Scoring Process for STAAR Constructed Responses

Constructed-Response Questions

The State of Texas Assessments of Academic Readiness (STAAR®) is a standardized academic achievement test designed to measure the extent to which a student has learned and is able to apply the defined knowledge and skills of the Texas Essential Knowledge and Skills (TEKS) for each tested grade, subject, and course. STAAR includes open-ended questions that reflect classroom test questions and allow students more ways to show their understanding of the TEKS. Two specific types of open-ended questions, known as extended constructed-response (ECR) questions and short constructed-response (SCR) questions, require students to draft written responses. ECR questions are included on all STAAR reading language arts (RLA) assessments—grades 3–8, English I, and English II. SCR questions are included on:

- grades 3–8 RLA,
- grades 5 and 8 science,
- grade 8 social studies,
- English I,
- English II,
- Biology, and
- U.S. History.

Student responses to SCR and ECR questions are evaluated using a holistic scoring process. The responses are evaluated as a whole based on pre-established criteria identified and detailed in the scoring rubrics. The scoring rubrics are used to determine the effectiveness of each written response. This document outlines the scoring process the Texas Education Agency (TEA) uses for STAAR SCRs and ECRs.

Overview of Scoring

Student responses to short constructed-response (SCR) questions and extended constructed-response (ECR) questions are scored using a hybrid scoring model. This means that the responses are initially scored by an automated scoring engine (ASE) and at least 25 percent of student responses are then routed to human scorers. The scores generated by human scorers are used throughout the scoring window to monitor scores that are being generated by the ASE. This process is similar to a second reader in human scoring models.

NOTE: Student responses to SCR and ECR questions for STAAR Spanish assessments are 100 percent human scored and do not get scored by the ASE.

Pre-scoring Activities

Before scoring begins, a few activities occur which optimize how TEA assesses the TEKS on STAAR.

- Developing New Questions
  Development activities begin with content and assessment experts developing the passages and questions that will assess the TEKS. Committees of Texas educators (e.g., teachers, instructional specialists, district coordinators, principals, etc.) review passages and questions and provide input

and feedback. Once passages and questions are reviewed and edited accordingly, they are approved by TEA and are ready for field testing.

- **Field Testing New Questions**
  Field testing questions is the process of gathering performance data on test questions, but it does not impact students' scores on STAAR. TEA conducts field testing of newly developed questions each year either through embedded field-test questions or through stand-alone field tests.

Scoring Activities

The steps outlined below each have an important role in the overall scoring process.

```
┌─────────────────┐
│ Anchor Approval │
│     Meeting     │
└─────────────────┘
    └─┐
   ┌──────────────────────┐
   │ Training Human Scorers│
   │       and ASE        │
   └──────────────────────┘
        └─┐
       ┌──────────┐
       │ Scoring  │
       └──────────┘
            └─┐
          ┌──────────┐
          │ Reporting│
          └──────────┘
               └─┐
             ┌──────────┐
             │ Rescores │
             └──────────┘
```

Step 1: Anchor Approval Meeting

A selected group of student responses are collected during field testing and are scored using scoring rubrics, i.e., the TEA-approved scoring criteria. Student responses from the field-tested questions are taken to anchor approval meetings. The purpose of anchor approval meetings is to apply the scoring rubrics to sample student responses and establish scoring boundaries. Student responses used to establish scoring boundaries are called "anchor responses", and these responses clearly exemplify each score point in the rubric. During the anchor approval meetings, educators identify anchor responses, provide consensus scores for these responses, and discuss the rationale for the scores these responses receive. These discussions are used to develop training notes or annotations for the responses. After the meeting, item sets intended for use as practice and qualification sets are created based on the approved anchor responses. Educators are trained and qualify as scorers using these item sets.

Step 2: Training Human Scorers and ASE

Both human scorers and the ASE are trained with TEA-approved student responses.

- **Training Human Scorers**
  Scorers are assigned to score specific grade/subject or course SCRs or ECRs based on their experience and educational background. For each question type, human scorers go through a

similar process. They study and review the anchor responses. Then scorers are given the opportunity to practice on the different ways students are responding to the question. Scoring directors and scoring supervisors are available for clarification and guidance during the process and support the scorers in their training as needed.

After completing practice sets, human scorers must demonstrate a complete understanding of the rubrics by successfully completing a qualification set. The qualification set is a quality measure to ensure that scorers are following the rubric and can accurately apply scores. Only scorers who complete the training and pass the qualification set may score student responses. Scorers who do not pass the qualification set are not allowed to score student responses for STAAR.

- o RLA SCR Training
  The SCRs for the writing section of the RLA assessment are scored using a generic 0- or 1-point rubric. The SCRs for the reading section use an item-specific 0- to 2-point rubric. Human scorers are trained to apply the 1-point rubric or 2-point rubric using anchor responses or anchor sets that include student responses at each score point. Scorers then complete practice sets where they score student responses for their assigned grade or course. Both the anchor and practice sets include annotations explaining why and how the score was assigned. The scoring rubric used for each SCR question on a STAAR released test can be found in the constructed response scoring guides available on the STAAR Reading Language Arts Resources webpage.
- o RLA ECR Training
  The ECRs for RLA are scored using an item-specific 5-point rubric that identifies scores based on two traits—development and organization of ideas (up to 3 points) and language conventions (up to 2 points). Human scorers are trained to apply the 5-point rubric using an anchor set that includes student responses at each score point. Scorers then complete practice sets where they score student responses for their assigned grade or course. Both the anchor and practice sets include annotations explaining why and how the score was assigned. The STAAR RLA scoring rubrics for ECRs are available on the STAAR Reading Language Arts Resources webpage.
- o Science and Social Studies SCR Training
  Both science and social studies SCRs are scored using 0- to 2-point item-specific rubrics. Human scorers are trained to apply the rubric by using the approved anchor sets and then must successfully score at least two practice sets. Both the anchor and practice sets include annotations explaining why and how the score was assigned. The scoring rubric used for each SCR question on a STAAR released test can be found in the constructed response scoring guides available on the corresponding STAAR Science Resources and STAAR Social Studies Resources webpages.

- Training the ASE
  The ASE is trained on student responses and human scores from the field-test data. It is trained to emulate how humans would score student responses for each constructed-response question. TEA evaluates the performance of the ASE for each question using a subset of student responses and corresponding scores that the ASE has not processed. Performance is examined relative to how humans would score the responses to each question; TEA requires the ASE to agree with

human scorers at the same rate human scorers agree with one another and that the distribution of ASE scores is similar to the distribution of human scores.
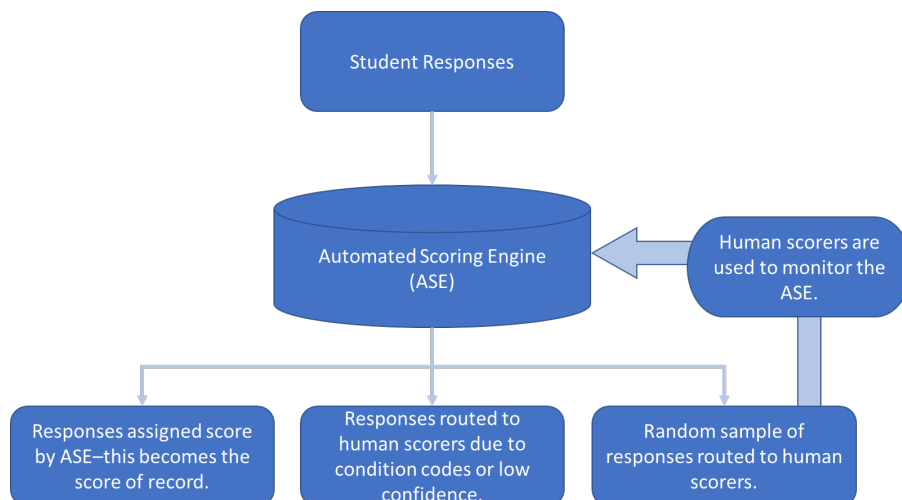
As part of the training process, the ASE calculates confidence values that indicate the degree to which the ASE is confident the score it has assigned matches the score a human would assign. The ASE also identifies student responses that should receive condition codes. Condition codes indicate that a response is blank, uses too few words, uses mostly duplicated text, is written in another language, consists primarily of stimulus material, uses vocabulary that does not overlap with the vocabulary in the subset of responses used to train the ASE, or uses language patterns that are reflective of off-topic or off-task responses.

Step 3: Scoring

All constructed-response questions for the English language version of STAAR are scored using a hybrid scoring model. All constructed-response questions for the Spanish language version of STAAR are scored using only human scorers.

The process for the hybrid scoring model is outlined below.

- All student responses are scored for the first time by the ASE.
- At least 25 percent of student responses for each grade/subject and course are then routed to trained human scorers. This set of second scores by human scorers is used to monitor the ASE scoring and is similar to what is done in human scoring to examine how often scorers agree with one another.
- Student responses that the ASE assigns condition codes (e.g., use words not seen in the responses used to train the ASE) or that are identified as "low confidence" are routed to trained human scorers. The low confidence responses are often those responses that are on the border between two score points. The purpose of this routing is to ensure that unusual or borderline responses receive fair and accurate scores.

```
┌──────────────────────┐
│  Student Responses   │
└──────────────────────┘
            │
            ▼
┌──────────────────────┐      ┌──────────────────────┐
│ Automated Scoring    │◄─────│ Human scorers are    │
│ Engine (ASE)         │      │ used to monitor the  │
│                      │      │ ASE.                 │
└──────────────────────┘      └──────────────────────┘
    │        │        │
    ▼        ▼        ▼
┌─────────┐ ┌─────────┐ ┌─────────┐
│Responses│ │Responses│ │Random   │
│assigned │ │routed to│ │sample of│
│score by │ │human    │ │responses│
│ASE—this │ │scorers  │ │routed to│
│becomes  │ │due to   │ │human    │
│the score│ │condition│ │scorers. │
│of record│ │codes or │ │         │
│         │ │low conf.│ │         │
└─────────┘ └─────────┘ └─────────┘
```

Scoring Process for STAAR Constructed Responses

Any student responses that are routed for human scoring maintain the score assigned by the human scorer as the score of record.

Throughout the scoring window, both ASE and human scoring results are monitored daily to ensure that scores consistently align to the scoring rubrics.

Step 4: Reporting

Results from STAAR administrations are sent to districts after all testing is complete. Reports for each student include the total number of points earned, the scale score, and the performance level. Item-level reports include the average number of points earned for each question on the assessment, including constructed-response questions. For more information about reporting of STAAR results, refer to the Interpreting Results page in the *District and Campus Coordinator Resources*.

Step 5: Rescores

If district personnel or a parent or guardian has concerns about a student's score on a constructed-response question, district testing personnel can request that the student's response be rescored for a fee. Rescore requests are sent for human scoring. The results of the rescore are provided to district testing personnel. If the score on the constructed-response question changes, the fee is waived.

The rescore request must be submitted in the Test Information Distribution Engine (TIDE) by district testing personnel within a specific window. Refer to the Calendar of Events for specific dates. For assistance in submitting rescore requests in TIDE, contact Texas Testing Support at (833) 601-8821 or TexasTestingSupport@cambiumassessment.com.