2024–2025 STAAR Through-year Assessment Pilot Technical Report

Table of Contents

1. Introduction 1
1.1 TTAP Intended Uses and Purposes 2
1.2 Test Design and Item Development 3
1.3 Blueprints 4
1.4 2024–25 TTAP Administration 6
1.5 Test Participation 7
1.6 Percentage of Students Taking Different Test Forms 15
2. TTAP Scores from 2024–25 17
2.1 Scaling and Equating 17
2.2 Scale Score Gain/Loss Between Opportunities 18
2.3 TTAP Performance Level 20
3. Reliability 23
3.1 Marginal Reliability 23
3.2 Classification Consistency and Accuracy 25
4. Validity 26
4.1 TTAP and STAAR Correlations 26
4.2 Prediction Agreement 28
5. Fairness 30
6. Reporting 31
6.1 Student-Level Reports 31
6.2 Campus-/District-Level Reports 36
7. Continuous Research and Improvement Plans 38
References 44
Appendix A: 2024–25 TTAP Administration Test Information Functions 46
Appendix B: Data Cleaning and Merging 51

Appendix C: Demographic Variable Recode 53

Appendix D: DOR Extract Variable Dictionary 56

Appendix E: Percentage of Students Routed to Different Paths 57

List of Tables

Table 1: Comparison Between STAAR Summative and TTAP Blueprints 5

Table 2: Summary of Reporting Categories 6

Table 3: 2024–25 STAAR TTAP Assessment Administration Schedules 6

Table 4: TTAP Assessments Administered in the 2024–25 School Year 7

Table 5: TTAP District, Campus, and Unique Student Participation for Each TTAP Assessment in 2024–25 7

Table 6: TTAP Participating Student Demographic Characteristics (Spanish Grade 3 Mathematics) 9

Table 7: TTAP Participating Student Demographic Characteristics (English Grade 3 Mathematics) 10

Table 8: TTAP Participating Student Demographic Characteristics (English Grade 6 Mathematics) 11

Table 9: TTAP Participating Student Demographic Characteristics (English Grade 7 Mathematics) 12

Table 10: TTAP Participating Student Demographic Characteristics (English Grade 8 Mathematics) 13

Table 11: TTAP Participating Student Demographic Characteristics (English Grade 8 Social Studies) 14

Table 12: Percentages of Students Taking Different Test Forms 15

Table 13: Student TTAP Score Growth Across Opportunities 18

Table 14: Effect Size of Student TTAP Scale Score Growth Across Opportunities 19

Table 15: Percentage of Student with Gain, Loss, or No Change TTAP Scale Scores Across Opportunities 20

Table 16: Student Performance Level Distribution Across Opportunities 22

Table 17: Test Reliabilities of TTAP and STAAR....24

Table 18: Classification Consistency and Accuracy....26

Table 19: Pearson Correlation Coefficients Between the TTAP and Summative Assessment Scale Scores....27

Table 20: Prediction Accuracy Summary (Opp. 1)....29

Table 21: Prediction Accuracy Summary (Opp. 2)....29

Table 22: Prediction Accuracy Summary (Opp. 3)....29

Table 23: DIF Classification Rules for Items....31

List of Figures

Figure 1: TTAP Design.... 4

Figure 2: Percentage of Students Routed to Different Paths (Grade 6 Mathematics, English as an example)....17

Figure 3: Individual Student Report (Overall Scores)....33

Figure 4: Individual Student Report (Reporting Category Level Scores)....35

Figure 5: Individual Student Report (Progress Monitoring)....36

Figure 6: District/Campus Report (Scale Score and Performance Level)....37

Figure 7: District/Campus Report (Percentage Correct)....37

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 2022–23 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 overarching 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 web page[1].

Legislative reports were produced in 2023 and 2024 to summarize results from Pilot Year 1 during school year 2022–23 and Pilot Year 2 during school year 2023–24. This technical report provides comprehensive information about the 2024–25 TTAP Assessments (Pilot Year 3), 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 2024–25 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, 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.
2) **TTAP Scores from 2024–25.** This section summarizes performance patterns in students' scale scores, performance levels, percentage correct scores by reporting

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

category and item difficulty level, and 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 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 to 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 2024–25. 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.

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 regarding 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 student's ability, and reduced test anxiety by presenting appropriately challenging items. Such tests provide a customized assessment experience that matches individual abilities and ensures 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.

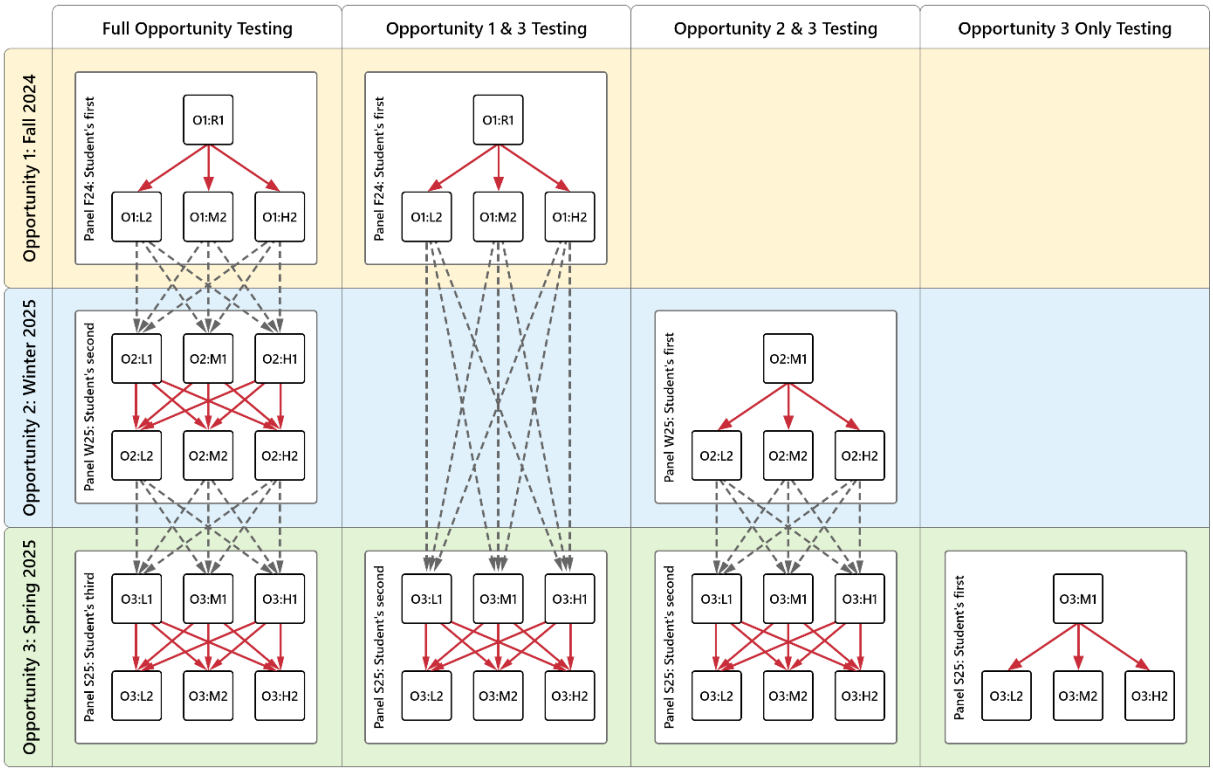
Each TTAP test had three opportunities administered in the 2024–25 school year. Each opportunity was a multistage assessment with two panels (stages). The multistage adaptive test is depicted in Figure 1. At Opportunity 1 (hereafter referred to as Opp. 1), students take a router form and then are routed to a form at the correct level of difficulty. In Opportunity 2 or 3 (hereafter referred to as Opp. 2 and Opp. 3), 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. 2 and Opp. 3 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 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 performance level difficulties: Does Not Meet, Approaches, Meets, and Masters, based on item parameters. These categories are then used in student-, campus-, and district-level reports to provide detailed insights into performance. The classification of items into performance levels is determined using item response theory (IRT). In IRT, items and students can be placed onto the same ability scale. Items at a student's performance level are 'just right': challenging enough to learn, not enough to frustrate.

[2] https://tea.texas.gov/student-assessment/reports-and-studies

Figure 1: TTAP Design



O1: Opp. 1, O2: Opp. 2, O3: Opp. 3
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 3 Mathematics	STAAR Redesign	7	11	7	5	30
	Through Year Opp. 1	7	11	5	3	26
	Through Year Opp. 2	7	11	5	3	26
	Through Year Opp. 3	7	11	5	3	26
	Through Year Total Counts	21	33	15	9	78
Grade 6 Mathematics	STAAR Redesign	9	13	6	8	36
	Through Year Opp. 1	7	11	6	6	30
	Through Year Opp. 2	7	11	6	6	30
	Through Year Opp. 3	7	11	6	6	30
	Through Year Total Counts	21	33	18	18	90
Grade 7 Mathematics	STAAR Redesign	5	15	11	7	38
	Through Year Opp. 1	4	13	10	5	32
	Through Year Opp. 2	4	13	10	5	32
	Through Year Opp. 3	4	13	10	5	32
	Through Year Total Counts	12	39	30	15	96
Grade 8 Mathematics	STAAR Redesign	3	16	15	6	40
	Through Year Opp. 1	2–3	12–14	11–13	4–6	32
	Through Year Opp. 2	2–3	12–14	11–13	4–6	32
	Through Year Opp. 3	2–3	12–14	11–13	4–6	32
	Through Year Total Counts	6–9	36–42	33–39	12–18	96
Grade 8 Social Studies	STAAR Redesign	17	9	9	5	40
	Through Year Opp. 1	13	8	8	5	34
	Through Year Opp. 2	13	8	8	5	34
	Through Year Opp. 3	13	7–8	8–9	5	34
	Through Year Total Counts	39	24–25	24-25	15	102

Table 2: Summary of Reporting Categories

Assessment	RC1	RC2	RC3	RC4
Grade 3 Mathematics	Numerical Representations and Relationships	Computations and Algebraic Relationships	Geometry and Measurement	Data Analysis and Personal Financial Literacy
Grade 6 Mathematics	Numerical Representations and Relationships	Computations and Algebraic Relationships	Geometry and Measurement	Data Analysis and Personal Financial Literacy
Grade 7 Mathematics	Probability and Numerical Representations	Computations and Algebraic Relationships	Geometry and Measurement	Data Analysis and Personal Financial Literacy
Grade 8 Mathematics	Numerical Representations and Relationships	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 2024–25 TTAP Administration

The 2024–25 TTAP assessments include three test opportunities. Table 3 represents TTAP assessment scopes and administration schedules.

Table 3: 2024–25 STAAR TTAP Assessment Administration Schedules

Content	Language	Grade	Opp. 1	Opp. 2	Opp. 3
Mathematics	Spanish	3	November 11–15, 2024	January 27–31, 2025	March 24–28, 2025
Mathematics	English	3			
Mathematics	English	6			
Mathematics	English	7			
Mathematics	English	8			
Social Studies	English	8			

In the 2024–25 school year, more than 145,000 TTAP assessments were administered. Table 4 reports 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 testers and students who did not meet attemptedness rules. A comprehensive list of these exclusion rules can be found in Appendix D. It is worth noting that the number of students who took grade 3 mathematics (Spanish) is relatively small, which could

potentially limit the interpretability of results. In contrast, the other four tests all have sample sizes exceeding 6,000, ensuring that meaningful results can be derived from the data.

Table 4: TTAP Assessments Administered in the 2024–25 School Year

Assessment	Grade	Opp. 1 (N)	Opp. 2 (N)	Opp. 3 (N)	Total N Took All Three Opps.	Total N Took at Least One Opp.
Mathematics (Spanish)	G3	197	202	224	167	246
Mathematics (English)	G3	6,775	6,533	6,667	6,073	7,131
Mathematics (English)	G6	12,203	11,543	11,681	10,799	12,775
Mathematics (English)	G7	7,732	7,460	7,566	6,888	8,191
Mathematics (English)	G8	6,167	6,057	6,085	5,268	6,812
Social Studies (English)	G8	16,691	15,857	16,198	14,702	17,562
Total		49,765	47,652	48,421	43,897	52,717

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 2024–25 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 2024–25

Assessment	Number of Districts	Number of Campuses	Number of Unique Students
Mathematics (Spanish) G3	12	30	246
Mathematics (English) G3	56	110	7,131
Mathematics (English) G6	60	112	12,775
Mathematics (English) G7	60	103	8,191
Mathematics (English) G8	51	96	6,812
Social Studies (English) G8	72	130	17,562
Total	84	231	46,600

In addition, the demographic characteristics of the 2024–25 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 2025 and those who participated in at least one TTAP assessment are presented in Table 6 through Table 11. For ease of reference in our analyses, the variable names and mapping can be found in Appendices C and D.

There are notable demographic differences between the students who took the Spanish version of the TTAP grade 3 mathematics assessment to those who took the English version. However,

it is important to acknowledge that the TTAP sample size for this assessment in Spanish is relatively small, which limits the significance of direct comparisons. For the other four assessments, in many demographic comparisons, the percentages within each category are close, with differences generally below 5%. However, there are exceptions. 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 higher representation of white students and a lower representation of Hispanic students in the mathematics and grade 8 Social Studies assessment of the TTAP sample compared to those who took STAAR.
2. There are slightly lower percentages of current limited English proficiency students in all TTAP samples except for grade 3 Spanish.
3. Across all tests there is an under-representation of economically disadvantaged testers, and in all grades except grade 3 Spanish and grade 8 Social Studies, Title 1 students are under-represented.

These observations provide insights into the demographic composition of TTAP assessment participants in relation to the broader student population, even though the differences are generally small. During sample selection, these differences did not go undetected, so sample sizes were increased beyond the original target of 5,000 to allow sampling down to a subset that falls within the 5% tolerance. For the purpose of this report, all tables are based on the full sample; in special studies that may be sensitive to sample representativeness, adjustments to the sample may be conducted to improve the representativeness of the sample to the STAAR population.

Table 6: TTAP Participating Student Demographic Characteristics (Spanish Grade 3 Mathematics)

Demographic	STAAR Spring 2025	TTAP 2024–25	Difference in Percentage
Number of Students	21,319	246	NA
Male	49.3	46.7	2.6
Female	50.6	53.3	2.7
Hispanic/Latino	97.8	97.2	0.6
American Indian or Alaska Native	0.3	0	0.3
Asian	0	0	0
Black or African American	0.1	0.4	0.3
Native Hawaiian or Pacific Islander	0	0.4	0.4
White	1.4	1.6	0.2
Two or More Races	0.1	0	0.1
Economically Disadvantaged	**87.4**	**74**	**13.4**
Title I, Part A Participants	94.1	94.3	0.2
Migrant	0.3	0	0.3
Current Limited English Proficient	97.8	99.6	1.8
Bilingual	83.7	86.6	2.9
English for Speakers of Other Languages (ESL) Participants	1.4	4.9	3.5
Special Education	10.8	9.3	1.5
Gifted/Talented Participants	5.6	0.8	4.8
At-Risk	95.3	97.2	1.9

Table 7: TTAP Participating Student Demographic Characteristics (English Grade 3 Mathematics)

Demographic	STAAR Spring 2025	TTAP 2024–25	Difference in Percentage
Number of Students	376,506	7,131	NA
Male	50.8	51.1	0.3
Female	49.2	48.9	0.3
Hispanic/Latino	**50.5**	**37.5**	**13**
American Indian or Alaska Native	0.3	0.3	0
Asian	6.1	10.1	4
Black or African American	13.1	9.8	3.3
Native Hawaiian or Pacific Islander	0.2	0.1	0.1
White	**25.9**	**36.7**	**10.8**
Two or More Races	3.7	5.2	1.5
Economically Disadvantaged	**59.7**	**47.5**	**12.2**
Title I, Part A Participants	**73.5**	**64.2**	**9.3**
Migrant	0.3	0.2	0.1
Current Limited English Proficient	**22.9**	**17.3**	**5.6**
Bilingual	**11.9**	**5.5**	**6.4**
ESL Participants	7	8.8	1.8
Special Education	20.5	19.9	0.6
Gifted/Talented Participants	10.4	11.2	0.8
At-Risk	47.7	34.7	13

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

Demographic	STAAR Spring 2025	TTAP 2024–25	Difference in Percentage
Number of Students	386,208	12,778	NA
Male	50.7	49.9	0.8
Female	49.3	50	0.7
Hispanic/Latino	**53.6**	**44.4**	**9.2**
American Indian or Alaska Native	0.3	0.4	0.1
Asian	5.4	6.9	1.5
Black or African American	12.6	12.5	0.1
Native Hawaiian or Pacific Islander	0.2	0.2	0
White	**24.5**	**31.6**	**7.1**
Two or More Races	3.1	3.7	0.6
Economically Disadvantaged	**60.9**	**52.2**	**8.7**
Title I, Part A Participants	**65.4**	**58.3**	**7.1**
Migrant	0.3	0.2	0.1
Current Limited English Proficient	23.6	22.8	0.8
Bilingual	2.9	1	1.9
ESL Participants	15.4	14	1.4
Special Education	17.4	16.9	0.5
Gifted/Talented Participants	10.9	10.6	0.3
At-Risk	53.8	53.6	0.2

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

Demographic	STAAR Spring 2025	TTAP 2024–25	Difference in Percentage
Number of Students	292,856	8,195	NA
Male	50.5	50.9	0.4
Female	49.5	49.1	0.4
Hispanic/Latino	**54.5**	**44.8**	**9.7**
American Indian or Alaska Native	0.3	0.3	0
Asian	4.5	5	0.5
Black or African American	13.6	11	2.6
Native Hawaiian or Pacific Islander	0.2	0.2	0
White	**23.5**	**34.7**	**11.2**
Two or More Races	2.9	3.7	0.8
Economically Disadvantaged	**63.8**	**52.4**	**11.4**
Title I, Part A Participants	**63.7**	**55.5**	**8.2**
Migrant	0.3	0	0.3
Current Limited English Proficient	26.3	21.3	5
Bilingual	0.8	0.8	0
ESL Participants	19.4	18.2	1.2
Special Education	18	18.4	0.4
Gifted/Talented Participants	5.9	9.6	3.7
At-Risk	**59.6**	**53.4**	**6.2**

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

Demographic	STAAR Spring 2025	TTAP 2024–25	Difference in Percentage
Number of Students	271,045	6,814	NA
Male	51	49.9	1.1
Female	49	50	1
Hispanic/Latino	**55.1**	**43.6**	**11.5**
American Indian or Alaska Native	0.3	0.4	0.1
Asian	3.9	5.9	2
Black or African American	14.3	10.9	3.4
Native Hawaiian or Pacific Islander	0.2	0.2	0
White	**22.9**	**34.8**	**11.9**
Two or More Races	2.9	3.8	0.9
Economically Disadvantaged	**64.3**	**50.8**	**13.5**
Title I, Part A Participants	**63**	**53.1**	**9.9**
Migrant	0.4	0.1	0.3
Current Limited English Proficient	**26.3**	**19.9**	**6.4**
Bilingual	0.7	0.5	0.2
ESL Participants	**21.1**	**15.8**	**5.3**
Special Education	17.6	18.8	1.2
Gifted/Talented Participants	5.1	7.3	2.2
At-Risk	66.1	61.1	5

Table 11: TTAP Participating Student Demographic Characteristics (English Grade 8 Social Studies)

Demographic	STAAR Spring 2025	TTAP 2024–25	Difference in Percentage
Number of Students	401,059	17,566	NA
Male	51.1	50.6	0.5
Female	48.8	49.3	0.5
Hispanic/Latino	53.1	50.2	2.9
American Indian or Alaska Native	0.3	0.4	0.1
Asian	5.8	6.6	0.8
Black or African American	12.7	9.4	3.3
Native Hawaiian or Pacific Islander	0.2	0.1	0.1
White	**24.7**	**30**	**5.3**
Two or More Races	3	3	0
Economically Disadvantaged	**58.7**	**52.4**	**6.3**
Title I, Part A Participants	61.2	61	0.2
Migrant	0.3	0.1	0.2
Current Limited English Proficient	23.2	20.9	2.3
Bilingual	0.7	0.8	0.1
ESL Participants	17.6	14.4	3.2
Special Education	13.1	13.1	0
Gifted/Talented Participants	11.5	12.6	1.1
At-Risk	55.5	55.4	0.1

1.6 Percentage of Students Taking Different Test Forms

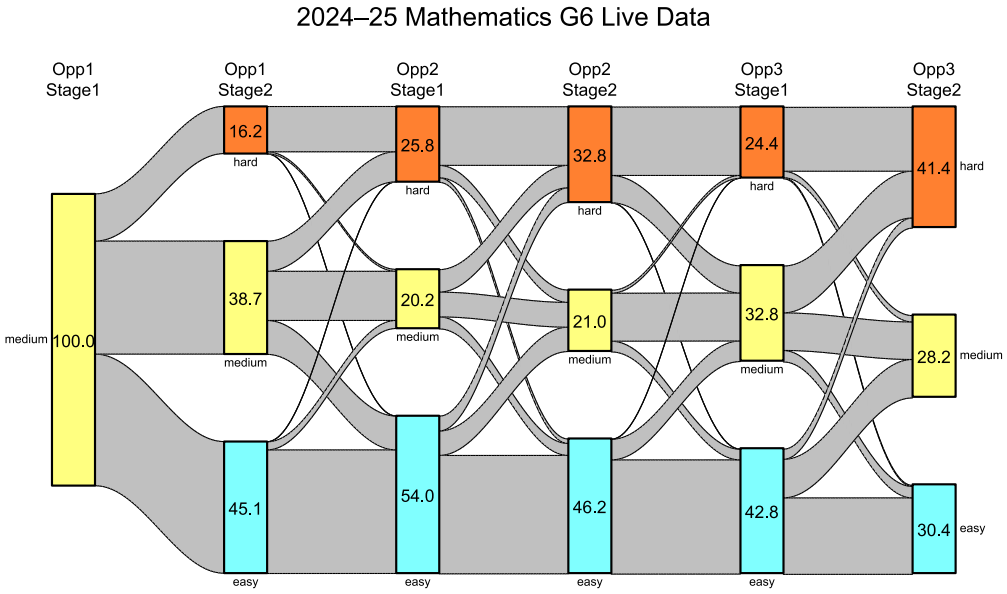
Table 12 lists the percentages of students who were routed to each of the Stage 1 and Stage 2 panels during the 2024–25 test administration. To illustrate the number of students routed to different panels during Stage 1 and 2, Figure 2 visually represents the percentage of students routed to various paths for the grade 6 mathematics test (English) serving as an example. Appendix E includes the visual representations for all tests and opportunities. Based on the numbers and percentages in Table 12, it is evident that a certain percentage of students switched difficulty levels between stages (e.g. medium-medium, medium-high, medium-low). For example, in the grade 6 mathematics test Opp. 3, of the 3,953 students who started in the medium module, 1,523remained there in stage 2, with the majority moving to the high module.

Table 12: Percentages of Students Taking Different Test Forms

Assessment	Route	Opp. 1		Opp. 2		Opp. 3	
		N	%	N	%	N	%
Mathematics Grade 3 (Spanish)	Low-Low	NA	NA	142	70.3	137	61.2
	Low-Medium	NA	NA	14	6.9	28	12.5
	Low-High	NA	NA	1	0.5	13	5.8
	Medium-Low	164	83.2	22	10.9	16	7.1
	Medium-Medium	31	15.7	21	10.4	20	8.9
	Medium-High	2	1	1	0.5	7	3.1
	High-Low	NA	NA	0	0	0	0
	High-Medium	NA	NA	1	0.5	0	0
	High-High	NA	NA	0	0	3	1.3
Mathematics Grade 3 (English)	Low-Low	NA	NA	1,889	28.9	1,791	26.9
	Low-Medium	NA	NA	559	8.6	526	7.9
	Low-High	NA	NA	76	1.2	362	5.4
	Medium-Low	3,216	47.5	602	9.2	402	6
	Medium-Medium	2,560	37.8	1,436	22	796	11.9
	Medium-High	999	14.7	405	6.2	1,176	17.6
	High-Low	NA	NA	20	0.3	15	0.2
	High-Medium	NA	NA	340	5.2	115	1.7
	High-High	NA	NA	1,206	18.5	1,484	22.3
Mathematics Grade 6 (English)	Low-Low	NA	NA	4,570	39.6	3,099	26.5
	Low-Medium	NA	NA	934	8.1	1,553	13.3
	Low-High	NA	NA	571	4.9	407	3.5
	Medium-Low	5,668	46.4	689	6	527	4.5
	Medium-Medium	4,658	38.2	1,050	9.1	1,523	13
	Medium-High	1,880	15.4	930	8.1	1,903	16.3
	High-Low	NA	NA	168	1.5	39	0.3
	High-Medium	NA	NA	465	4	244	2.1
	High-High	NA	NA	2,168	18.8	2,388	20.4

Assessment	Route	Opp. 1		Opp. 2		Opp. 3	
		N	%	N	%	N	%
Mathematics Grade 7 (English)	Low-Low	NA	NA	2,620	35.1	2,590	34.2
	Low-Medium	NA	NA	1,266	17	859	11.4
	Low-High	NA	NA	147	2	77	1
	Medium-Low	4,098	53	768	10.3	538	7.1
	Medium-Medium	2,189	28.3	951	12.7	1,269	16.8
	Medium-High	1,448	18.7	725	9.7	616	8.1
	High-Low	NA	NA	30	0.4	46	0.6
	High-Medium	NA	NA	161	2.2	377	5
	High-High	NA	NA	793	10.6	1,196	15.8
Mathematics Grade 8 (English)	Low-Low	NA	NA	2,208	36.4	1,798	29.5
	Low-Medium	NA	NA	1,018	16.8	627	10.3
	Low-High	NA	NA	263	4.3	38	0.6
	Medium-Low	3,699	60	314	5.2	795	13.1
	Medium-Medium	2,154	34.9	965	15.9	1,216	20
	Medium-High	316	5.1	784	12.9	883	14.5
	High-Low	NA	NA	23	0.4	11	0.2
	High-Medium	NA	NA	101	1.7	37	0.6
	High-High	NA	NA	382	6.3	681	11.2
Social Studies Grade 8 (English)	Low-Low	NA	NA	5,349	33.7	5,651	34.9
	Low-Medium	NA	NA	1,600	10.1	841	5.2
	Low-High	NA	NA	429	2.7	430	2.7
	Medium-Low	8,230	49.3	1,726	10.9	1,830	11.3
	Medium-Medium	5,037	30.2	2,994	18.9	1,453	9
	Medium-High	3,426	20.5	1,192	7.5	3,010	18.6
	High-Low	NA	NA	206	1.3	127	0.8
	High-Medium	NA	NA	767	4.8	319	2
	High-High	NA	NA	1,596	10.1	2,539	15.7

Figure 2: Percentage of Students Routed to Different Paths (Grade 6 Mathematics, English as an example)



2. TTAP Scores from 2024–25

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 Processes[3]. This

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

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 13. In general, students' average scale scores exhibited an increase across opportunities. One exception to this occurred at grade 3 mathematics (Spanish), which is likely due to the small sample size. The Opp. 2 score at social studies was a positive gain but was a much smaller effect compared to other tests and grades. The standard deviation (SD) increased across the three test opportunities. In all tests and grades, the mean growth from Opp. 1 to Opp. 2 was smaller than the increase from Opp. 2 to Opp. 3, except for grade 8 mathematics, which showed higher growth at first, followed by lower growth.

Table 13: Student TTAP Score Growth Across Opportunities

Assessment	Opportunity	N	Mean	SD	Min	25th P	50th P	75th P	Max
Mathematics Grade 3 (Spanish)	Opp. 1	197	533.244	92.266	182	488	514	582	928
	Opp. 2	202	530.807	99.586	335	447	510	592	785
	Opp. 3	224	589.911	107.578	366	503	589	668	916
Mathematics Grade 3 (English)	Opp. 1	6,775	658.194	147.001	104	561	645	737	1,404
	Opp. 2	6,533	688.07	152.307	241	580	676	776	1,404
	Opp. 3	6,667	733.639	170.204	318	608	724	840	1,404
Mathematics Grade 6 (English)	Opp. 1	12,206	949.25	122.526	445	860	950	1022	1,581
	Opp. 2	11,545	967.282	140.549	497	864	969	1060	1,664
	Opp. 3	11,683	993.347	157.884	549	882	986	1092	1,664
Mathematics Grade 7 (English)	Opp. 1	7,735	998.367	126.02	490	915	988	1073	1,650
	Opp. 2	7,461	1012.978	142.347	555	919	1007	1095	1,719
	Opp. 3	7,568	1037.844	155.748	611	921	1025	1126	1,719
Mathematics Grade 8 (English)	Opp. 1	6,169	1065.963	112.769	777	987	1054	1124	1,801
	Opp. 2	6,058	1091.421	138.342	729	1005	1076	1157	1,801
	Opp. 3	6,086	1107.314	144.358	737	1013	1089	1177	1,801
Grade 8 Social Studies (English)	Opp. 1	16,693	904.627	39.716	789	872	904	931	1,103
	Opp. 2	15,859	906.309	48.643	736	870	907	943	1,154
	Opp. 3	16,200	916.986	52.945	739	875	917	954	1,155

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 14. 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.

The mean effect sizes between adjacent opportunities in Table 14 are around 0.18, implying that the average growth in scale scores was relatively small. For grade 3, effect sizes from Opp. 2 to Opp. 3 were relatively larger than other grade levels. Overall, the effect sizes from Opp. 2 to Opp. 3 tended to be larger than those from Opp. 1 to Opp. 2, suggesting that students showed more progress from winter to spring than they did from fall to winter. Mathematics grade 8 showed the opposite pattern with smaller gains from winter to spring. The effect sizes reflecting annual growth, specifically from Opp. 1 to Opp. 3, range from small to medium. The magnitude of growth reduced as grade level increased. In the prior year, the growth slope at grade 7 was an outlier and was nearly flat, but this year, the slope is more consistent with other grades.

Table 14: Effect Size of Student TTAP Scale Score Growth Across Opportunities

Assessment	Opp. 2 vs. 1	Opp. 3 vs. 2	Opp. 3 vs. 1
Grade 3 Mathematics (Spanish)	-0.025	0.57	0.565
Grade 3 Mathematics (English)	0.2	0.282	0.474
Grade 6 Mathematics (English)	0.137	0.174	0.312
Grade 7 Mathematics (English)	0.109	0.167	0.279
Grade 8 Mathematics (English)	0.202	0.112	0.319
Grade 8 Social Studies (English)	0.038	0.21	0.264

In addition to scale scores, students receive a gain, loss, or no change score that reflect their scale score changes across opportunities. Table 15 presents the percentage of students who experienced gains, losses, or no changes in their scaled scores across opportunities. The trends and patterns of growth from Table 14 are similarly reflected in the percentages of students who gained, lost, or experienced no change in their scale scores from Opp. 2 vs. 1, Opp. 3 vs. 2, and Opp. 3 vs. Opp. 1 presented in Table 15.

Table 15: Percentage of Student with Gain, Loss, or No Change TTAP Scale Scores Across Opportunities

Assessment	Opp. 2 vs. 1 Percentage of Gain/Loss/No Change			Opp. 3 vs. 2 Percentage of Gain/Loss/No Change			Opp. 3 vs. 1 Percentage of Gain/Loss/No Change		
	Loss %	Gain %	No Change %	Loss %	Gain %	No Change %	Loss %	Gain %	No Change %
Grade 3 Mathematics (Spanish)	50	49.5	0.5	23.7	76.3	0	26.6	72.9	0.5
Grade 3 Mathematics (English)	36.6	62.6	0.8	31.1	68.7	0.3	22	77.7	0.3
Grade 6 Mathematics (English)	39.8	59.7	0.5	37.8	62.2	0.1	30.5	68.6	1
Grade 7 Mathematics (English)	42	57.1	0.9	34.7	64	1.3	30.1	69.5	0.4
Grade 8 Mathematics (English)	37.9	60.4	1.7	41.4	58.2	0.4	33.8	65.6	0.6
Grade 8 Social Studies (English)	46.9	51.7	1.3	35	63.6	1.4	33.9	65	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 16 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. 3 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. 3 and STAAR show similar trends.

- Currently Does Not Meet Grade Level
- Currently Approaches Grade Level
- Currently Meets Grade Level
- Currently Masters Grade Level

Table 16: Student Performance Level Distribution Across Opportunities

Assessment	Performance Levels	Opp. 1 (N)	Opp. 2 (N)	Opp. 3 (N)	STAAR (N)	Opp. 1 (%)	Opp. 2 (%)	Opp. 3 (%)	STAAR (%)
Grade 3 Mathematics (Spanish)	1	180	174	160	175	91.4	86.1	71.4	71.1
	2	16	26	48	52	8.1	12.9	21.4	21.1
	3	0	2	15	17	0	1	6.7	6.9
	4	1	0	1	2	0.5	0	0.4	0.8
	Total	197	202	224	246	100	100	100	100
Grade 3 Mathematics (English)	1	3,701	2,989	2,381	1,990	54.6	45.8	35.7	27.9
	2	1,707	1,829	1,605	1,632	25.2	28	24.1	22.9
	3	940	1,126	1,619	1,784	13.9	17.2	24.3	25
	4	427	589	1,062	1,725	6.3	9	15.9	24.2
	Total	6,775	6,533	6,667	7,131	100	100	100	100
Grade 6 Mathematics (English)	1	4,765	4,401	3,901	3,337	39	38.1	33.4	26.1
	2	4,873	3,816	3,719	4,406	39.9	33.1	31.8	34.5
	3	2,182	2,585	2,758	3,131	17.9	22.4	23.6	24.5
	4	386	743	1,305	1,904	3.2	6.4	11.2	14.9
	Total	12,206	11,545	11,683	12,778	100	100	100	100
Grade 7 Mathematics (English)	1	4,185	3,712	3,177	3,335	54.1	49.8	42	40.7
	2	2,050	1,868	1,815	1,738	26.5	25	24	21.2
	3	1,266	1,496	1,974	2,009	16.4	20.1	26.1	24.5
	4	234	385	602	1,113	3	5.2	8	13.6
	Total	7,735	7,461	7,568	8,195	100	100	100	100
Grade 8 Mathematics (English)	1	2,876	2,362	2,251	2,255	46.6	39	37	33.1
	2	2,427	2,252	2,066	1,767	39.3	37.2	33.9	25.9
	3	709	1,116	1,222	1,751	11.5	18.4	20.1	25.7
	4	157	328	547	1,041	2.5	5.4	9	15.3
	Total	6,169	6,058	6,086	6,814	100	100	100	100
Grade 8 Social Studies (English)	1	9,523	8,392	7,378	7,220	57	52.9	45.5	41.1
	2	5,010	4,394	4,428	4,475	30	27.7	27.3	25.5
	3	1,729	2,068	2,554	2,592	10.4	13	15.8	14.8
	4	431	1,005	1,840	3,279	2.6	6.3	11.4	18.7
	Total	16,693	15,859	16,200	17,566	100	100	100	100

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 17 provides a comparison of the marginal reliability coefficients for TTAP and STAAR during the 2024–25 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. 1 exhibited lower reliabilities, while Opp. 3 demonstrated higher reliabilities. The general trend across the TTAP tests is that reliability increased as mean scores increased across opportunities. Reliability probably increased because the ability distribution was better aligned to the item pools near the end of the school year, which typically increases score precision and reliability. Opp. 3 showed comparable levels of reliability when compared to STAAR across all tests when all groups were aggregated together.

Upon examining reliabilities at the subgroup level, there was a general pattern of comparability across subgroups, with a few exceptions. Opp. 3 reliabilities are 3 to 8 points higher than STAAR for Asian Americans across all mathematics tests. At grade 8 mathematics, Opp. 3 reliabilities were 2 to 5 points lower than STAAR across ethnic/racial groups. For some subgroups, such as Black or African American, Hispanic or Latino, and female students, the reliabilities tended to be lower than other subgroups. In past years, reliabilities for multiple subgroups fell below 0.7; however, only one subgroup, Black students at grade 8 mathematics, show reliability below that threshold. This clearly demonstrates that increasing the lengths of Opp. 1 and Opp. 2 has resulted in higher reliabilities.

Table 17: Test Reliabilities of TTAP and STAAR

Assessment	Group	N	Opp. 1	Opp. 2	Opp. 3	STAAR
Grade 3 Mathematics (Spanish)	All	246	NA	0.71	0.751	0.742
	Ethnic: H	239	NA	NA	0.752	0.745
Grade 3 Mathematics (English)	All	7,131	0.857	0.877	0.897	0.891
	Ethnic: A	721	0.879	0.879	0.869	0.803
	Ethnic: B	701	0.77	0.809	0.86	0.866
	Ethnic: H	2,672	0.794	0.827	0.87	0.879
	Ethnic: T	371	0.869	0.877	0.901	0.883
	Ethnic: W	2,614	0.838	0.855	0.88	0.875
	Sex: F	3,488	0.839	0.861	0.889	0.888
	Sex: M	3,644	0.867	0.886	0.901	0.891
Grade 6 Mathematics (English)	All	12,778	0.824	0.875	0.897	0.891
	Ethnic: A	881	0.846	0.884	0.902	0.867
	Ethnic: B	1,600	0.782	0.847	0.869	0.856
	Ethnic: H	5,673	0.792	0.854	0.881	0.874
	Ethnic: T	472	0.819	0.878	0.894	0.889
	Ethnic: W	4,033	0.807	0.859	0.881	0.882
	Sex: F	6,398	0.806	0.866	0.89	0.89
	Sex: M	6,383	0.836	0.881	0.902	0.891
Grade 7 Mathematics (English)	All	8,195	0.844	0.883	0.901	0.902
	Ethnic: A	409	0.893	0.906	0.913	0.868
	Ethnic: B	902	0.783	0.853	0.888	0.884
	Ethnic: H	3,672	0.817	0.865	0.888	0.893
	Ethnic: T	303	0.851	0.89	0.906	0.904
	Ethnic: W	2,847	0.836	0.874	0.891	0.898
	Sex: F	4,023	0.826	0.863	0.89	0.897
	Sex: M	4,173	0.857	0.896	0.91	0.906
Grade 8 Mathematics (English)	All	6,814	0.816	0.875	0.893	0.906
	Ethnic: A	405	0.92	0.922	0.92	0.84
	Ethnic: B	743	0.669	0.807	0.828	0.876
	Ethnic: H	2,969	0.737	0.808	0.856	0.891
	Ethnic: T	262	0.759	0.841	0.861	0.901
	Ethnic: W	2,371	0.791	0.843	0.877	0.9
	Sex: F	3,417	0.794	0.859	0.883	0.901
	Sex: M	3,404	0.833	0.888	0.902	0.909
Grade 8 Social Studies (English)	All	17,566	0.789	0.866	0.885	0.901
	Ethnic: A	1,160	0.801	0.859	0.88	0.864
	Ethnic: B	1,654	0.736	0.845	0.863	0.881
	Ethnic: H	8,819	0.739	0.849	0.866	0.885
	Ethnic: T	519	0.804	0.873	0.894	0.904
	Ethnic: W	5,265	0.796	0.853	0.877	0.894
	Sex: F	8,672	0.766	0.853	0.874	0.895
	Sex: M	8,902	0.805	0.877	0.894	0.904

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

3.2 Classification Consistency and Accuracy

Information regarding classification consistency and accuracy has been derived from actual test outcomes from the 2024–25 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 18 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 2025. The classification consistency and accuracy values for TTAP are comparable to those observed in the STAAR assessments. For all TTAP tests, except for the grade 3 mathematics (Spanish) assessment, which has higher classification consistency and accuracy, the classification consistency ranges from 0.68 to 0.72, while the classification accuracy falls between 0.77 and 0.79.

Table 18: Classification Consistency and Accuracy

Assessment	Opps	N	Classification Consistency	Classification Accuracy
Grade 3 Mathematics (Spanish)	Opp. 1	197	0.884	0.917
	Opp. 2	202	0.879	0.915
	Opp. 3	224	0.787	0.841
	STAAR	246	0.757	0.826
Grade 3 Mathematics (English)	Opp. 1	6,775	0.712	0.788
	Opp. 2	6,533	0.691	0.772
	Opp. 3	6,667	0.692	0.775
	STAAR	7,131	0.699	0.783
Grade 6 Mathematics (English)	Opp. 1	12,206	0.684	0.773
	Opp. 2	11,545	0.697	0.781
	Opp. 3	11,683	0.694	0.778
	STAAR	12,778	0.688	0.775
Grade 7 Mathematics (English)	Opp. 1	7,735	0.709	0.783
	Opp. 2	7,461	0.714	0.788
	Opp. 3	7,568	0.715	0.789
	STAAR	8,195	0.724	0.802
Grade 7 Mathematics (English)	Opp. 1	6,169	0.683	0.767
	Opp. 2	6,058	0.686	0.769
	Opp. 3	6,086	0.694	0.777
	STAAR	6,814	0.714	0.796
Grade 8 Social Studies (English)	Opp. 1	16,693	0.703	0.779
	Opp. 2	15,859	0.713	0.786
	Opp. 3	16,200	0.7	0.776
	STAAR	17,566	0.716	0.793

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 19 shows the Pearson correlations between TTAP and STAAR scores by opportunity, subject, and grade.

Table 19 also showcases patterns of associations across different opportunities (Opp. 1, Opp. 2, Opp. 3) and STAAR. The correlation values from grade 3 mathematics (Spanish) are in the range between 0.52–0.72, indicating a relationship that varies from moderate to strong.

Across the various other values in the table, the correlations between Opp. 1, Opp. 2, and Opp. 3 are moderately strong, generally ranging between 0.76–0.86. This suggests a consistent positive relationship in scores across these opportunities. The correlations between Opp. 1, Opp. 2, Opp. 3, and STAAR are also moderately strong, ranging from 0.76–0.87. It is notable that the correlation values between Opp. 3 and STAAR tend to be higher than those between Opp. 1 or Opp. 2 with STAAR, showing that Opp. 3 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 grade 3 mathematics (Spanish), where the sample size is relatively small.

Table 19: Pearson Correlation Coefficients Between the TTAP and Summative Assessment Scale Scores

Assessment	Opp.	Opp. 1	Opp. 2	Opp. 3	STAAR
Grade 3 Mathematics (Spanish)	Opp. 1	1	0.62	0.522	0.566
	Opp. 2		1	0.622	0.627
	Opp. 3			1	0.718
	STAAR				1
Grade 3 Mathematics (English)	Opp. 1	1	0.828	0.812	0.808
	Opp. 2		1	0.843	0.843
	Opp. 3			1	0.873
	STAAR				1
Grade 6 Mathematics (English)	Opp. 1	1	0.823	0.822	0.792
	Opp. 2		1	0.862	0.829
	Opp. 3			1	0.854
	STAAR				1
Grade 7 Mathematics (English)	Opp. 1	1	0.827	0.829	0.809
	Opp. 2		1	0.863	0.842
	Opp. 3			1	0.86
	STAAR				1
Grade8 Mathematics (English)	Opp. 1	1	0.798	0.8	0.778
	Opp. 2		1	0.852	0.832
	Opp. 3			1	0.87
	STAAR				1
Grade 8 Social Studies (English)	Opp. 1	1	0.755	0.762	0.763
	Opp. 2		1	0.83	0.814
	Opp. 3			1	0.839
	STAAR				1

4.2 Prediction Agreement

Beginning with the 2024–25 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 the following:

- 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

Receiver operating characteristic (ROC) curves were employed to predict students' STAAR performance level based on their TTAP scale score. These curves were used 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 20–Table 22 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 20, Table 21, and Table 22 present the predicted accuracy by opportunities for grades 6 and 7 in mathematics, and grade 8 social studies. New test titles for 2024–25 (grade 3 and grade 8 mathematics) could not be included for lack of data from 2023–24. 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 20–Table 22 that are highlighted in green show cells with convincing evidence (>=0.8); yellow denotes 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 "Approaches" performance level often marginally lag behind predictions for the other two performance levels.

Table 20: Prediction Accuracy Summary (Opp. 1)

Assessment	Cut	Performance Level	Accuracy	Specificity	Sensitivity	AUC
Mathematics 6	922	Approaches	0.74	0.82	0.72	0.85
	974	Meets	0.80	0.85	0.72	0.87
	1002	Masters	0.80	0.79	0.84	0.90
Mathematics 7	991	Approaches	0.76	0.78	0.75	0.85
	1016	Meets	0.79	0.78	0.81	0.89
	1065	Masters	0.83	0.82	0.92	0.94
Social Studies 8	896	Approaches	0.81	0.82	0.8	0.89
	915	Meets	0.83	0.83	0.83	0.91
	927	Masters	0.81	0.80	0.89	0.92

Table 21: Prediction Accuracy Summary (Opp. 2)

Assessment	Cut	Performance Level	Accuracy	Specificity	Sensitivity	AUC
Mathematics 6	918	Approaches	0.81	0.82	0.80	0.89
	996	Meets	0.82	0.82	0.81	0.90
	1067	Masters	0.85	0.85	0.86	0.93
Mathematics 7	1018	Approaches	0.79	0.83	0.75	0.87
	1048	Meets	0.83	0.84	0.82	0.90
	1099	Masters	0.85	0.84	0.90	0.94
Social Studies 8	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 22: Prediction Accuracy Summary (Opp. 3)

Assessment	Cut	Performance Level	Accuracy	Specificity	Sensitivity	AUC
Mathematics 6	922	Approaches	0.84	0.84	0.84	0.92
	992	Meets	0.85	0.83	0.88	0.94
	1104	Masters	0.87	0.86	0.91	0.95
Mathematics 7	1005	Approaches	0.83	0.85	0.82	0.91
	1041	Meets	0.84	0.82	0.88	0.93
	1131	Masters	0.89	0.89	0.91	0.96
Social Studies 8	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

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

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 item DIF. 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)

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 23. Table 23 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 23: 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, families, and students 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 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. 1: No gain or loss score is reported since students do not have scores from previous opportunities.
    - Opp. 2
        - Difference in scale scores between Opp. 2 and Opp. 1.
    - Opp. 3
        - Difference in scale scores between Opp. 3 and Opp. 1.
        - Difference in scale scores between Opp. 3 and Opp. 2.
- **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

Figure 3: Individual Student Report (Overall Scores)

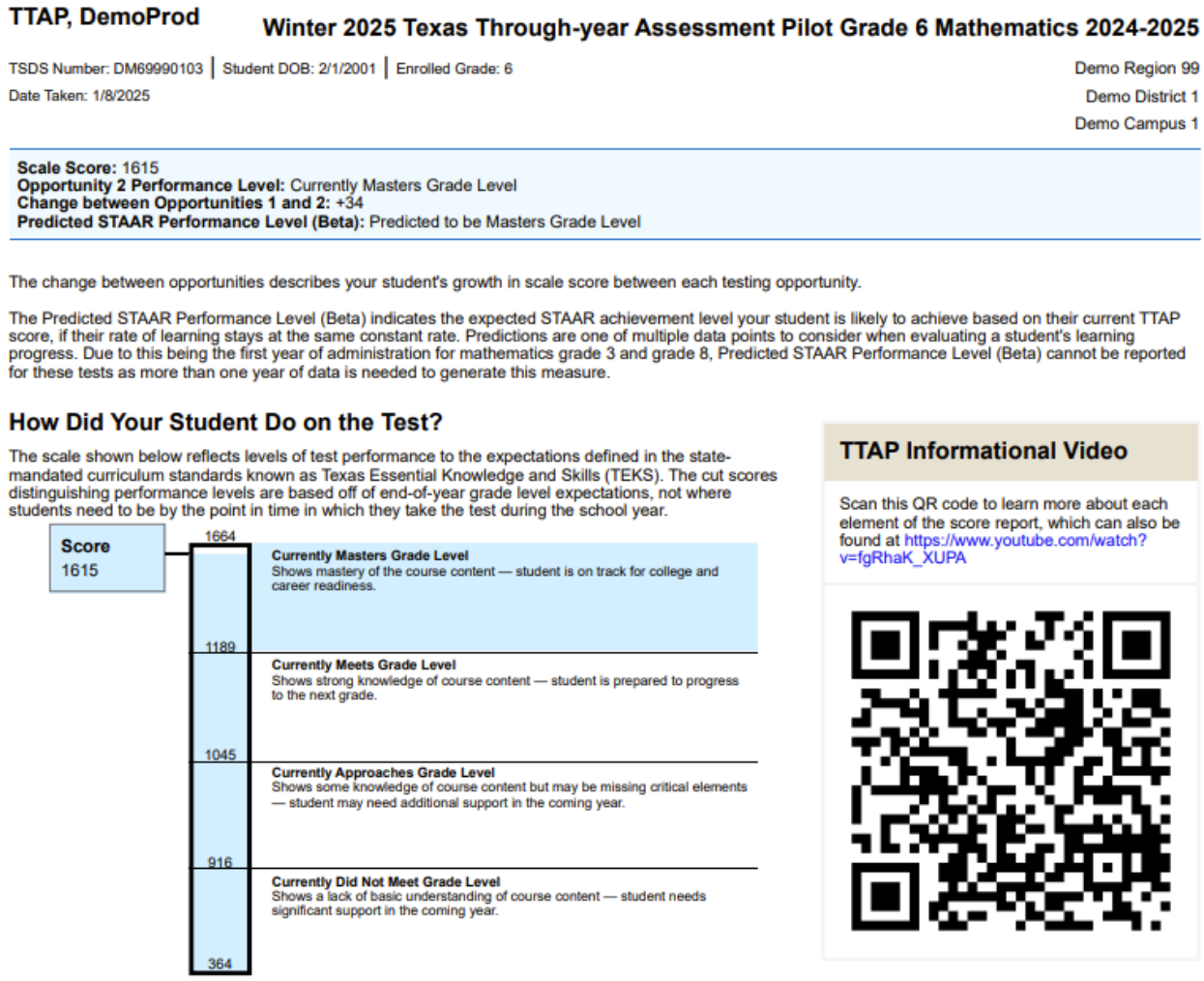


Figure 4 illustrates that in addition to the previously mentioned scores, students also receive detailed information within each reporting category. This includes the following:

- **Total Points Earned and Item Difficulty Level (Did Not Meet/Approaches/Meets/Masters) 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 (Did Not Meet/Approaches/Meets/Masters). This granular level of detail allows students, educators, and parents to identify specific areas of strength and weakness within each reporting category.
- **Total Points Possible by Item Difficulty Level (Did Not Meet/Approaches/Meets/Masters).** Students are informed of the total number of points that could have been earned on items categorized by difficulty level.
- **Percentage Points Correct by Item Difficulty Level (Did Not Meet/Approaches/Meets/Masters).** This metric indicates the percentage of points the student actually earned out of the total possible points, categorized by item difficulty level (Did Not Meet/Approaches/Meets/Masters). It offers a nuanced understanding of

students' performance relative to the difficulty of the items attempted. Item difficulty levels are based on mapping each item into performance levels based on its difficulty. The process of mapping items into performance levels was done after items were developed using empirical item difficulties. Using empirical difficulties ensures that students who have mastered a performance level have a 50%–or better–probability of answering the items in lower levels correctly.

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

How Did Your Student Perform on Items at Each Difficulty Level?

The table below shows the points earned, points possible, and percent correct of items the students answered correctly by item difficulty level. The through-year assessments consist of test items that are aligned to standards at different degrees of rigor. The items are categorized as Did Not Meet, Approaches, Meets, or Masters to help teachers and parents better understand the types of items that students are currently answering correctly. Items categorized as Did Not Meet align to easiest knowledge and skills within the standards presented in familiar contexts. Items categorized as Approaches align to easier knowledge and skills within the standards presented in familiar contexts. Items categorized as Meets align to knowledge and skills within the standards presented at the level of rigor expected by the standard. Items categorized as Masters align to the more difficult knowledge and skills within the standards presented in varied and unfamiliar contexts. Students will generally respond correctly to items whose difficulty level is below or matches their overall ability. To grow, students need to be presented tasks that are at the same or next higher difficulty level during instruction and assessment opportunities. Note that students may not see items of all difficulty levels as TTAP is an adaptive test, matching students to items based on their demonstrated ability. For more information about item difficulty, download the supplemental materials along with the ISR in CRS. N/A reported in the percent correct column signifies that students did not receive with any items at this difficulty level.

Item Difficulty Level	Total Points Earned	Total Points Possible	Percent Points Correct
Masters	23	24	96%
Meets	7	7	100%
Approaches	2	2	100%
Did Not Meet	0	0	N/A

How Did Your Student 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 Difficulty Level*	Standard Key	Student Expectation	Points
Masters	6.1.7.B	Distinguish between expressions and equations verbally, numerically, and algebraically.	1/1
Masters	6.1.7.D	Generate equivalent expressions using the properties of operations: inverse, identity, commutative, associative, and distributive properties.	1/1
Meets	6.1.4.G	Generate equivalent forms of fractions, decimals, and percents using real-world problems, including problems that involve money.	1/1
Meets	6.1.4.G	Generate equivalent forms of fractions, decimals, and percents using real-world problems, including problems that involve money.	1/1
Masters	6.1.7.D	Generate equivalent expressions using the properties of operations: inverse, identity, commutative, associative, and distributive properties.	1/1
Masters	6.1.7.D	Generate equivalent expressions using the properties of operations: inverse, identity, commutative, associative, and distributive properties.	1/1
Masters	6.1.2.D	Order a set of rational numbers arising from mathematical and real-world contexts.	1/1

2. Computations and Algebraic Relationships

Item Difficulty Level*	Standard Key	Student Expectation	Points
Masters	6.2.10.A	Model and solve one-variable, one-step equations and inequalities that represent problems, including geometric concepts.	1/1
Meets	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.	2/2
Masters	6.2.3.D	Add, subtract, multiply, and divide integers fluently.	1/1
Masters	6.2.3.E	Multiply and divide positive rational numbers fluently.	1/1
Approaches	6.2.9.C	Write corresponding real-world problems given one-variable, one-step equations or inequalities.	1/1
Masters	6.2.3.B	Determine, with and without computation, whether a quantity is increased or decreased when multiplied by a fraction, including values greater than or less than one.	1/1
Meets	6.2.3.B	Determine, with and without computation, whether a quantity is increased or decreased when multiplied by a fraction, including values greater than or less than one.	1/1
Masters	6.2.9.A	Write one-variable, one-step equations and inequalities to represent constraints or conditions within problems.	1/1
Masters	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.	1/1
Masters	6.2.3.E	Multiply and divide positive rational numbers fluently.	1/1
Masters	6.2.9.B	Represent solutions for one-variable, one-step equations and inequalities on number lines.	2/2

3. Geometry and Measurement

Item Difficulty Level*	Standard Key	Student Expectation	Points
Masters	6.3.8.B	Model area formulas for parallelograms, trapezoids, and triangles by decomposing and rearranging parts of these shapes.	1/1
Masters	6.3.8.D	Determine solutions for problems involving the area of rectangles, parallelograms, trapezoids, and triangles and volume of right rectangular prisms where dimensions are positive rational numbers.	1/1
Masters	6.3.4.H	Convert units within a measurement system, including the use of proportions and unit rates.	1/1
Masters	6.3.8.C	Write equations that represent problems related to the area of rectangles, parallelograms, trapezoids, and triangles and volume of right rectangular prisms where dimensions are positive rational numbers.	0/1

*See description of item difficulty levels in the section titled "How did your student perform on items at each difficulty level?" on page 3.

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.

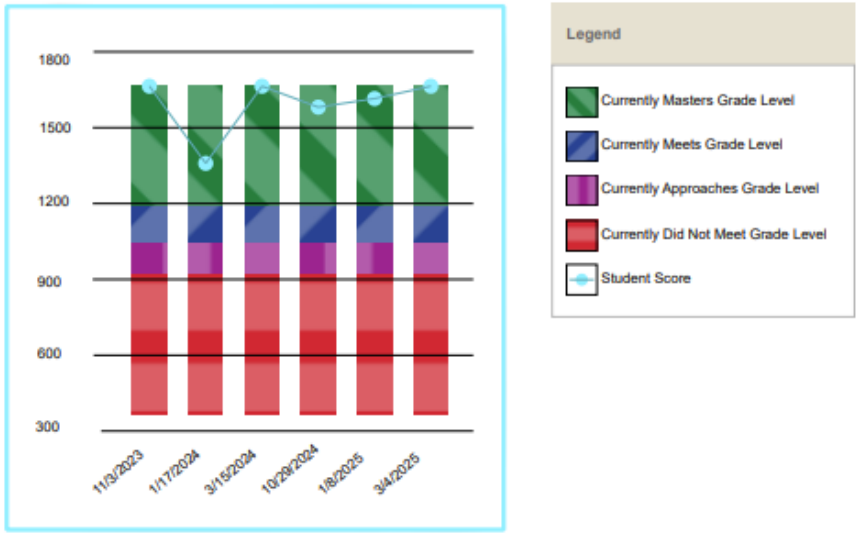
This allows for progress monitoring over time, facilitating identification of trends and areas for improvement.

Figure 5: Individual Student Report (Progress Monitoring)

Your Student's Progress

Longitudinal Trend Chart Information

The chart below reports your student's performance over time. The shaded areas in multiple colors indicate the scale score range in each performance level. Each mark on the graph represents your student's score and indicates whether he or she met the standards that year. The cut-off scores for each performance level remain the same for each test title within the school year. However, the performance level cut-off scores vary between content areas and grade levels.



Your Student's Progress

Date	Test Administration	Assessment Name	Scale Score	Performance Level
11/3/2023	Opportunity 1	Fall 2023 Texas Through-year Assessment Pilot Grade 6 Mathematics	1664	Currently Masters Grade Level
1/17/2024	Opportunity 2	Winter 2024 Texas Through-year Assessment Pilot Grade 6 Mathematics	1359	Currently Masters Grade Level
3/15/2024	Opportunity 3	Spring 2024 Texas Through-year Assessment Pilot Grade 6 Mathematics	1664	Currently Masters Grade Level
10/29/2024	Opportunity 1	Fall 2024 Texas Through-year Assessment Pilot Grade 6 Mathematics	1581	Currently Masters Grade Level
1/8/2025	Opportunity 2	Winter 2025 Texas Through-year Assessment Pilot Grade 6 Mathematics	1615	Currently Masters Grade Level
3/4/2025	Opportunity 3	Spring 2025 Texas Through-year Assessment Pilot Grade 6 Mathematics	1664	Currently Masters Grade Level

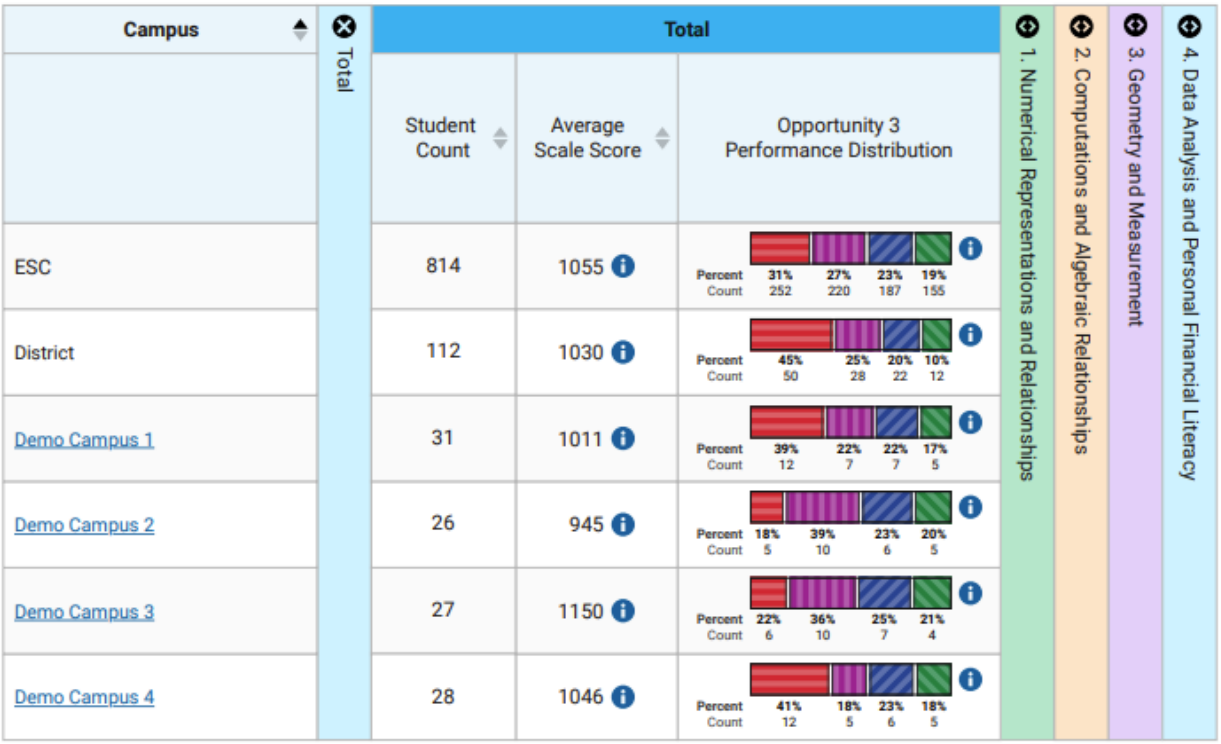
Campus-/District-Level Reports

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 Scale Score).** Figure 6 shows a score representing 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).** In Figure 6, the Opportunity Performance Distribution 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.

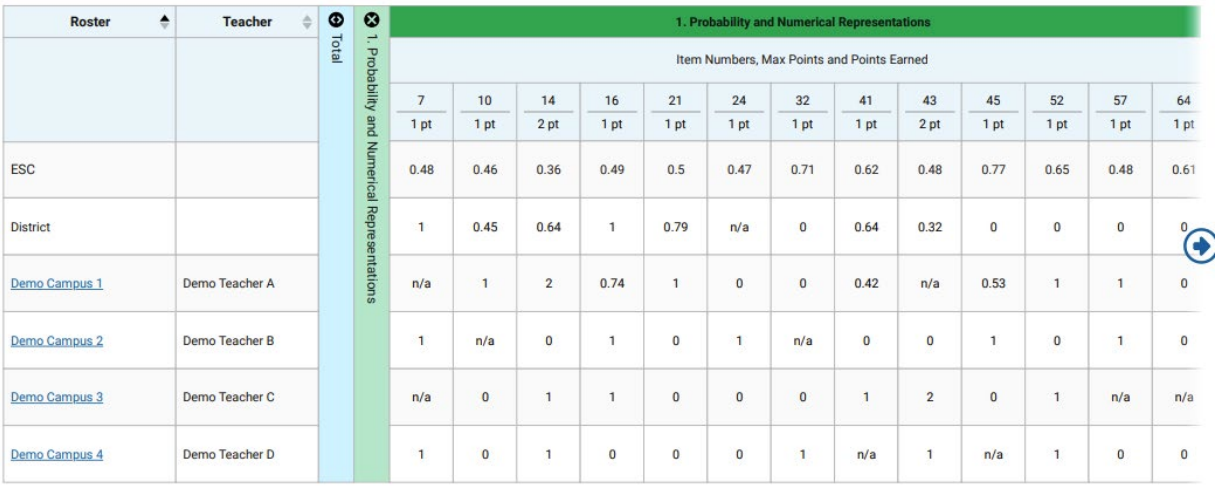
Figure 6: District/Campus Report (Scale Score and Performance Level)



Campus	Total	Total			1. Numerical Representations and Relationships	2. Computations and Algebraic Relationships	3. Geometry and Measurement	4. Data Analysis and Personal Financial Literacy
		Student Count	Average Scale Score	Opportunity 3 Performance Distribution				
ESC		814	1055	Percent 31% 27% 23% 19% Count 252 220 187 155				
District		112	1030	Percent 45% 25% 20% 10% Count 50 28 22 12				
Demo Campus 1		31	1011	Percent 39% 22% 22% 17% Count 12 7 7 5				
Demo Campus 2		26	945	Percent 18% 39% 23% 20% Count 5 10 6 5				
Demo Campus 3		27	1150	Percent 22% 36% 25% 21% Count 6 10 7 4				
Demo Campus 4		28	1046	Percent 41% 18% 23% 18% Count 12 5 6 5				

- **Mean Raw Score by Item.** Figure 7 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 7: District/Campus Report (Percentage Correct)



Roster	Teacher	Total	1. Probability and Numerical Representations												
			Item Numbers, Max Points and Points Earned												
			7 (1 pt)	10 (1 pt)	14 (2 pt)	16 (1 pt)	21 (1 pt)	24 (1 pt)	32 (1 pt)	41 (1 pt)	43 (2 pt)	45 (1 pt)	52 (1 pt)	57 (1 pt)	64 (1 pt)
ESC			0.48	0.46	0.36	0.49	0.5	0.47	0.71	0.62	0.48	0.77	0.65	0.48	0.61
District			1	0.45	0.64	1	0.79	n/a	0	0.64	0.32	0	0	0	0
Demo Campus 1	Demo Teacher A		n/a	1	2	0.74	1	0	0	0.42	n/a	0.53	1	1	0
Demo Campus 2	Demo Teacher B		1	n/a	0	1	0	1	n/a	0	0	1	0	1	0
Demo Campus 3	Demo Teacher C		n/a	0	1	1	0	0	0	1	2	0	1	n/a	n/a
Demo Campus 4	Demo Teacher D		1	0	1	0	0	0	1	n/a	1	n/a	1	0	0

7. Continuous Research and Improvement Plans

In this section, we summarize TTAP special studies conducted in year 2024–25. These studies evaluate the reliablity and validity of TTAP assessments and its compariability with STAAR. We review the objectives and key findings of each study 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.

Eight special studies were conducted in school year 2024–25. These studies included the following:

- *Study 1.1 Reliability and Comparability Across Opportunities*
- *Study 2.1 Research Outcomes on Linear Composite + Opp. 3 Scoring Combination on Classification Consistency with STAAR Using Simulations*
- *Study 2.2 Who Benefits from the LC + Opp. 3 Within and Across years?*
- *Study 3.1. ROC Curves*
- *Study 3.2.1 Routing Study (Alternate study responsive to TTAC feedback)*
- *Study 3.2.3 Impact of Routing Decision on Student Outcome with Similar-Performing Students as Regression Discontinuity (To address TTAC feedback)*
- *Study 3.3 Detecting Potential Fall Collusion Strategies to Inflate Student Growth of Through-Year Measures*
- *Study 4.1 Item Difficulty Modeling*

Study 1.1 Reliability and Comparability Across Opportunities

We conducted simulations to compare the technical characteristics of TTAP Opportunities and STAAR. Our comparisons focused on ability recovery, classification accuracy, performance-level distribution, and reliability of both assessments. Our analysis revealed that, in the non-growth condition, the three TTAP opportunities showed similar technical properties with STAAR in the aforementioned aspects. In the growth condition, only TTAP Opportunity 3 showed similar technical properties with STAAR, which is expected because summative assessments are optimized to function based on student performance at the end of the year.

STAAR and TTAP opportunities had similar correlations and were comparable in the non-growth condition. While the correlations were slightly lower for the earlier TTAP opportunities in the growth condition, the correlations were higher than those from the Pilot Year 2 Design (Lim & Chen, 2024) and much more similar to Opportunity 3. Additionally, the average biases from TTAP Opportunities 1–3 and STAAR were close to 0, indicating that the TTAP opportunities and STAAR did not consistently overestimate or underestimate students' abilities.

In general, because TTAP Opportunity 3 forms were shorter than STAAR, the shorter TTAP Opportunity 3 forms were more efficient in achieving similar levels of reliability and classification accuracy. This finding aligns with results from earlier studies (Armstrong et al., 2004; Chen et al., 2023; Patsula, 1999). More specifically, all TTAP opportunities had reliabilities exceeding 0.80 in the five titles both for the growth and non-growth conditions, and they exceeded .85 for Opportunity 3.

In the non-growth condition, the performance-level classifications of TTAP Opportunities 1–3 and STAAR resulted in similar percentages of students being distributed into the four performance levels, with the difference with STAAR being 3% at most. In the growth condition, the difference with STAAR was larger in TTAP Opportunities 1 and 2, mainly because there was an assumption, based on observed trend data, that the students' abilities, on average, increased with each opportunity. This study shows positive evidence that TTAP Opportunity 3 and STAAR would likely provide comparable interpretations of student ability if administered within the STAAR testing window and with the assumption that the students, their motivation, and administration conditions were also comparable to STAAR.

Study 2.1 Research Outcomes on Linear Composite + Opp. 3 Scoring Combination on Classification Consistency with STAAR Using Simulations

This simulation study is an extension of two previous studies on this topic (Gianopulos & Schneider, 2023 & Gianopulos, 2024) using simulated data informed by Pilot Year 2 and designed around TTAP new blueprints and titles with method modifications based on feedback from the Texas Technical Advisory Committee (TTAC). Eight methods for producing linear composite scores were compared. To control for ability differences between STAAR and the third TTAP Opportunity (Opp. 3), tests were simulated as if they were administered during the STAAR testing window. First, all methods were compared in their ability to recover true theta under a Gain condition and a No Gain condition. Then, the effects of missingness on true theta recovery were examined.

Results changed depending on the condition. Of the studied methods, the multiple regression (MR) linear composite (LC) scores and the MR with the help-not-hurt (HNH) rule tended to show the best performance in terms of theta recovery and classification accuracy under the Gain condition. In contrast, under the No Gain condition, the opportunity weighted, and reliability weighted methods showed the best performance. Similar to prior findings, linear composites tended to underestimate STAAR under the Gain condition, while the MAX tended to overestimate STAAR. The MAX, Opp. 3, and HNH methods in general showed less variation in performance compared to prior studies. The added value of any LC method beyond Opportunity 3 in this study is less dramatic than previously shown. When Opp. 1 or Opp. 2 were missing, LC scores generally maintained similar performance to non-missingness, but performance declined whenever Opp. 3 was missing. From a policy perspective, the results from the missingness analysis suggest that an Opp. 3 score should always be a requirement.

Study 2.2 Who Benefits From the LC + Opp. 3 Within and Across Years?

Linear composite scores express a final summative score as a linear combination of longitudinal scores. Ideally, linear composite scores would work equally well across all subgroups. Three linear composite scores–multiple regression, opportunity weighted, and reliability weighted scores– were compared in terms of how consistently they predicted STAAR across 28 subgroups of interest. A series of increasingly complex linear mixed models were compared. If the model that allowed group-specific slopes and intercepts emerged as the best fitting model, this was taken as evidence that a single linear composite would not work equally well across subgroups.

Findings revealed that certain groups do need different intercepts and slopes to avoid differential prediction error. These groups include Asian-American students, at risk students, male students with disabilities, and most especially, female students with disabilities. Asian students were unique in that the direction of the prediction bias was negative (underestimated predictions), while the other groups had positive bias (over-predictions). Given the observed differences in growth patterns of ethnic/racial groups, it is inferred that the root cause for this predictive invariance is differential growth patterns by group. Given that academic growth is construct-relevant and the observed effect sizes are trivial to small, the variability in the performance of linear composites does not warrant concern.

This study also devised a new TTAP-derived progress measure to mirror the current measure used with STAAR. The new TTAP progress measure used within-year scores, rather than across-year scores. Given the different time horizons of each measure, it is not surprising that the two progress measures only moderately agree. The TTAP progress measure increases the number of students in the Expected growth category and decreases the number in the Accelerated category compared to the STAAR progress measure. Whether the TTAP progress measure is more or less accurate is beyond the scope of this study, but this question could be studied further with simulations.

Study 3.1. ROC Curves

The purpose of this study was to examine how well the ROC predictions performed in school year 2024–25. The ROC prediction study (Chen & Gianopulos, 2024) was conducted in the summer of 2024 using data from school year 2023–24. The prediction cut scores for mathematics grades 6 and 7 and social studies grade 8 were used to predict STAAR performance levels. This study compared these predictions to the observed STAAR performance level classifications from the spring of 2025. Only mathematics grades 6 and 7, and social studies grade 8 are given in both 2023–24 and 2024–25, so the comparisons are limited to that. Accuracy ranged from .53 to .88. Specificity ranged from .44 to .99, and Sensitivity ranged from .45 to .92. Grade 6 Opp. 1 showed the greatest discrepancies between expected and observed. The observed accuracy averaged across tests declined 7%, average sensitivity declined 3.5%, and average specificity declined 13%. Sensitivity is the ability of a measure to detect true positive classifications, while specificity is the test's ability to accurately predict true negatives. These results indicate the ROC predictions performed as expected at most grades and most tests, especially for the Meets performance level, but predictions degraded more at the Approaches and Masters levels, Opp. 1, and grade 6 than other conditions.

Study 3.2.1 Routing Study (Alternate Study Responsive to TTAC Feedback)

As a multistage assessment comprising three opportunities, each with two stages, TTAP makes routing decisions to place students into one of three possible difficulty tiers at each stage. As a continuation of the routing study conducted during the 2023–24 pilot year, this study examined psychometric properties related to routing decisions in TTAP, as well as their relationship to performance-level classifications.

A series of research questions were investigated. One primary question focused on the impact of incorrect routing decisions on ability estimates. A simulation was conducted using simulated students generated based on real data. Students who should have been routed to extreme difficulty categories (i.e., Low/High) were intentionally re-routed to the opposite categories (i.e., High/Low). Several performance measures were calculated, including ability estimation performance (RMSE), reliability when a routing decision is made, and performance-level classification performance (e.g., accuracy). Comparisons between simulated and live data were made to evaluate the extent to which routing behavior in the simulated data resembled that of live test administrations, ensuring that conclusions from this study could be applied to real-world data. The study found that when students were intentionally re-routed, the student's theta was largely recovered in Stage 2.

A second research question examined the impact of administering only Opportunity 3, as opposed to all three opportunities. A simulation was conducted in which students were given only Opportunity 3, with all students placed in the Medium difficulty category for Stage 1. The study found that while Stage 1 ability estimation performance in the Opportunity 3-only condition

was slightly lower than in the all-opportunity condition, Stage 2 ability estimation performance was similar between the two conditions.

Study 3.2.3 Impact of Routing Decision on Student Outcome with Similar-Performing Students as Regression Discontinuity (To address TTAC feedback)

In a through-year multistage test, one concern that may arise is about undesirable effects of routing decisions. In particular, for two students who have the same true ability, one may be routed to take a medium module and another may be routed to take an easy module, because of inherent estimation error in the ability estimates used for routing decisions. A regression discontinuity analysis was conducted on simulated and live data from TTAP 2024–25. This was modeled as a multiple-cutoff discontinuity using two cutoffs to represent the two routing cuts for placing students into one of three difficulty tiers (Easy, Medium, Hard). The predictor variable was ability estimates, allowing routing decision points to be modeled as sharp discontinuities. The outcome variable was the end-of-year ability estimate from Opportunity 3 Stage 2. Because there are many possible routing points across the course of TTAP test administration, ability estimates from each of the stages were used as the predictor variable, repeating the discontinuity regression analysis as needed.

Discontinuity effect estimates were evaluated against an effect size criteria. Specifically, if the effects were smaller than what was needed to produce a one-point raw score difference in Opportunity 3, it was considered as not meaningful. In simulated data, among the 30 effects analyzed (5 test titles * 3 stages * 2 cuts), most of the effects did not reach the criteria, with only one reaching a meaningful effect. This suggests that routing decisions made throughout the course of TTAP did not have undeservingly increasing/decreasing effects on end-of-year student outcomes in the simulated conditions. In live data, 16 out of 30 effects reached the criteria. Further comparisons between live and simulated data suggested that the live data result may be affected by extreme values being present in the score distribution.

Study 3.3 Detecting Potential Fall Collusion Strategies to Inflate Student Growth of Through-Year Measures

In this study, we compared different methods to detect unexpectedly low scores or fast response times in the Opp. 1 test event. Since a TTAP gain score is the simple difference between Opp. 3 and Opp. 1, the resulting gain score is vulnerable to *sandbagging*: the act of intentionally lowering a pre-test score to inflate the gain score (Ho, 2012). To detect sandbagging and possible collusion, we defined expectations based on IRT model predictions and various general linear regression models using past STAAR data to predict Opp. 1 scores and response times. Absent teacher-level labels, we aggregated indices at the school level to detect possible collusion. In addition to using the standard methods in the forensic reports, we developed a new index, the school-level weighted regression (SLWR), to evaluate school residual response times as an indicator of collusion to sandbag. Further, we implemented rapid guessing indices at the student level and aggregated these metrics to the school level.

While the percentage of low-scoring students was higher than expected and a few schools were flagged, the absence of benchmark metrics makes it unclear whether the results reflect

collusion, sandbagging, or simply lower motivation. Unexpectedly, schools flagged for lower response times were not the same as those flagged for low scores. This finding may have differed had the unit of analysis been classrooms rather than schools. This study gives benchmarks for underperformance, response time, rapid guessing, and misfit across two tests: a high-motivation test (STAAR) and a low-motivation pilot test (Opp. 1). Some schools had unusually high rates of low or high Opp. 1 scores (see Figures 4 and 5) which could serve as an early warning system. Although not the primary focus, an interesting pattern emerged among grade 8 students: high achievers maintained effort on STAAR but declined on Opp. 1 as item position increased, while low achievers showed declining effort across both tests. Differences in the test engagement patterns across subgroups like these, as measured by rapid guessing, may be useful to improve models that predict end-of-year status using interim scores. Although this study began as a search for evidence of collusion, we found statistically significant patterns—negative or positive —that schools should investigate or celebrate.

Study 4.1 Item Difficulty Modeling

This study investigated item features associated with STAAR reading items which are a subset of the reading language arts construct to determine which features were associated with item classifications categorized into four performance levels. In this item difficulty modeling (IDM) study, item features were included as predictors that represented text-task interactions. Two IDM models were used: a random forest classifier model was used to determine how accurately the final set of features predicted item performance-level classification. We found that 37% of the items were accurately classified to their performance level; however, the correlations of the Range performance level descriptors to performance levels was smaller than found in a previous study. Final features retained in the model in order of importance were the number of difficult words in the stem and distractors, Range performance level descriptors, amount of stimulus text needed to answer the item, and item type. Implications for additional research related to Range performance level descriptors for STAAR reading items are discussed because the models could not differentiate between Levels 2 and 3.

8. 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.com.

Masters, 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). doi: https://doi.org/10.7275/an9m-2035

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].

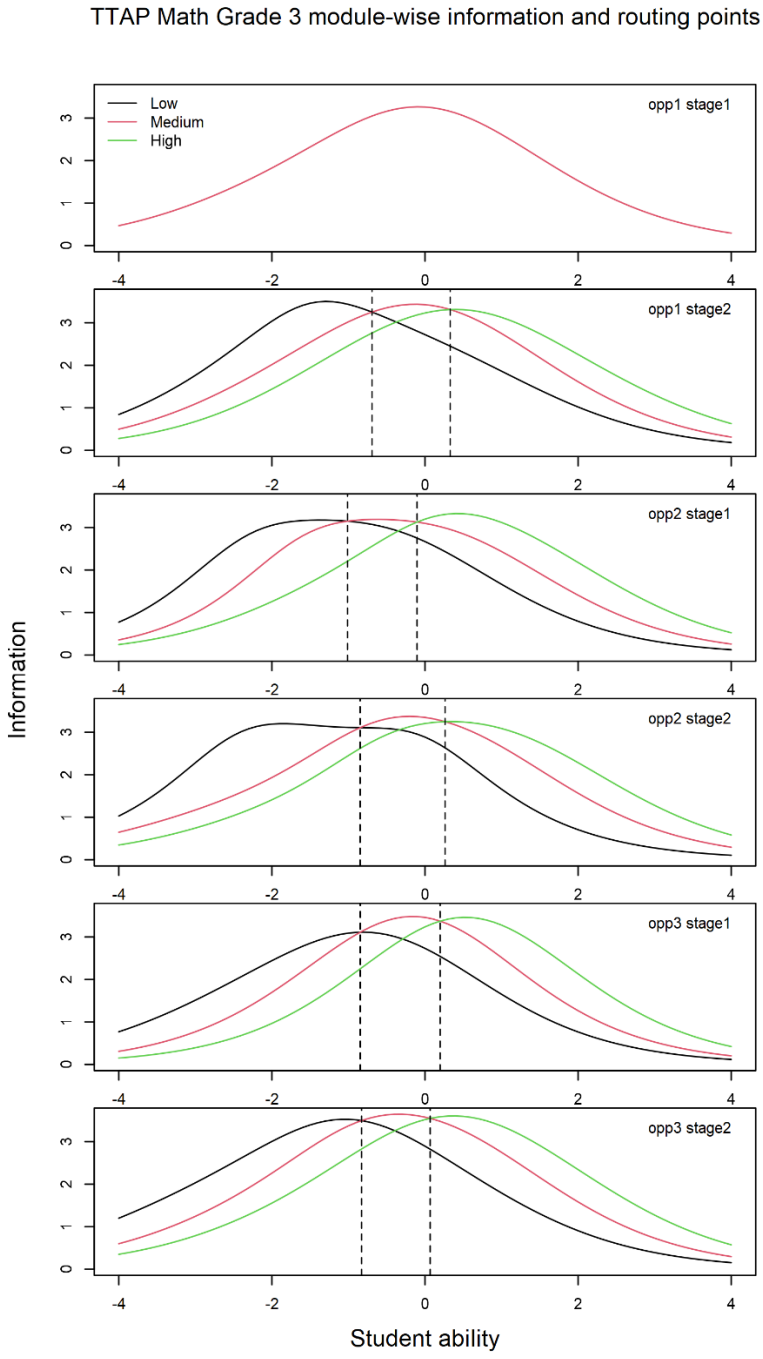
Spearman, C. (1910). Correlation calculated with faulty data. *British Journal of Psychology*, *3*(3), 271–295. https://doi.org/10.1111/j.2044-8295.1910.tb00206.x

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(1950)3:1<32::aid-cncr2820030106>3.0.co;2-3

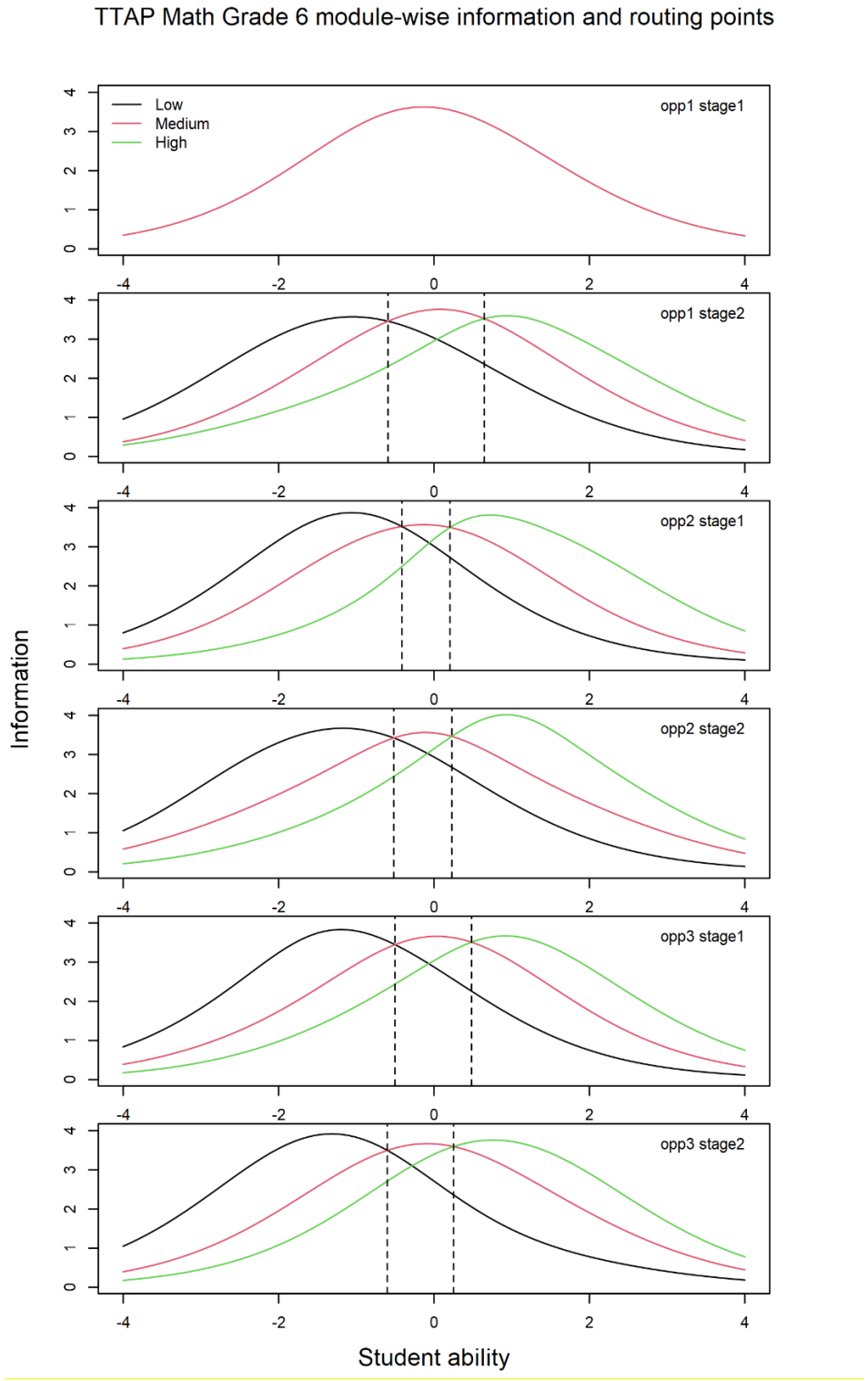
Appendix A: 2024–25 TTAP Administration Test Information Functions

Figure A.1: TTAP 2024–25 Test Information Function (Grade 3 Mathematics)



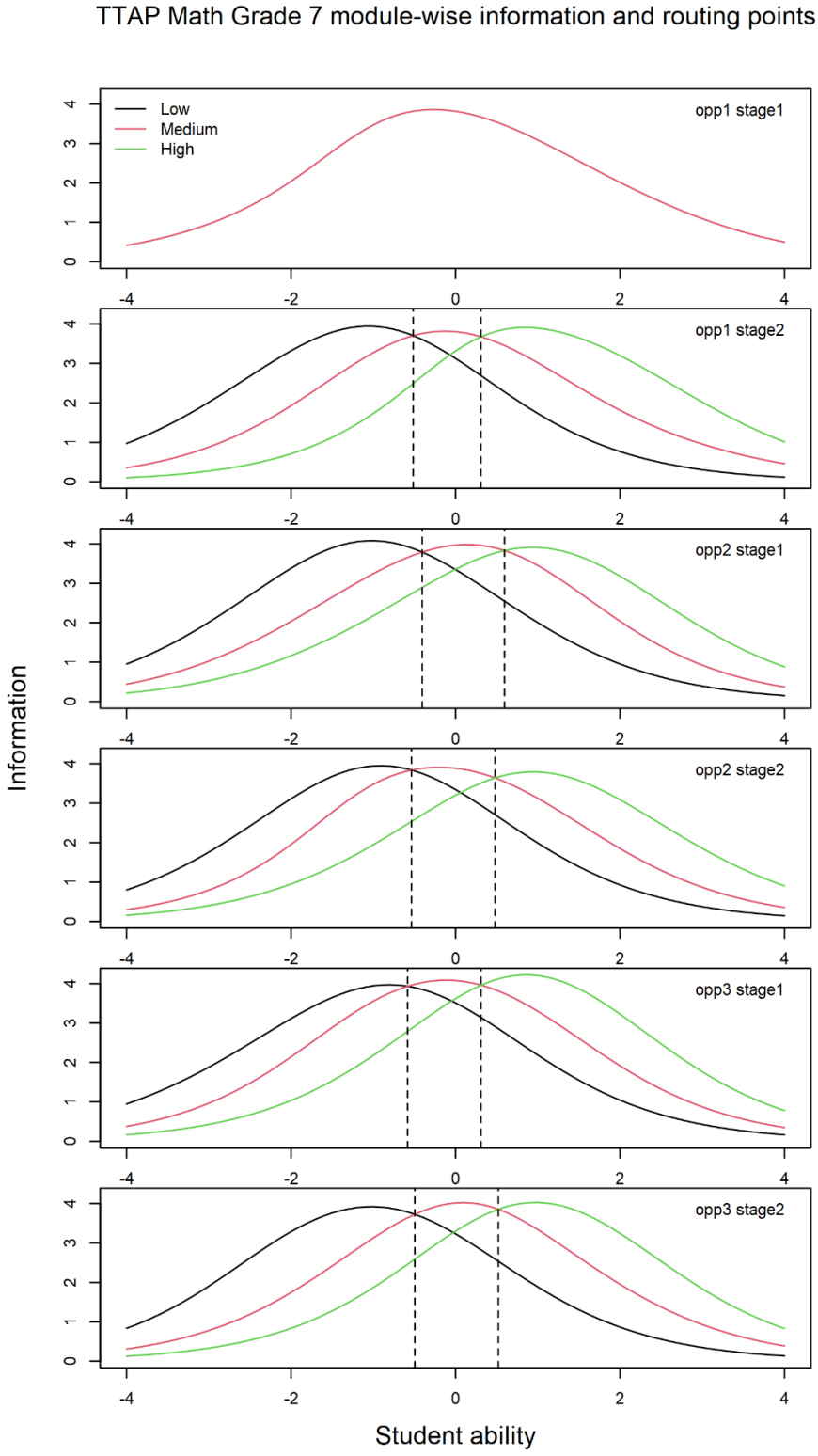
Note: The single curve in the top-most panel is medium. In all other panels, the first, second, and third curves are easy, medium, and difficult modules respectively.

Figure A.2: TTAP 2024–25 Test Information Function (Grade 6 Mathematics)



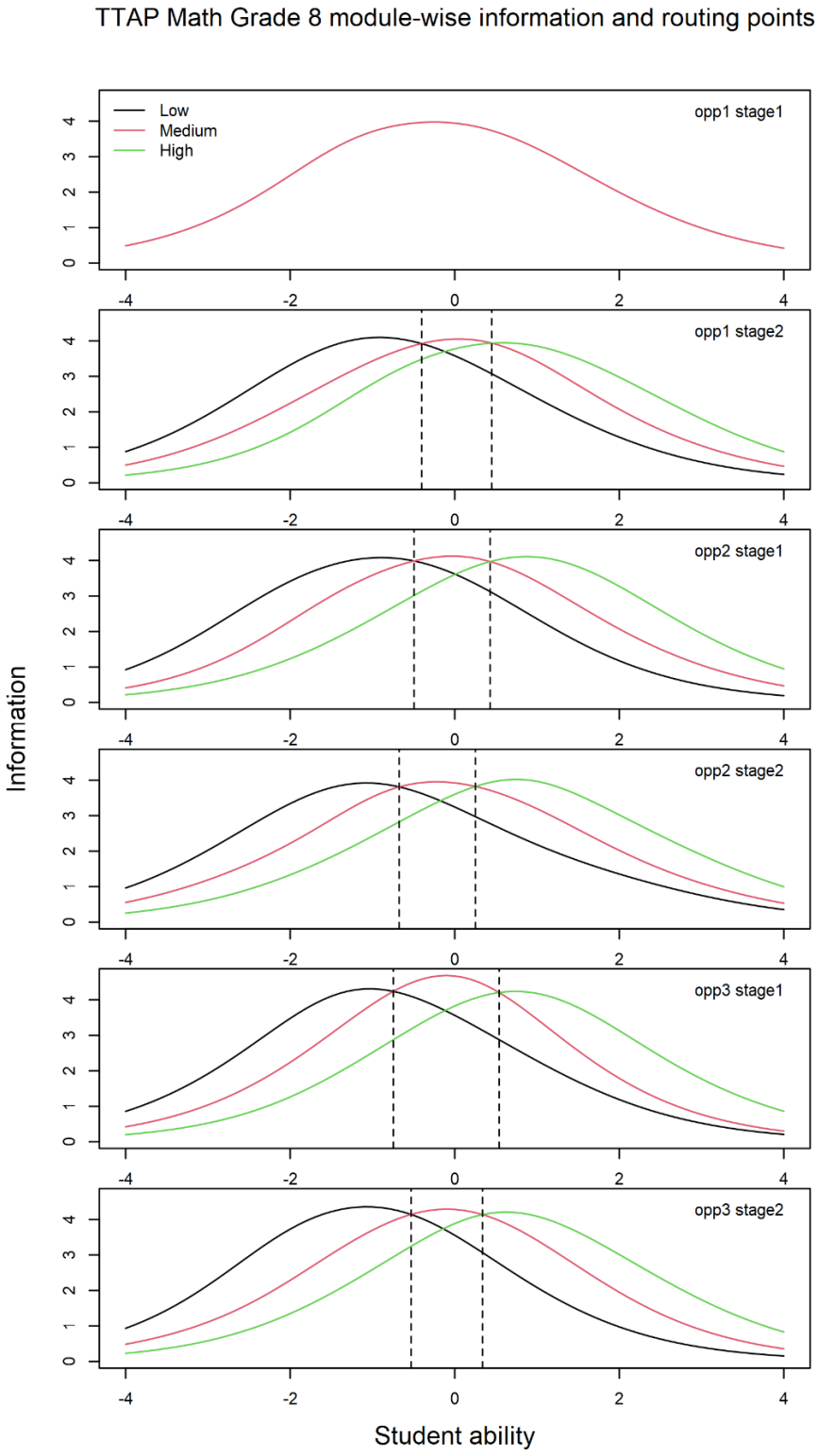
Note: The single curve in the top-most panel is medium. In all other panels, the first, second, and third curves are easy, medium, and difficult modules respectively.

Figure A.3: TTAP 2024–25 Test Information Function (Grade 7 Mathematics)



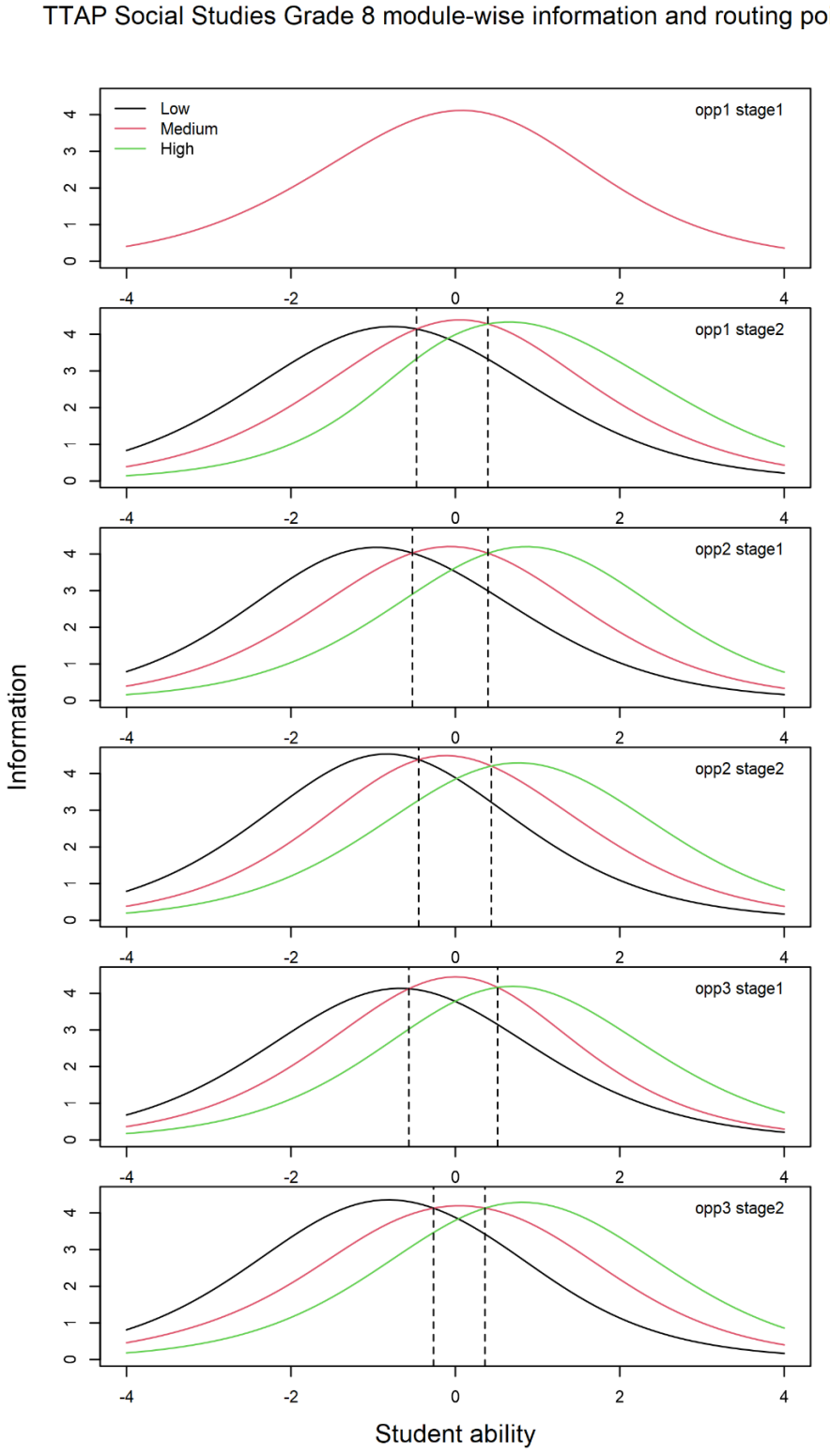
Note: The single curve in the top-most panel is medium. In all other panels, the first, second, and third curves are easy, medium, and difficult modules respectively.

Figure A.4: TTAP 2024–25 Test Information Function (Grade 8 Mathematics)



Note: The single curve in the top-most panel is medium. In all other panels, the first, second, and third curves are easy, medium, and difficult modules respectively.

Figure A.4: TTAP 2024–25 Test Information Function (Grade 8 Social Studies)



Note: The single curve in the top-most panel is medium. In all other panels, the first, second, and third curves are easy, medium, and difficult modules respectively.

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 3 mathematics.
    - For grade 3 mathematics, 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 3 mathematics.
- 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 records.
- Keep only records with respective DISCREPANCYINDICATOR value of 0.
    - Use “DISCREPANCYINDICATORMATHEMATICS” for mathematics.
    - 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 (Texas Student Data System [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/RACE REPORTING CATEGORY	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 economically 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 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-approved full bilingual program	2, 3, 4, 5 = Bilingual 0 = Otherwise

Summative Data Variables	Values/Definitions	Recode for Analysis
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 the Test Information Distribution Engine (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 an actual student.	Keep values of *0*.

Appendix E: Percentage of Students Routed to Different Paths

Figure E.1: Sankey Plot Grade 3 Mathematics

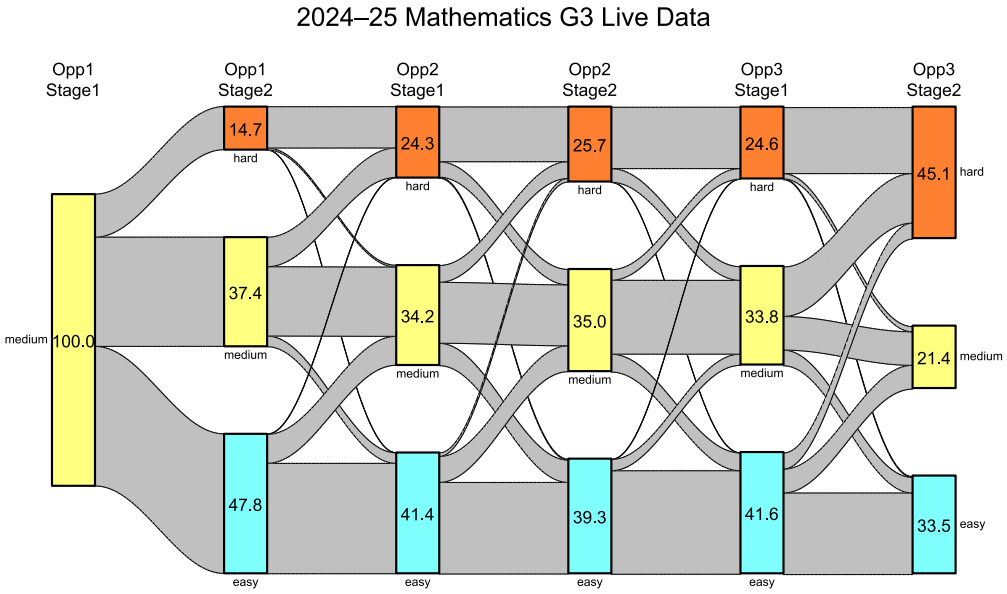


Figure E.2: Sankey Plot Grade 6 Mathematics, English

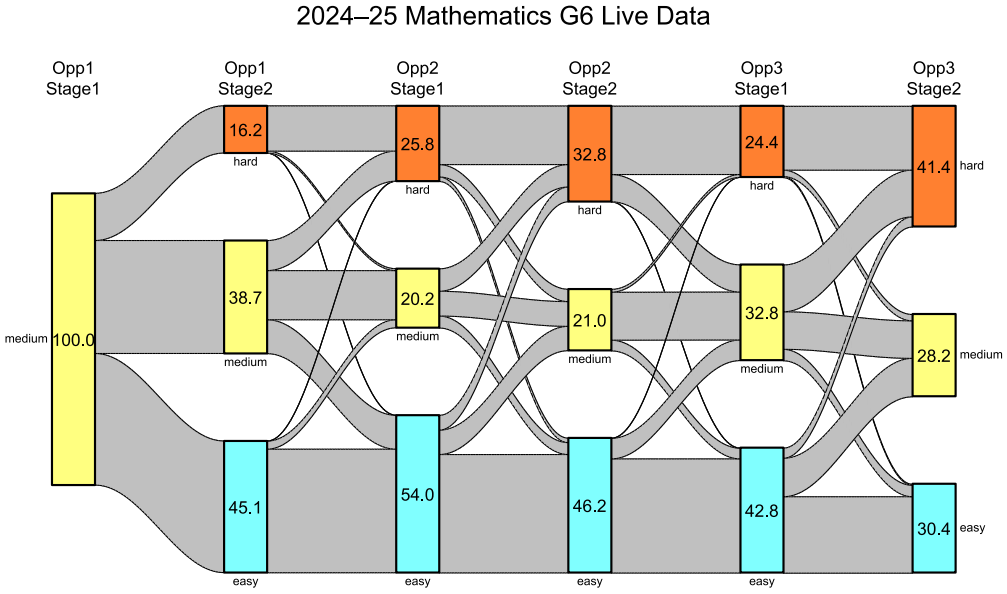


Figure E.3: Sankey Plot Grade 7 Mathematics, English

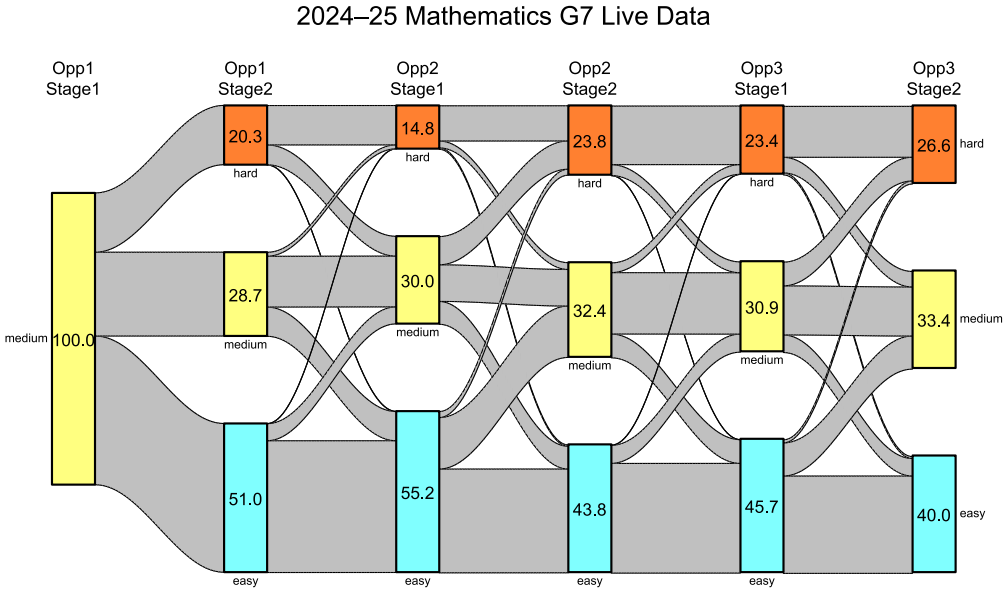


Figure E.4: Sankey Plot Grade 8 Mathematics, English

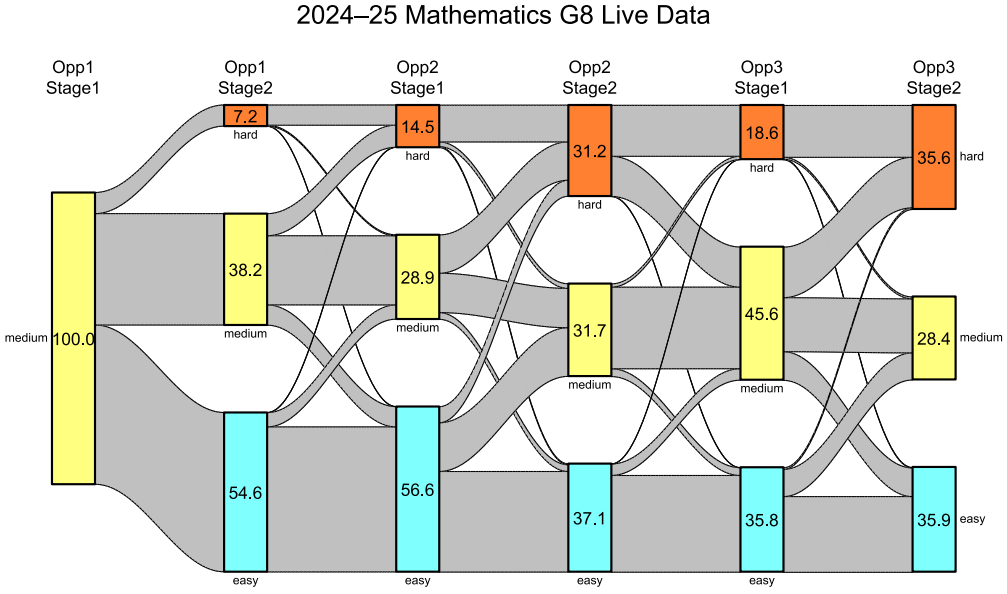


Figure E.5: Sankey Plot Grade 8 Social Studies, English