December 2015 STAAR Constructed Response Scoring: Questions and Answers

Were there problems with the accuracy of scores on constructed response questions in the December 2015 administration?

Although some district staff are concerned that December 2015 scoring of short answer questions was not accurate, all of ETS's and TEA's analysis of the quality of the scoring, using reference data (field test of those items, or previous administrations of item) indicates that the scores generated by the raters meet the same standards for accuracy and reliability as past administrations. There is no reason to assume that December hand-scoring yielded incorrect results.

How do TEA and ETS work to ensure the consistency and fairness of ratings of constructed response questions?

Constructed response (Open-ended) questions on standardized tests are scored differently than those on classroom tests created by teachers. TEA and ETS take significant steps to ensure the consistency and fairness of the scoring. These steps include:

- **Use of clear scoring rubrics.** A scoring rubric is a written description of the types of responses that should be placed into each score category. For the short answer constructed response questions, the scoring rubrics focus on the quality of a student's comprehension of the text, and the student's ability to use text to support his or her argument. For these questions, there are four score categories ranging from 0 to 3. The role of these rubrics is to set a standard and make sure all papers are judged against the same criteria. Rubrics for the items in question are included when test forms are released.
- **Use of benchmark papers.** In addition to the use of rubrics, TEA and ETS provide raters with benchmark papers that provide concrete examples of the types of responses that are to be included in each rating category.
- **Extensive training.** Potential raters must complete online training courses before they are allowed to score actual student responses.
- **Qualification tests for raters.** Raters are not allowed to score until they have proven they can rate a set of responses accurately and reliably.
- **Ongoing monitoring.** Raters never score without a supervisor assigned to them. The supervisor checks the ratings on a sample of their papers and makes sure they are still assigning scores accurately. The supervisors also review rater statistics. Raters whose statistics indicate they are having difficulty get additional training and ongoing monitoring until their performance returns to program accuracy standards. Raters not showing improvement are dismissed from future scoring.

How are the constructed response questions graded?

All constructed responses on STAAR tests are scored on a scale of 0 to 3 by at least two raters. For the short answer constructed response questions, if the first two raters have exact agreement, that is the final score. If the two raters do not have exact agreement on a score, the response is sent to a third rater. If the third rater agrees with one of the first two ratings, that becomes the final score. If the third rater disagrees with the first two raters, the response is sent to a fourth rater for resolution. In the infrequent cases this does not resolve the rating, responses in question are sent to TEA for a final

decision. For reference, on the short answer constructed response questions in December 2015, the raters' initial two ratings agreed 83% of the time for English I and 82% of the time for English II.

If the scores change on a rescore, does that mean the initial scoring was done incorrectly?

In constructed response scoring, raters who are equally well-trained and competent may not agree perfectly on the score to be awarded to a particular response. For example, what one rater sees as a high "2" response, another rater may legitimately see as a low "3" response. Such variations are an integral part of the individual scoring of complex responses, and no amount of training or experience will completely eliminate small differences in the way raters evaluate a test taker's work. Such variability is not a "mistake" in scoring that has to be corrected, but is rather an inherent component of any individualized scoring of complex responses. So a change during a rescore request does not mean an initial score was "wrong," just that sometimes trained raters disagree.

For rescore requests processed to date, the overwhelming majority of initial scores were confirmed. As of April 21, 2016, TEA has received requests to rescore 5,896 short answer constructed responses from 3,260 students. 92.5% were confirmed. Of the remaining 444 responses, 306 went up and 138 went down. This difference is not surprising given that the vast majority of the rescore requests involved zero scores, which could not go down.

Note that a change in a score assigned to one or even more constructed responses does not necessarily mean that a student's passing status will be affected. Of the 3,260 students for whom rescores were requested, 130 resulted in a student moving from fail to pass.

If TEA and ETS have a thorough scoring process, why were there so many more zero scores on the December 2015 test than there were on tests used in December 2014?

There are a number of possible explanations for this difference other than the quality of scoring. The most important point is that different constructed response questions were used on the December 2014 and 2015 tests, and these questions had different patterns of scores. These questions are different and are based on different passages, and thus may be expected to show different difficulty. In addition, test-taking populations vary: the December 2015 test-taking population had a large percentage of third-time test takers who might be expected to score lower on these questions. Based on previous administrations of the questions used in December 2015, the proportion of zero responses appears reasonable.

What it does suggest, however, is that trying to compare the constructed response raw score distributions across administrations may be misleading. Changes in the spread of these scores cannot be taken as evidence that students are performing better or worse, or that the quality of scoring has changed.

If the December 2015 constructed response questions were harder than those used in December 2014, doesn't that make the test unfair?

No. What matters for passing or failing is the scale score on the test, not the scores on individual constructed response questions (part of the raw score which also includes the number of multiple-choice questions answered correctly). Large-scale assessment programs employ a procedure called equating to help ensure that differences in question difficulty do not, on average, disadvantage test

takers on a given form of an assessment. Stated simply, the raw score required to pass a harder test form will be lower than that for an easier test form, but the scale score would be the same.

What should districts look for when deciding when to ask for a rescore? Why shouldn't a district request rescores for all its students?

ETS, like all testing companies, has substantial, well-documented and time-tested procedures to ensure that all aspects of students' tests are accurately scored. The procedures for constructed response scoring, which are briefly described above, are designed to minimize the variability that comes with using human judgment to assign scores.

From a practical perspective, though, the most important thing to consider for each student is whether he or she passed, or could have passed, the test as a whole. The only students who might benefit from the results of a rescore are those who meet the conditions noted in the question above — they are close to the passing score, and an extra point or two may put them over the threshold. It should be noted, though, that most rescores result in no change to the original score. The table below shows the results of rescores for the December 2015 administration as of April 21, 2016. Remember that these were a select group for whom the district had reason to believe the original scores were marginal because of other evidence, like high essay scores or high grades in school.

When deciding whether or not to ask for a rescore, it is important to think about the ways in which scoring is conducted. ETS raters are trained and monitored to make scoring judgments against a specific set of criteria. If an untrained rater disagrees with a rating given to a response, that person should be sure that the disagreement is because of a belief that the rating was an error. In such a circumstance, external knowledge of the individual student should not be brought to bear: even if that person believes that the student "must have meant" something else. Trained raters cannot act upon such a belief. More importantly, a disagreement should be based on the rating and not with the rubric or item.

(continued on next page)

Summary of Rescore Requests for December 2015 STAAR EOC Administration as of April 21th, 2016

			Old Score	Rescore	Count	Sum	%
English I	Single Selection	No Change	0	0	1,297	1,727	94%
			1	1	346		
			2	2	84		
		Rescore is Higher	0	1	57	78	4%
			0	2	11		
			1	2	10		
		Rescore is Lower	1	0	25	34	2%
			2	0	8		
			2	1	1		
		Total Rescore Requests			1,839		
	Paired Selections	No Change	0	0	1,216	1,745	93%
			1	1	322		
			2	2	207		
		Rescore is Higher	0	1	73	99	5%
			0	2	18		
			1	2	8		
		Rescore is Lower	1	0	19	34	2%
			2	0	15		
		Total Rescore Requests			1,878		
	Total English I Rescore Requests (Candidates)				2,052		
English II	Single Selection	No Change	0	0	613	982	91%
			1	1	233		
			2	2	136		
		Rescore is Higher	0	1	48	62	6%
			0	2	7		
			1	2	7		
		Rescore is Lower	1	0	17	31	3%
			2	0	14		
		Total Rescore Requests			1,075		
	Paired Selections	No Change	0	0	514	998	90%
			1	1	205		
			2	2	279		
		Rescore is Higher	0	1	49	67	6%
			0	2	8		
			1	2	10		
		Rescore is Lower	1	0	18	39	4%
			2	0	21		
		Total Rescore Requests			1,104		
	Total English II Rescore Requests (Candidates)				1,208		
Grand Total (Candidates)					3,260		

How were the raters trained for December 2015 vs previous administrations?

The training provided for raters was conducted using the identical training materials and procedures used when the test form was administered in July 2014. The training covered the same topics and processes such as the use of training sets, anchor papers, and annotations. The only difference in the training was in presentation. ETS uses a more interactive model where raters engage with the content through modules rather than in a traditional classroom style. That allows raters to spend more time with the material than they could in a traditional classroom setting. Overall, the duration and depth of the training was comparable for both the previous vendor and ETS.

Why does ETS charge $25 for a rescore?

There are two reasons to charge for a rescores Raters must be paid to rescore, and after the scores are generated they have to be merged into the data files and new reports generated.

Second, the TEA's policy has consistently been to allow districts to request rescores if they truly have reason to suspect the score is in question. The fee helps to ensure that districts are thoughtful about making these requests, rather than asking for rescores for all students. For instance, a district may want to request a rescore for a student who was very close to passing the entire test, has historically high performance in reading or writing, and receives a very low score on the constructed response items. For this student, an increase of one or two points in the constructed response items would mean the student passes the test required for graduation. If the student's score changes as a result of a rescore, there is no charge for the rescore. If the student's score does not change, there is a charge for the rescore.

News media have said that districts had already paid ETS thousands of dollars for rescores. Is this true?

No. To date, ETS has not invoiced any district for any rescores, even those that did not result in a change. Some districts sent in checks with their rescore requests, but ETS returned them with a clarification that they had to wait for invoicing.

What are TEA and ETS doing to address the December2015 rescore concerns?

Commissioner Mike Morath wants to be transparent and open in showing educators the process for scoring constructed responses. As a result, he has chosen to provide every district the opportunity to come to Austin to view their students' responses to the short answer items so districts can see the quality of the process.

In addition, ETS will waive all the rescore fees for requests that were submitted before April 22, 2016 for the December 2015 administration. If districts, after examining the students' responses and scores received want to request additional rescores for the December 2015 administration, the charge will be reduced from $25 to $10 per student whose rescore results in no change. This is a one-time situation, only affecting the December 2015 administration. The March and May 2016 administrations and all future administrations will return to the $25 fee for rescores that do not result in a score change.

Because some of the students are seniors, ETS will prioritize rescore requests for students who need to pass English I and/or English II to graduate in May or June, did not pass in December, and were close enough to the passing requirement that a rescore could potentially benefit them. ETS will report the results of these rescores to districts by May 20th for rescore requests received by May 10th. Also, only one rescore per student on any given open-ended response will be considered.

All other rescore requests for the December administration will be done after the priority rescores noted above. The results of the other rescores will be reported before the July 11th retest.

While it is possible that a few scores may change for some students, the previous table shows that this is unlikely. ETS and TEA are committed to accurately scoring constructed-response questions and to provide Texas educators with the means to see this for themselves.