Hybrid Scoring Key Questions
March 2024
This presentation is intended to address the key questions the agency has received about implementing hybrid scoring.

1. What prompted the move to hybrid scoring?

2. How did TEA communicate these changes to stakeholders?

3. How does hybrid scoring fit into the assessment process?

4. Why are we seeing differences in ECR scores?
Key questions around hybrid scoring

1. What prompted the move to hybrid scoring?

2. How did TEA communicate these changes to stakeholders?

3. How does hybrid scoring fit into the assessment process?

4. Why are we seeing differences in ECR scores?
In 2023, STAAR was redesigned to improve alignment to the classroom experience.

In effective classrooms, teachers are...

<table>
<thead>
<tr>
<th>1</th>
<th>Coherently building students’ <strong>background knowledge and vocabulary</strong> in all subject areas...</th>
<th>Prioritize <strong>cross-curricular passages</strong> in RLA that reference topics that students have learned about in other classes</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>Asking students to <strong>write about what they read using evidence from text</strong>...</td>
<td>Include <strong>writing in all RLA tests</strong>, reflecting our updated TEKS, and having students <strong>write text-based responses</strong></td>
</tr>
<tr>
<td>3</td>
<td>Providing <strong>various open-ended formats</strong> for students to respond to questions...</td>
<td>Add new, <strong>non-multiple-choice questions</strong> that are more like questions teachers ask in class</td>
</tr>
<tr>
<td>4</td>
<td>Supporting the learning needs of all students by providing <strong>appropriate accommodations</strong>...</td>
<td>Move to online assessments that provide a <strong>full suite of robust accommodations</strong> for students with specific learning needs</td>
</tr>
<tr>
<td>5</td>
<td>Moving to <strong>online assessments</strong> supports all the changes above and provides faster test results to support accelerated learning.</td>
<td></td>
</tr>
</tbody>
</table>

**TEA**

The STAAR redesign added more open-ended questions that are similar to the questions teachers ask in class.

By making the test more aligned with the classroom experience, this increased the number of constructed response questions students access.

<table>
<thead>
<tr>
<th>Content Area</th>
<th>Item Type*</th>
<th>Pre-STAAR Redesign</th>
<th>Post-STAAR Redesign</th>
</tr>
</thead>
<tbody>
<tr>
<td>STAAR RLA (G4, G7, E1, E2)</td>
<td>SCR</td>
<td>-</td>
<td>1-2</td>
</tr>
<tr>
<td>STAAR RLA (Remaining titles)</td>
<td>ECR</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>STAAR RLA (Remaining titles)</td>
<td>SCR</td>
<td>-</td>
<td>1-2</td>
</tr>
<tr>
<td>STAAR Science</td>
<td>SCR</td>
<td>-</td>
<td>1-2</td>
</tr>
<tr>
<td>STAAR Social Studies</td>
<td>SCR</td>
<td>-</td>
<td>1-2</td>
</tr>
</tbody>
</table>

*SCR = Short Constructed Response; ECR = Extended Constructed Response
The significant increase in written responses required the move to hybrid scoring to meet budget and timeline limitations.

With 6-7x more constructed responses to grade annually for STAAR, maintaining full human scoring would have cost $15-20M more per year.
Key questions around hybrid scoring

1. What prompted the move to hybrid scoring?
2. How did TEA communicate these changes to stakeholders?
3. How does hybrid scoring fit into the assessment process?
4. Why are we seeing differences in ECR scores?
TEA communicated the transition with stakeholders in advance.

Aug. 2022

Testimony at House Public Education Committee
- Commissioner testified that automated scoring is a necessary step to control costs while ensuring reliability
- Relying solely on human scoring for the increased number of constructed responses as part of STAAR redesign would require $15-20M/year

Sept. 2023

Statewide Announcement to Testing Coordinators
- TEA announced the implementation of SY23-24 hybrid scoring during the annual test coordinator training
- This presentation was recorded and posted in the Texas Assessment Program Learning Management System for district personnel access

Nov. 2023

Presentations at Conferences and with Stakeholder Groups
- TEA presented about the hybrid scoring model at the 2023 Texas Assessment Conference
- TEA provided information in other stakeholder group presentations, e.g., Texas Science Education Leadership Association (TSELA), Texas Social Studies Supervisors Association (TSSSA), Texas Council of Teachers of English Language Arts (TCTELA), Coalition of Reading and English Supervisors of Texas (CREST)

Dec. 2023

Release of STAAR Scoring Process Document
- The detailed scoring process document was published on TEA's website

Example slide from TEA presentations:

TEA ensures that the scoring model for constructed-response questions is valid and reliable. We are consistently exploring how to improve the process.

Beginning in December 2023, TEA will implement a hybrid-scoring model that incorporates automated scoring alongside our human expert scorers.

- Hybrid scoring is not artificial intelligence
- Educators, including Texas educators, will continue to be involved in all parts of the scoring process
- Educator committees will give input on what scoring methods should be used
- Human experts will provide feedback and adjust the scoring model along with our human expert scorers

More information will be available in the next few months.
Key questions around hybrid scoring

1. What prompted the move to hybrid scoring?

2. How did TEA communicate these changes to stakeholders?

3. How does hybrid scoring fit into the assessment process?

4. Why are we seeing differences in ECR scores?
Creating high-quality state assessments is a rigorous process that includes educators across multiple steps.

**Assessment design**
1. Assessment design framework is developed*
2. Assessment blueprints are developed*
3. Educator advisory committees provide feedback*

**Passage and item development**
4. Professional item writers develop new passages & items
5. TEA content specialists review passages & items
6. Educator external review committees review passages & items

**Test construction**
7. Items are field tested
8. Field tested items and statistical data are reviewed
9. Items with good data are added to the item bank
10. Operational test forms are created from item bank
11. Items are accommodated
12. Educator “rangelining” to support consistent grading of essays

**Admin & QC**
13. Assessments are administered
14. Performance review

**Field testing**
7. Items are field tested
8. Field tested items and statistical data are reviewed
9. Items with good data are added to the item bank
10. Operational test forms are created from item bank
11. Items are accommodated
12. Educator “rangelining” to support consistent grading of essays

**Scoring and reporting**
15. Standard setting is completed with educator input*
16. Assessments are scored
17. Score reporting occurs
18. Technical reports are written

*Does not occur every year

It takes over two years from assessment design to scoring and reporting.

Thus, any changes to assessments must be planned far in advance.
Field testing serves as a building block for assessment construction and is not related to the move to hybrid scoring.

Field testing collects data on items to ensure that they are unbiased and measure what they’re supposed to.

**Assessment design**
1. Assessment design framework is developed*
2. Assessment blueprints are developed*
3. Educator advisory committees provide feedback*

**Passage and item development**
4. Professional item writers develop new passages & items
5. TEA content specialists review passages & items
6. Educator external review committees review passages & items

**Field testing**
7. Items are field tested
8. Field tested items and statistical data are reviewed
9. Items with good data are added to the item bank

**Test construction**
10. Operational test forms are created from item bank
11. Items are accommodated
12. Educator "rangepinding" to support consistent grading of essays

**Admin & QC**
13. Assessments are administered
14. Performance review

**Scoring and reporting**
15. Standard setting is completed with educator input*
16. Assessments are scored
17. Score reporting occurs
18. Technical reports are written

*Does not occur every year
The purpose of field testing is to populate a bank that is diverse and free of bias.

To best measure student performance, TEA needs to build an item bank that represents all items from the least to most difficult spectrum:

When we field test, we are testing the item, not the students.
Most field test items are embedded within the STAAR assessments, with the exception of extended constructed response (ECR) items. A student typically interacts with field test items within their STAAR test. They would not know which items are a part of their assessment and which items are field test items. Field test items do not count towards a student’s score.

STAAR tests have embedded field test items.

Extended constructed response (ECR) items cannot be placed within a STAAR test due to its length. Therefore, Stand Alone Field Tests (SAFT) have been established for ECRs. This is the same process for field testing ECRs before and after the STAAR redesign.

Selected students take an assessment that includes the ECR field test mixed in with other items. This is to simulate a testing environment and provide items that help link the field test items back to the STAAR scale.
After the field test event, each constructed response item is scored against a rubric by two humans.

**Short Constructed Response (SCR)**
- Rubrics are developed alongside passages and reviewed during educator item review meetings
- Rubrics range from 0-2 points
- There are 2 human scorers involved
  - RLA writing SCRs use 0-1 point rubrics
  - RLA reading SCRs use 0-2 point rubrics
  - Science and Social Studies SCRs use 0-2 point rubrics
  - The first human’s score is the score used
  - The second human’s score is kept for quality assurance purposes

**Extended Constructed Responses (ECR)**
- Rubrics were established during STAAR redesign and are static
- Rubrics are 5 points in total
  - Development and organization (max 3 pts)
  - Language and conventions (max 2 pts)
- Scores from each human scorer is combined for a max score of 10
  - If the scores are adjacent, scores are summed up
  - If not, it goes through adjudication process where new score received is doubled

Rubrics are employed to ensure consistency in scoring for items that are open-ended.
(See examples in the appendix.)
All of the humans involved in the scoring process are highly trained and calibrated.

To qualify as a rater, one must have a 4-year college degree and experience teaching at the assigned grade level.

Raters undergo rigorous training to learn how to use the standardized rubric to score student responses.

To pass training, they must accurately rate example responses that have already been scored.

Each certified rater’s performance is calibrated at regular intervals to ensure that all responses are graded consistently across Texas.

More on this later
Field test constructed response items continue to be scored against a rubric by two humans.

### Short Constructed Response (SCR)

Two humans score each response.
Scorer 1’s rating is the score of record.
Scorer 2’s rating is used for the purpose of auditing / quality control.

### Extended Constructed Responses (ECR)

**Scenario 1:** If two raters show exact or adjacent (within 1 point) agreement on a response, then the scores are summed together to create the score of record.

<table>
<thead>
<tr>
<th>Scorer 1</th>
<th>Scorer 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>4</td>
</tr>
</tbody>
</table>

= 8

**Scenario 2:** If two raters show more than 1 point difference in rating a response, then a scoring leader takes over and assigns the score of record.

<table>
<thead>
<tr>
<th>Scorer 1</th>
<th>Scorer 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>2</td>
</tr>
</tbody>
</table>

x 2

= 8

Scoring Leader
Within an administration period, any rater that does not meet standard is removed from scoring.

3% of what scorers rate are validity papers that are inserted for scorer calibration during the scoring window.

If this scorer doesn’t pass recalibration, they are exited.

If a scorer does not maintain at least at 65% exact agreement and 95% adjacent agreement during the scoring window, they cannot remain as a scorer for that admin. If they fail re-calibrations, they will have to try again in the next administration.
Pre-equating is a part of the test construction process that occurs after field testing.

**Assessment design**
1. Assessment design framework is developed*
2. Assessment blueprints are developed*
3. Educator advisory committees provide feedback*

**Passage and item development**
4. Professional item writers develop new passages & items
5. TEA content specialists review passages & items
6. Educator external review committees review passages & items

**Field testing**
7. Items are field tested
8. Field tested items and statistical data are reviewed
9. Items with good data are added to the item bank

**Test construction**
10. Operational test forms are created from item bank
11. Items are accommodated
12. Educator "rangefinding" to support consistent grading of essays

**Admin & QC**
13. Assessments are administered
14. Performance review

**Scoring and reporting**
15. Standard setting is completed with educator input*
16. Assessments are scored
17. Score reporting occurs
18. Technical reports are written

*Does not occur every year

Equating is the statistical process by which the performance of items from different administrations can be compared by placing the items on the same scale. Equating ensures students taking the STAAR receive the correct scale score.
Equating ensures that the STAAR test is the same level of difficulty each year.

While individual items can be easier or harder in a given year, the mix of item difficulty is balanced across years by using field test results.

Click to see STAAR FAQ video: “How do we know the STAAR test is the same level of difficulty from year to year?”
Scoring and reporting is the final process of the assessment lifecycle.

**Assessment design**
1. Assessment design framework is developed*
2. Assessment blueprints are developed*
3. Educator advisory committees provide feedback*

**Passage and item development**
4. Professional item writers develop new passages & items
5. TEA content specialists review passages & items
6. Educator external review committees review passages & items

**Test construction**
12. Educator "ranging" to support consistent grading of essays
11. Items are accommodated
10. Operational test forms are created from item bank

**Field testing**
9. Items with good data are added to the item bank
8. Field tested items and statistical data are reviewed
7. Items are field tested

**Admin & QC**
13. Assessments are administered
14. Performance review
15. Standard setting is completed with educator input*

**Scoring and reporting**
16. Assessments are scored
17. Score reporting occurs
18. Technical reports are written

*Does not occur every year

We will go into detail on the human scoring process and how the automated scoring engine replicates this process consistently with heavy human oversight.
Prior to hybrid scoring, all constructed response items were scored the same way as field testing, but the process has changed.

### Scoring Process Today

<table>
<thead>
<tr>
<th>Item Type</th>
<th>Field Testing</th>
<th>Operational Scoring</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-Constructed Response Items</td>
<td>Machine Scored</td>
<td>Machine Scored</td>
</tr>
</tbody>
</table>
| Short Constructed Response (SCR) Items | 2 human scorers (one score of record, one for auditing) | Hybrid Scoring
All scores go through auto scoring engine; 25% are double human scored (one score of record, one for auditing). |
| Extended Constructed Response (ECR) Items | 2 human scorers (scores are combined for total score) | Hybrid Scoring
All scores go through auto scoring engine; 25% are double human scored (scores are combined for total score). |

While this process has changed, TEA is ensuring that the quality of scoring remains the same.
Up until the CR scoring event itself, each part of the CR scoring prep process relies solely on human input.

- All constructed responses are double human scored.
- Humans identify anchor responses which are field test responses that exemplify responses at each rubric score point.
- Human scorers are trained through the anchor responses.
- The automated scoring engine is programmed through ~3,000 hand-scored field test responses and human-identified anchor responses.
- Hybrid scoring: Human raters support the automated scoring engine through calibration checks and scoring non-standard responses during the administration window.
The auto scoring engine (ASE) goes through a rigorous programming process that is led and checked by humans.

*For each item being scored...*

- The engine uses a sample of ~3,000 human scored responses from the field test for programming.
- The engine analyzes the responses to identify common patterns and is programmed to emulate how humans would score.
- TEA evaluates the performance for each item and compares it to how humans would score.
- The engine is monitored throughout the scoring cycle to ensure that it remains calibrated to the anchor set.

Similar to human scorers who need to be constantly calibrated throughout the scoring window, there is a parallel process for the ASE.
TEA conducted a proof-of-concept study with STAAR Spring 2023 operational data before implementing hybrid scoring.

- Spring 2023 constructed response items were scored entirely by humans. The study was conducted after score reports were sent out to districts.
- The study “re-scored” constructed response items with the automated scoring engine and compared how closely the engine performed to humans.
- Five subgroups were examined (Male, Female, Black, Hispanic/Latino, White).
- The proof-of-concept study was successful and found that the automated scoring engine met the performance criteria to be implemented operationally.
- A detailed technical report on the Spring 2023 study can be found on the Assessment Reports and Studies webpage.

*Report can be found under "Additional Reports and Studies".

**Example: Spring STAAR 2023 ECR Conventions – Exact Agreement for ASE Model 2**

<table>
<thead>
<tr>
<th>Item</th>
<th>Human-Human</th>
<th>Human-Engine</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>72%</td>
<td>74%</td>
<td>2%</td>
</tr>
<tr>
<td>2</td>
<td>71%</td>
<td>72%</td>
<td>1%</td>
</tr>
<tr>
<td>3</td>
<td>67%</td>
<td>66%</td>
<td>-1%</td>
</tr>
<tr>
<td>4</td>
<td>67%</td>
<td>72%</td>
<td>5%</td>
</tr>
<tr>
<td>5</td>
<td>67%</td>
<td>73%</td>
<td>6%</td>
</tr>
<tr>
<td>6</td>
<td>71%</td>
<td>76%</td>
<td>5%</td>
</tr>
</tbody>
</table>

TEA analyzed a number of criteria, including exact agreement (shown here), adjacent agreement, and score distributions to ensure the engine performs as expected.
Similarly, STAAR December 2023 data confirms the ASE performed as expected.

December 2023 SCR Exact Agreement Results

<table>
<thead>
<tr>
<th>Grade</th>
<th>Item</th>
<th>Subject</th>
<th>Human-Human Exact Agreement</th>
<th>Human-Engine Exact Agreement</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>1</td>
<td>Biology</td>
<td>96.5%</td>
<td>97.1%</td>
<td>0.6%</td>
</tr>
<tr>
<td>9</td>
<td>2</td>
<td>Biology</td>
<td>95.8%</td>
<td>95.1%</td>
<td>-0.7%</td>
</tr>
<tr>
<td>9</td>
<td>3</td>
<td>Read</td>
<td>82.6%</td>
<td>86.8%</td>
<td>4.3%</td>
</tr>
<tr>
<td>9</td>
<td>4</td>
<td>USH</td>
<td>86.2%</td>
<td>92.4%</td>
<td>6.2%</td>
</tr>
<tr>
<td>9</td>
<td>5</td>
<td>USH</td>
<td>95.1%</td>
<td>97.1%</td>
<td>2.0%</td>
</tr>
<tr>
<td>9</td>
<td>6</td>
<td>Write</td>
<td>95.6%</td>
<td>97.3%</td>
<td>1.7%</td>
</tr>
<tr>
<td>10</td>
<td>8</td>
<td>Read</td>
<td>76.2%</td>
<td>79.1%</td>
<td>2.9%</td>
</tr>
<tr>
<td>10</td>
<td>9</td>
<td>Write</td>
<td>92.2%</td>
<td>97.0%</td>
<td>4.8%</td>
</tr>
</tbody>
</table>

December 2023 ECR Exact Agreement Results

<table>
<thead>
<tr>
<th>Grade</th>
<th>Item</th>
<th>Subject</th>
<th>Human-Human Agreement</th>
<th>Human-Engine Agreement</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>7</td>
<td>Read</td>
<td>Convention 69.3%</td>
<td>Convention 73.9%</td>
<td>4.6%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Ideas 66.5%</td>
<td>Ideas 70.5%</td>
<td>4.0%</td>
</tr>
<tr>
<td>10</td>
<td>10</td>
<td>Read</td>
<td>Convention 78.4%</td>
<td>Convention 87.9%</td>
<td>9.5%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Ideas 73.5%</td>
<td>Ideas 77.1%</td>
<td>3.6%</td>
</tr>
</tbody>
</table>

- TEA used the hybrid scoring approach to score all constructed responses in December 2023.
- Constructed response items saw similar human-engine exact agreement to human-human exact agreement.
- TEA will continue to monitor hybrid scoring during every administration to ensure ASE produces accurate scores.

Lower exact agreement levels are expected with ECRs compared to SCRs due to larger range of possible points.
The Texas hybrid scoring model uses an automated scoring engine to augment the work of human scorers.

**Scenario 1:**
Auto scoring engine assigns score of record

- Approx. 75% responses

**Scenario 2:**
Engine flags responses for double human scoring (assignment of condition codes or low confidence*)

- At least 25% responses

**Scenario 3:**
Random sample of responses for double human scoring

Human scorers are used to monitor the engine during the admin window for quality control purposes.

Note: Any student responses that are routed for human scoring maintain the score assigned by humans as the score of record. Human scoring will also go through the adjudication process if needed.

*Condition codes that get sent for human scoring are those flagged for unusual patterns; low confidence responses are often those responses that are on the border between two score points.
The ASE assigns condition codes to some responses, which are each routed to two trained human scorers.

Condition codes indicate that a response uses just a few words, uses mostly duplicated text, is written in another language, consists primarily of text from the passage, uses vocabulary that does not overlap with the vocabulary in the subset of responses used to program the ASE, or uses language patterns that are reflective of off-topic or off-task responses.

The purpose of this routing is to ensure that these unusual responses receive fair and accurate scoring. The score assigned by the human scorer is kept as the score of record for any student response that is routed for human scoring.

Note: Any student responses that are routed for human scoring maintain the score assigned by humans as the score of record. Human scoring will also go through the adjudication process if needed.
Automated scoring technology* is over a decade old and is widely used, including in Texas.

10+ years

amount of time technology for automated scoring engine has been around

180K+

Texas students

annually use the Texas Success Initiative Assessment (TSIA) to meet their graduation requirement, which relies on automated scoring technology

21+

states

currently employ auto scoring for their state assessments

*This kind of technology is different from AI in that AI is a computer using progressive learning algorithms to adapt, allowing the data to do the programming and essentially teaching itself. Instead, the automated scoring engine is a closed database with student response data accessible only by TEA and, with strict contractual privacy controls, its assessment contractors, Cambium and Pearson.
The Texas Assessment Program continues to strive for assessment development transparency.

**Assessment design**
- 1. Assessment design framework is developed*
- 2. Assessment blueprints are developed*
- 3. Educator advisory committees provide feedback*

**Passage and item development**
- 4. Professional item writers develop new passages & items
- 5. TEA content specialists review passages & items
- 6. Educator external review committees review passages & items

**Test construction**
- 7. Items are field tested
- 8. Field tested items and statistical data are reviewed
- 9. Items with good data are added to the item bank

**Field testing**
- 10. Operational test forms are created from item bank
- 11. Items are accommodated

**Admin & QC**
- 12. Educator "rangefinding" to support consistent grading of essays
- 13. Assessments are administered
- 14. Performance review

**Scoring and reporting**
- 15. Standard setting is completed with educator input*
- 16. Assessments are scored
- 17. Score reporting occurs
- 18. Technical reports are written

*Does not occur every year

It is uncommon for states to release all test items for primary spring test administrations on an annual basis, yet Texas has made that commitment.

June and December STAAR tests cannot be released annually because items are re-used in later tests. To release June and December tests annually, TEA would need to develop and field test 3x the number of items.
Key questions around hybrid scoring

1. What prompted the move to hybrid scoring?

2. How did TEA communicate these changes to stakeholders?

3. How does hybrid scoring fit into the assessment process?

4. Why are we seeing differences in ECR scores?
There are two timelines and changes to account for when comparing ECR scores.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Pre-STAAR Redesign</td>
<td>Post-STAAR Redesign</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>The STAAR Redesign implementation went live Spring 2023 and drove two things – a change in the type of ECR questions asked, and as a result, the scoring rubric.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Human Scoring</td>
<td>Hybrid Scoring</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>The hybrid scoring transition started in December 2023, and student responses went from fully human scoring to hybrid scoring.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The STAAR redesign changed ECRs and their rubrics, thus impacting ECR scores.
STAAR Redesign: Based on stakeholder feedback, ECRs were redesigned to ask students to use evidence from text.

**Pre-STAAR Redesign**
*Writing responses to standalone prompts*

Read the following quotation.

The reason most people never reach their goals is that they don’t define them, or ever seriously consider them as believable or achievable. Winners can tell you where they are going, what they plan to do along the way, and who will be sharing the adventure with them.

—Denis Waitley

Successful people often set clear goals and understand the specific steps needed to achieve them. Think carefully about this statement.

Write an essay explaining the best way to achieve a goal.

Be sure to—
- clearly state your thesis
- organize and develop your ideas effectively
- choose your words carefully
- edit your writing for grammar, mechanics, and spelling

**STAAR Redesign Implementation**
*Writing responses using evidence from text*

Read the selection and choose the best answer to each question.

**Back in Time: The National Road**

1. The National Road, in many places known as Route 40, was built between 1811 and 1838 to reach the western settlers. It was the first federally funded road in U.S. history. George Washington and Thomas Jefferson believed that a tame Appalachian road was necessary for settling the young country. In 1835 Congress authorized construction of the road and President Jefferson signed the act establishing the National Road. It would connect Cumberland, Maryland, to the Ohio River.

   In 1811 the first contract was awarded and the first 10 miles of road built. By 1818 the road was completed to following, and mail coaches began using the road. By 1831 the federal government assumed part of the road’s responsibility to the states through which it ran. Toll gates and tollhouses were then built by the states, with the federal government taking responsibility for road repairs.

   As work on the road progressed, a settlement pattern developed that is still visible. Original towns and villages are found along the National Road, many barely touched by the passage of time. The road, also called the Cumberland Road, National Pike, and other names, became Main Street in these early settlements, earning the nickname “The Main Street of America.” The height of the National Road’s popularity came in 1822 when it was celebrated in song, stories, painting, and poetry. During the 1840s popularity soared again.

   Travelers and draysmen, verstaked sound, craddled the mass and traveled along the route. Route robotics against resistence from frontier hosts to the east most certainly.

   ![Map of Historic National Road]

2. The National Road was built to facilitate trade and transportation.
   
   - True
   
   - False

   3. The National Road was completed in 1818.
   
   - True
   
   - False

   4. The National Road was used to transport goods during the 1840s.
   
   - True
   
   - False

   5. Toll gates were built by the states with the federal government taking responsibility for road repairs.
   
   - True
   
   - False

Spring 2022 English 1 EOC Example

Spring 2023 English EOC Example
STAAR Redesign: In addition, the educator-approved new rubrics for writing prompts introduce a possible score of zero.

Pre-STAAR Redesign

Lowest score that could be assigned to a response was a 2.

Zeros were only reserved for unscorable responses (i.e., blanks, random characters).

STAAR Redesign Implementation

Students could earn a zero through the rubric or through a non-scoreable response.

Example rubric from the 2023 English I Constructed Response Scoring Guide
To make a proper comparison, we isolated the scoring data to only include retesters who existed in each test administration.

<table>
<thead>
<tr>
<th></th>
<th>Pre-STAAR Redesign</th>
<th>Human Scoring</th>
<th>Post-STAAR Redesign</th>
<th>Hybrid Scoring</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zero on ECR</td>
<td>9%</td>
<td>10%</td>
<td>11%</td>
<td>50%</td>
</tr>
<tr>
<td>Approaches or Above on EOC</td>
<td>21%</td>
<td>16%</td>
<td>22%</td>
<td>33%</td>
</tr>
<tr>
<td>Number of Retesters</td>
<td>82,755</td>
<td>6,932</td>
<td>125,320</td>
<td>95,551</td>
</tr>
<tr>
<td>Zero on ECR</td>
<td>7%</td>
<td>9%</td>
<td>9%</td>
<td>62%</td>
</tr>
<tr>
<td>Approaches or Above on EOC</td>
<td>28%</td>
<td>8%</td>
<td>25%</td>
<td>28%</td>
</tr>
<tr>
<td>Number of Retesters</td>
<td>59,617</td>
<td>44,687</td>
<td>89,918</td>
<td>64,857</td>
</tr>
</tbody>
</table>

Percent of Retesters Who Received a Zero Score on the Extended Constructed Response Item, Percent of Retesters Who Achieved Approaches Grade Level or Better on the EOC as a Whole, and Number of Retesters
The STAAR redesign resulted in notably higher zero rates on ECRs in 2023 but did not impact overall test performance.

<table>
<thead>
<tr>
<th></th>
<th>Pre-STAAR Redesign</th>
<th>Post-STAAR Redesign</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Human Scoring</td>
<td>Hybrid Scoring</td>
</tr>
<tr>
<td>Zero on ECR</td>
<td>9%</td>
<td>10%</td>
</tr>
<tr>
<td>Approaches or Above on EOC</td>
<td>21%</td>
<td>16%</td>
</tr>
<tr>
<td>Number of Retesters</td>
<td>82,755</td>
<td>6,932</td>
</tr>
</tbody>
</table>

Due to the change in ECR questions and introducing a new possible score of zero on the rubric, Spring 2023 has higher zero rates than Spring 2022. No change in scoring had occurred.

However, our test equating process (see slide 19) means that while individual items can be easier or harder in a given year, the mix of item difficulty is balanced across years, so overall performance on the test stayed the same or increased.

Percent of Retesters Who Received a Zero Score on the Extended Constructed Response Item, Percent of Retesters Who Achieved Approaches Grade Level or Better on the EOC as a Whole, and Number of Retesters.
The variation in zero rates seen during the move to hybrid scoring are much smaller and are considered normal.

<table>
<thead>
<tr>
<th>Year</th>
<th>Pre-STAAR Redesign</th>
<th>Post-STAAR Redesign</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Human Scoring</td>
<td>Hybrid Scoring</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zero on ECR</td>
<td>82,755</td>
<td>6,932</td>
</tr>
<tr>
<td>Approaches or Above on EOC</td>
<td>59,617</td>
<td>44,687</td>
</tr>
</tbody>
</table>

However, the mix of item difficulty is balanced across years through equating (see slide 19). Overall performance on the test has stayed largely consistent.

Individual items (like ECRs) can be easier or harder in a given year, so we expect to see some variation in ECR zero rates across tests.

Percent of Retesters Who Received a Zero Score on the Extended Constructed Response Item, Percent of Retesters Who Achieved Approaches Grade Level or Better on the EOC as a Whole, and Number of Retesters.
For transparency purposes, TEA provided LEAs with additional December 2023 English 1 and English 2 ECR results information. Unlike Spring STAAR tests, which are released annually, June and December STAAR tests are not released so items can be re-used in later tests. To support transparency, TEA developed a two-step process in March to give districts more insight into student performance on December 2023 ECRs:

**Step 1 (444 LEAs):** Upon request, TEA provided frequency distributions of reasons for an ECR receiving a score of 0 within the requesting district –

<table>
<thead>
<tr>
<th>Reason</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>No response</td>
<td>14%</td>
</tr>
<tr>
<td>Not scored due to a condition code (not enough words, duplicated text, written in another language, consisting mostly of copied text from the passage, or writing is off-topic or off-task)</td>
<td>30%</td>
</tr>
<tr>
<td>Received a score of 0 according to the rubric (eligible for Step 2)</td>
<td>55%</td>
</tr>
</tbody>
</table>

**Step 2 (109 LEAs):** After Step 1, LEAs can also opt to schedule an appointment to view the responses in person

Only responses that receive a score of zero are made available for viewing;

Received positive feedback from district personnel on this process
Thank you!

- More details on the hybrid scoring study performed on 2023 STAAR data can be found in "Assessment Reports and Studies" under the section titled Additional Reports and Studies.

- If you have questions about hybrid scoring, contact TEA using the Student Assessment Help Desk.
ECR Rubric Example: English 1

More RLA scoring resources can be found here: https://tea.texas.gov/student-assessment/testing/staar/staar-reading-language-arts-resources
## SCR Rubric Examples: US History and Grade 8 Science

### United States History Prompt
**Prompt:** What was President Theodore Roosevelt’s Big Stick policy **AND** what was an example of this policy?

### Item-Specific Rubric
**Score: 2**
Response includes specific details in reference to a description and one example:

**Description:**
- The Big Stick policy used military readiness and diplomacy to protect the Western Hemisphere from foreign intervention.

**Examples:**
- Roosevelt used this diplomacy to restrain European countries from threatening Latin American countries.
- Roosevelt issued this policy to enforce the Monroe Doctrine and become the international police power of the Americas.
- The United States increasingly used force to justify intervention in several countries, including securing the Panama Canal Zone, Cuba, Nicaragua, Haiti, and the Dominican Republic.

**Score: 1**
Response provides only half of the correct details.

**Score: 0**
Does not provide a response, or the response is incorrect or irrelevant.

---

### Grade 8 Science Short Constructed Response Prompt

**Prompt**
Sodium sulfate (Na₂SO₄) is used to produce many products.
Which elements are represented in the formula **AND** how many atoms of each element are represented in the formula?

Read the question carefully. Then enter your answer in the box provided.

### Item-Specific Rubric
**Score: 2**
The student response includes:
- There are a total of 7 atoms representing three elements in the formula **AND**
- The elements are sodium (2 atoms), sulfur (1 atom), and oxygen (4 atoms)

**Score: 1**
The student answers half of the question correctly.

**Score: 0**
The response is incorrect or irrelevant.

---

More social studies scoring resources can be found here: [https://tea.texas.gov/student-assessment/testing/staar/staar-social-studies-resources](https://tea.texas.gov/student-assessment/testing/staar/staar-social-studies-resources)

More science scoring resources can be found here: [https://tea.texas.gov/student-assessment/testing/staar/staar-science-resources](https://tea.texas.gov/student-assessment/testing/staar/staar-science-resources)