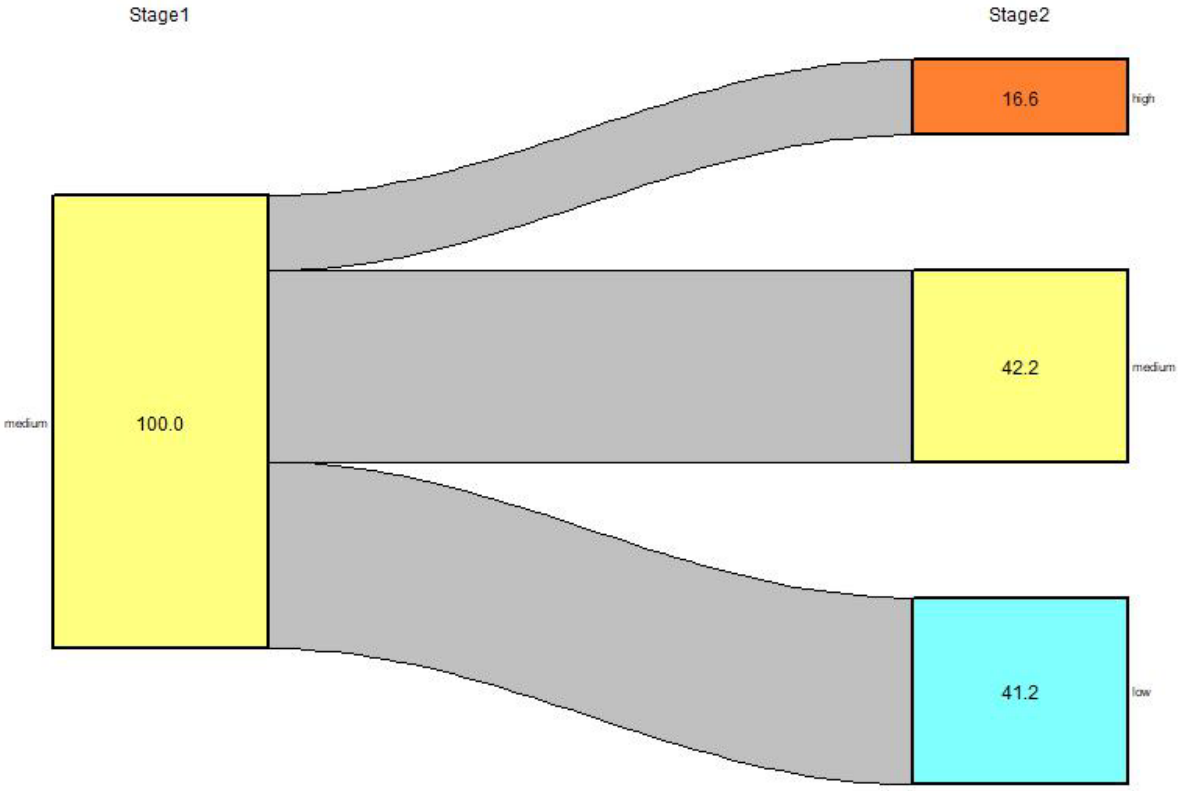


Figure E.5: Opp. I, Grade 8 Social Studies, English

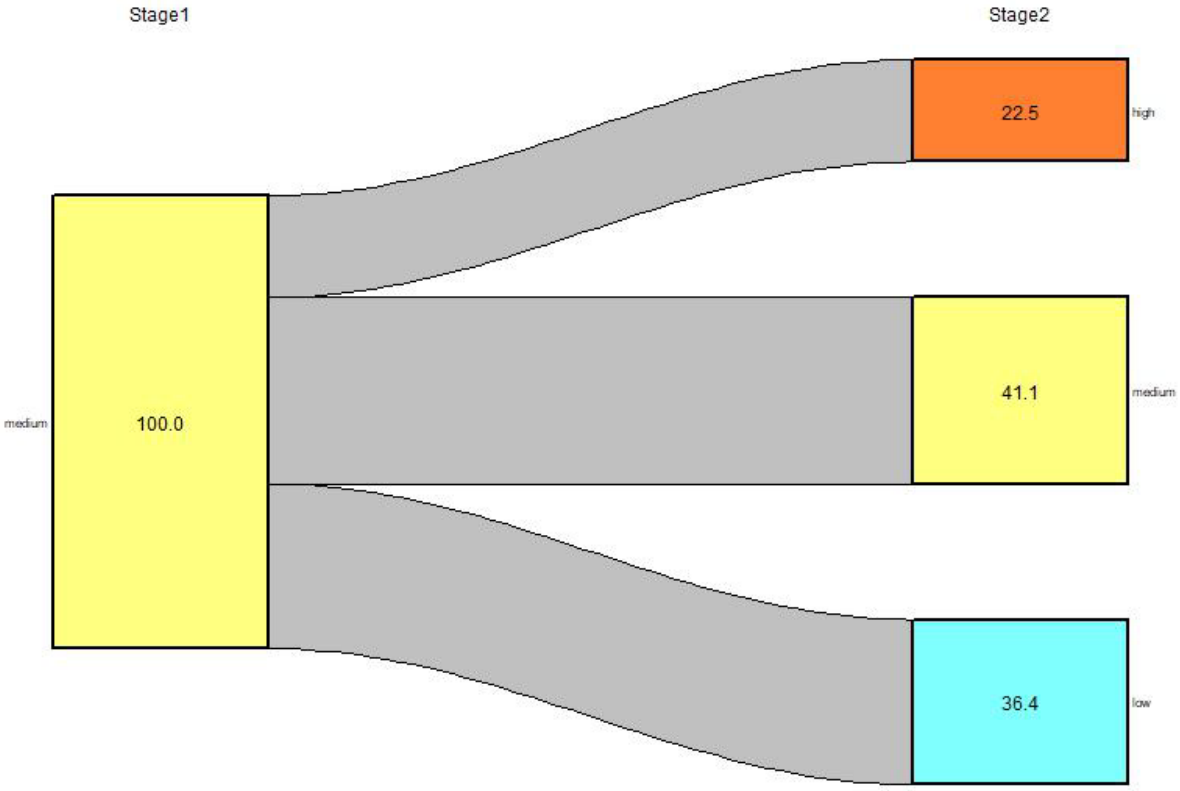


Figure E.6: Opp. II, Grade 5 Science, Spanish

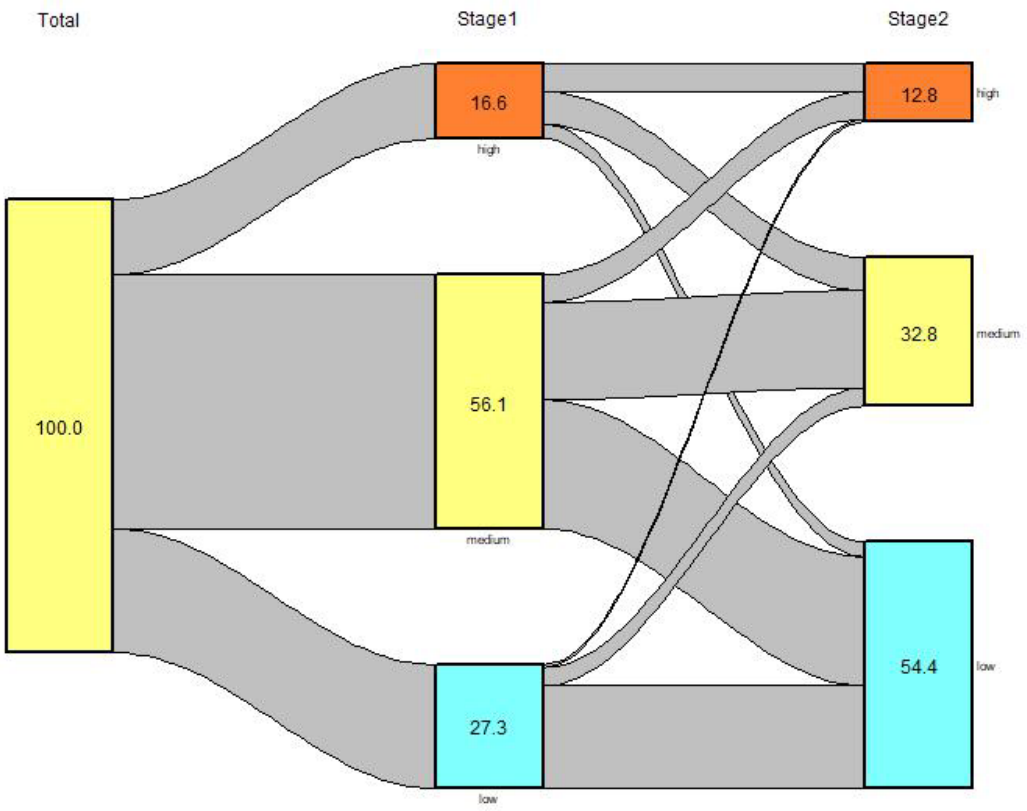


Figure E.7: Opp. II, Grade 5 Science, English

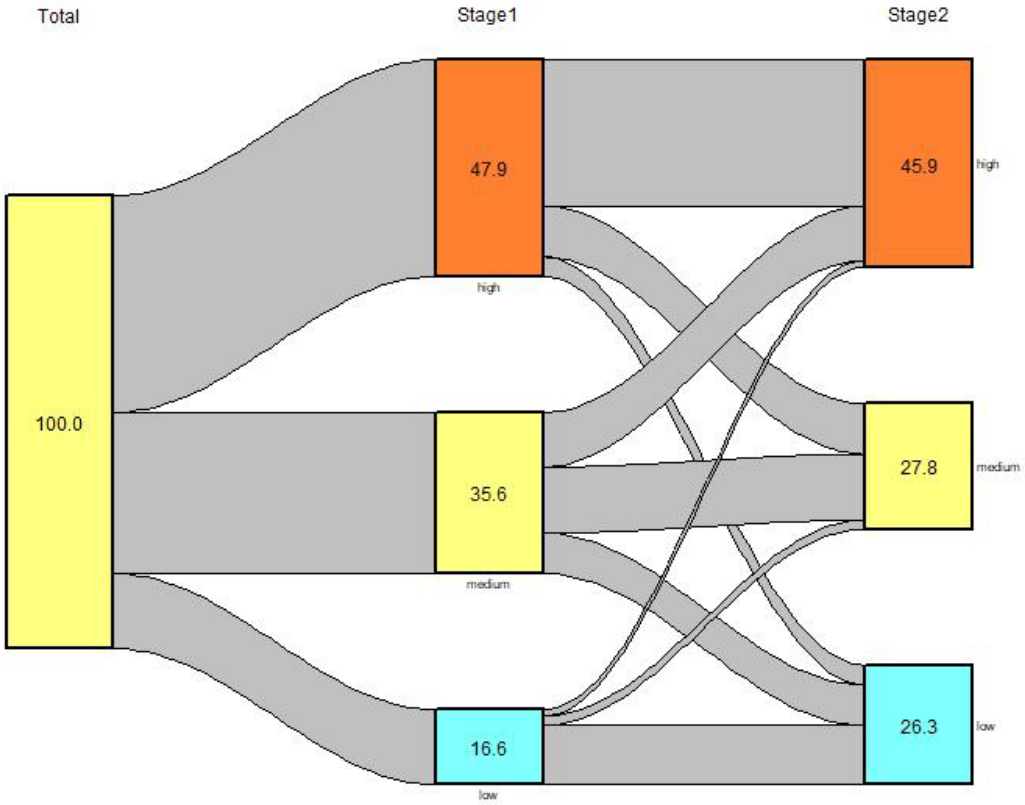


Figure E.8: Opp. II, Grade 6 Mathematics, English

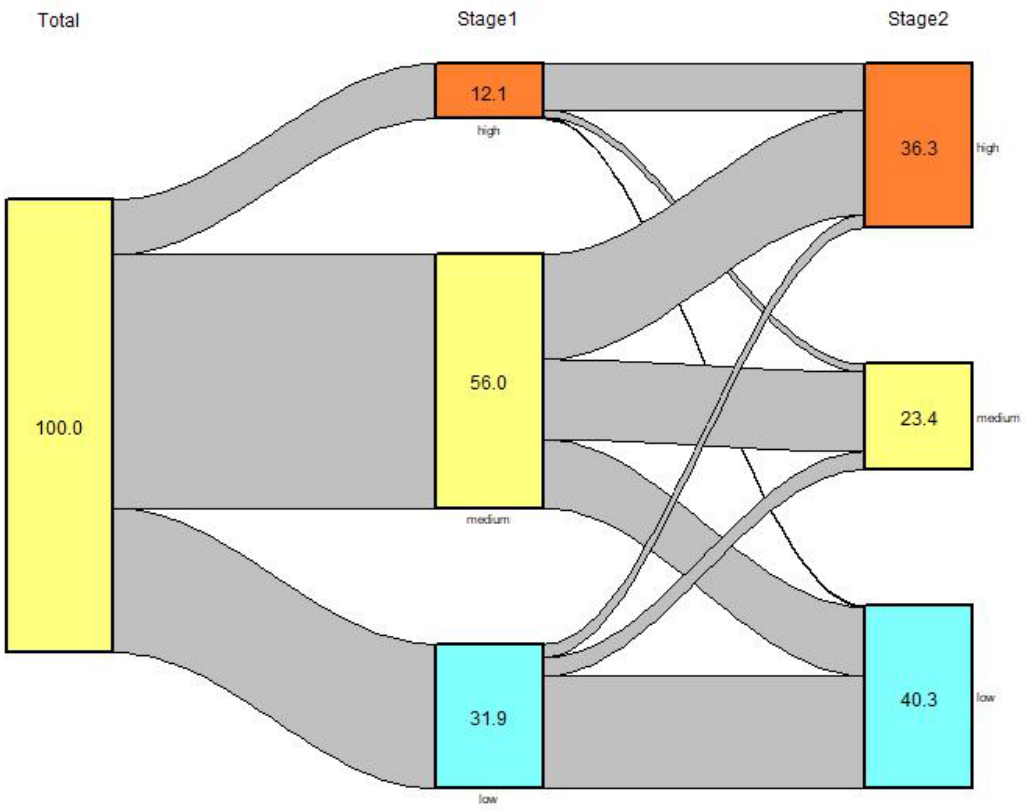


Figure E.9: Opp. II, Grade 7 Mathematics, English

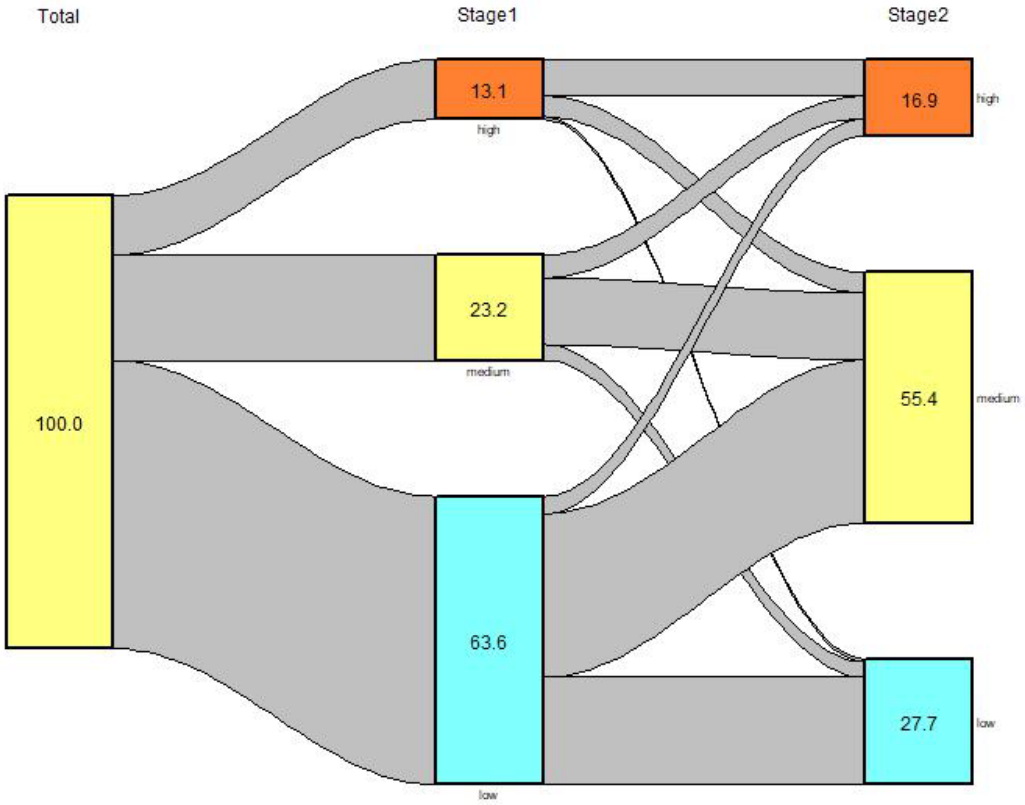


Figure E.10: Opp. II, Grade 8 Social Studies, English

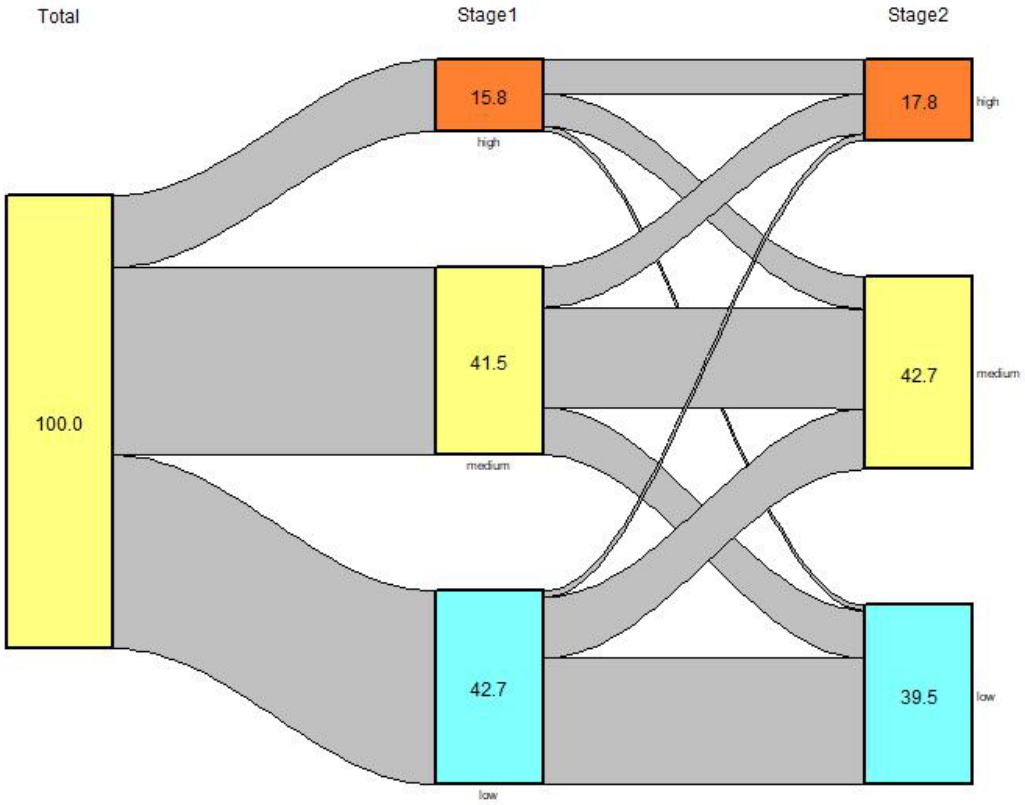


Figure E.11 Opp. III, Grade 5 Science, Spanish

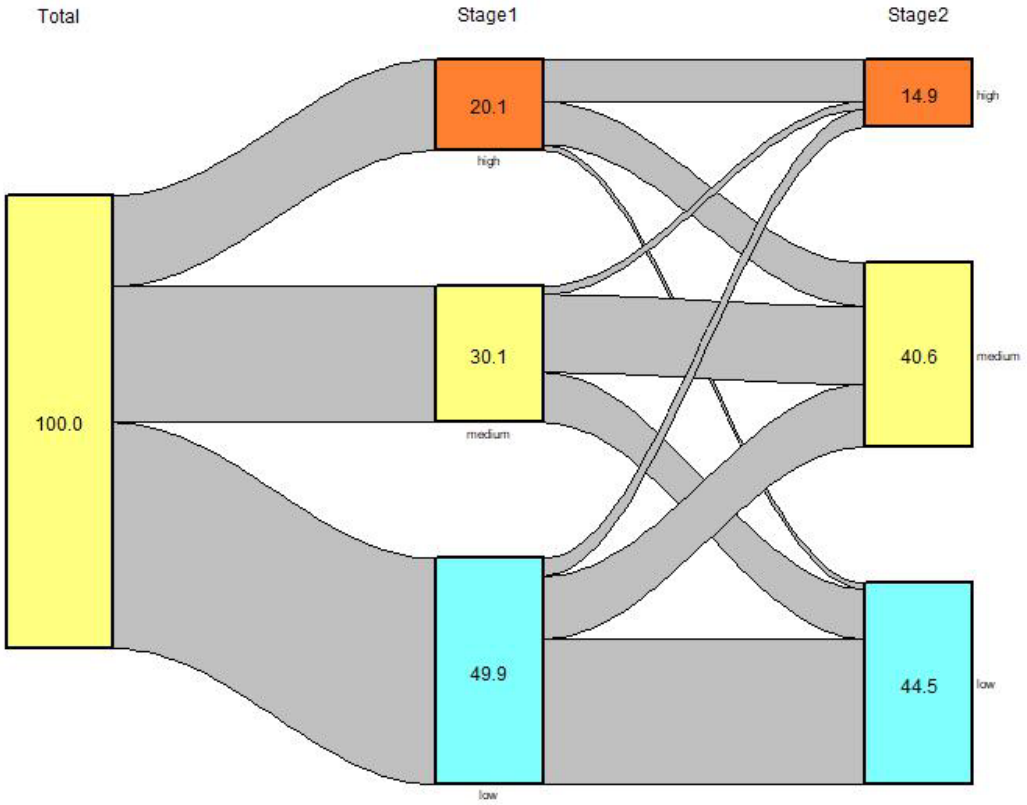


Figure E.12: Opp. III, Grade 5 Science, English

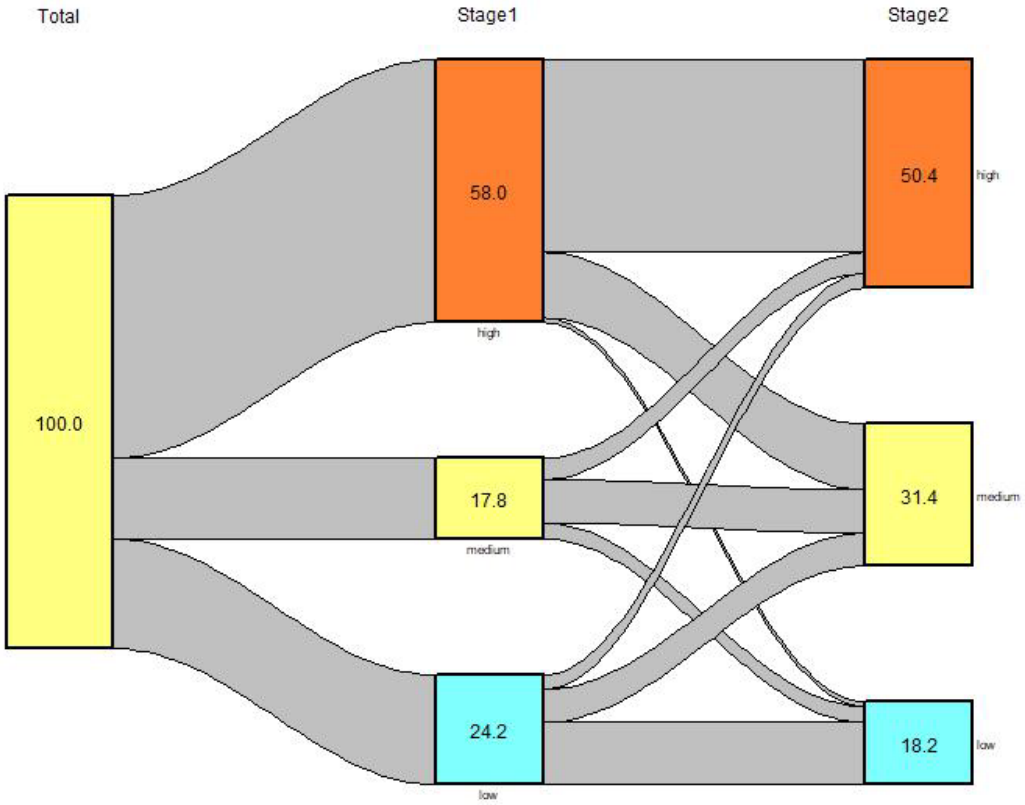


Figure E.13: Opp. III, Grade 6 Mathematics, English

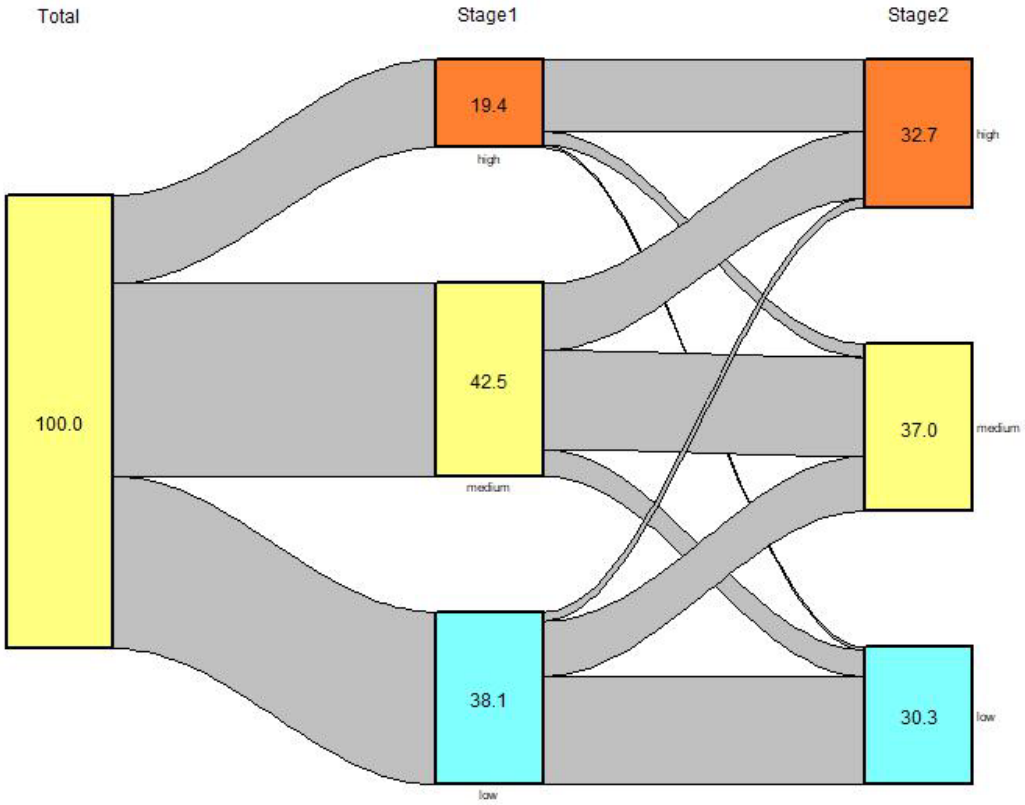


Figure E.14: Opp. III, Grade 7 Mathematics, English

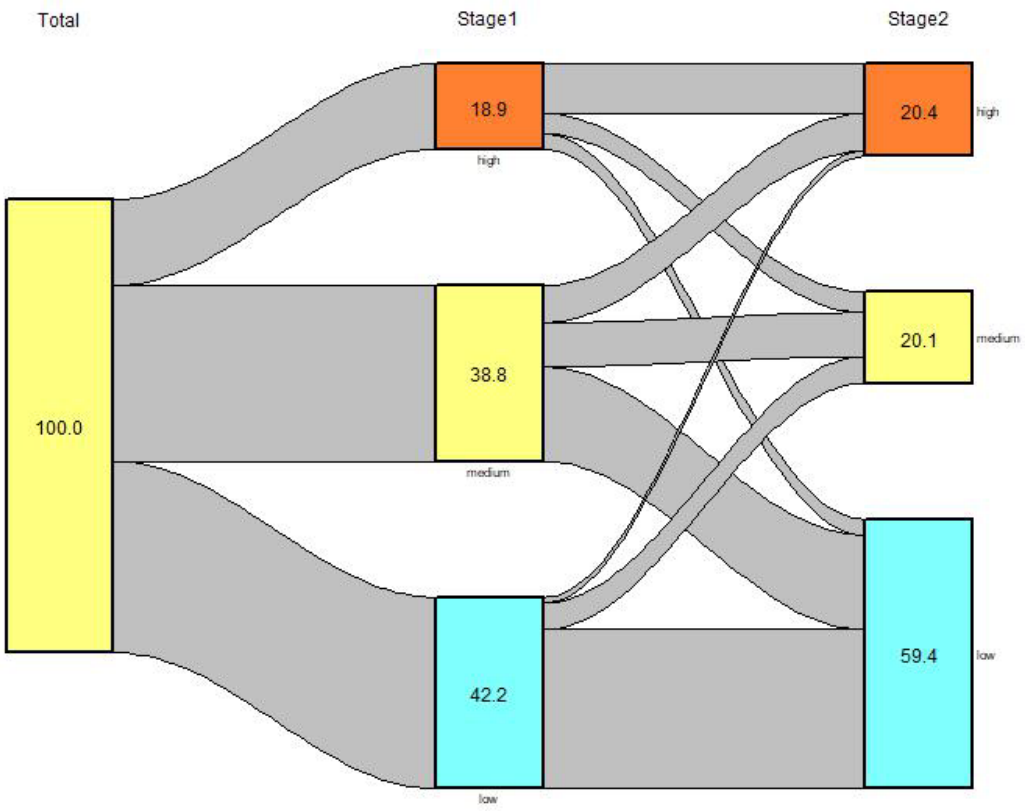


Figure E.15: Opp. III, Grade 8 Social Studies, English