**Inter-Rater-Reliability**

**Background**

ODE’s technical documentation plan (see page 136 in the 2016-17 Technical Report), included an Inter-Rater Reliability (IRR) study for the 2017-18 school year. Pursuant to Hallgren, K. A. (2012) the assessment of IRR may be necessary to demonstrate consistency among observational ratings provided by multiple assessors. The results of the study will be used to address the requirements within the USED’s Peer Review process (Critical Element 4.1). A sample of Oregon’s Qualified Assessors (QAs) who administer the paper/pencil version of the Oregon Extended Assessment (ORExt) were observed to determine reliability of administration and scoring. We did not include the tablet administration or the Oregon Observational Rating

**Methods**

QTs in districts across the state observed a sample of their respective QAs using the observation protocol and enter their data online. The QA reads the item stem and the student selects from three possible answer choices (A, B, or C) then, the QA records the answer choice. QTs (observer) records the students answer choice, then records the answer choice recorded by the QA for agreement. Only the English Language Arts Writing porting of the ORExt requires any additional analysis by the assessor to determine if the answer meets (1) or doesn’t meet (0). Districts from across the state of Oregon participated in the study, matching the state’s student population demographics, including large, medium, and small districts, across all regions.

The observation protocol was completed for the identified QA, but the student(s) and content area(s) observed were selected by the QT or QA. BRT researchers contacted district-level QTs at the beginning of the test window, which runs from February 15 - April 26, 2018, to arrange observations that could hopefully be completed within one school day.

In addition to addressing inter-rater reliability, the study also evaluated test administration procedures. The methods, results, and interpretation are provided below, in addition to recommended next steps. The observation was composed of three sections:

* **First**, QT’s reviewed ORExt paper/pencil test preparation and administration using the rubric (Appendix 4.1 InterRater\_Observation\_Form). Test preparation/administration domains were rated on a four-point scale from *Inappropriate (I)* to *Exemplary (E)*:
  + ***Inappropriate (I)*** denotes a level of concern that could clearly affect the accuracy of the test results gathered from the test administration. Ratings at this level require substantive retraining of the QA involved.
  + ***Somewhat Appropriate (SA)*** rating denotes a level that includes some minor aspects that could be improved, but the accuracy of the test results are likely not compromised.
  + ***Appropriate (A)*** denotes a level that is consistent with all test administration requirements.
  + ***Exemplary (E)*** level performance suggests that the QA incorporated approaches to test administration that could become models for best practice.
* **Second**, QT’s scored the student alongside the QA using the scoring sheet (Appendix 4.1 InterRater\_Observation\_Form). QT’s compared results after this observation to ensure that the QA entered accurate data.
* **Finally**, QT’s observed the QA completing the data entry process to ensure that no errors are made during data entry and document the number of errors (Appendix 4.1 InterRater\_Observation\_Form).

**Inter-rater Agreement Results**

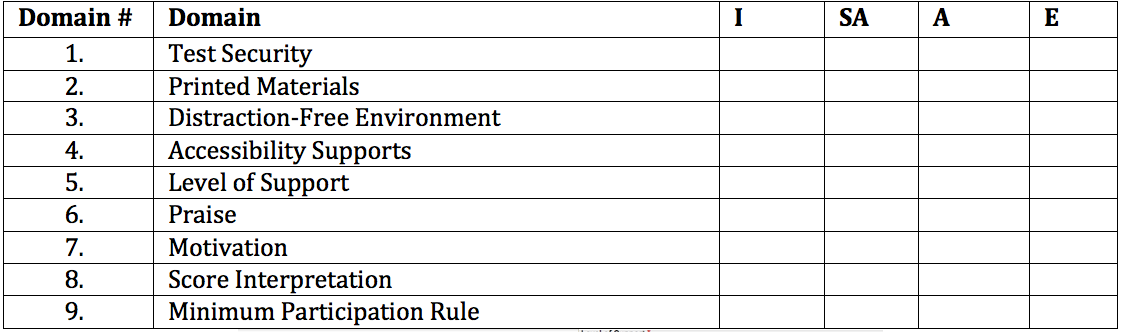
Thirty-three Qualified Trainers from around Oregon participated in the Inter-Rater-Reliability study by doing at least one observation on the Oregon Extended Assessment via paper/pencil administration. Of the thirty-three observations, 48.5% were English Language Arts, 33.3% were Mathematics, and 18.2% were Science. Observations were done at individual student’s typical testing location. The study found a 99.3 Inter-Rater Reliability percentage agreement between the test observers and test administrators on student item (answer) selection.

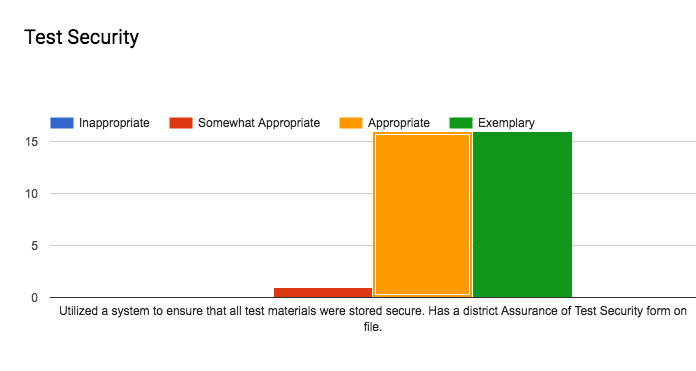
**Qualified Assessor Testing Preparation and Administration Rubric**

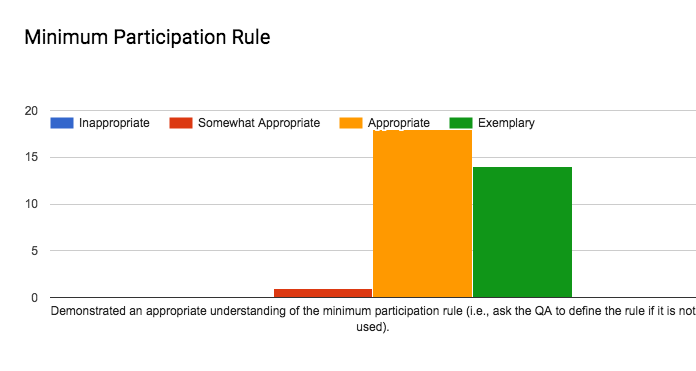
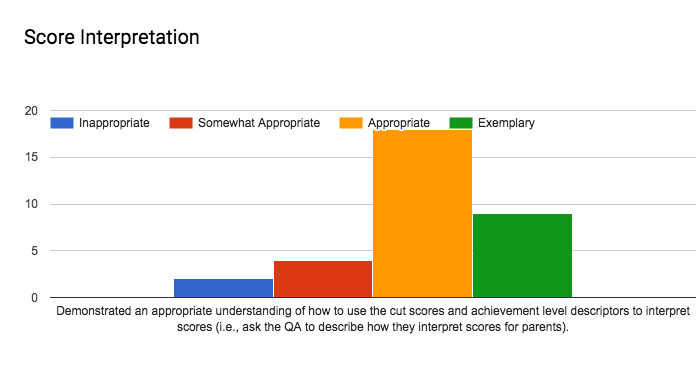
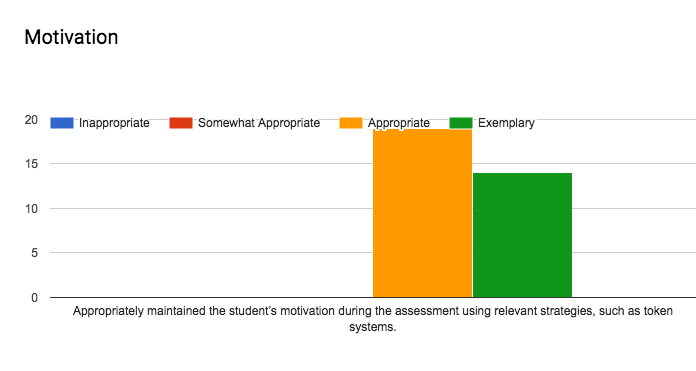
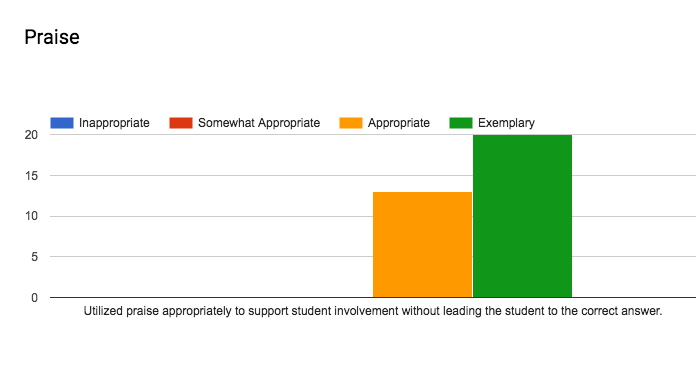
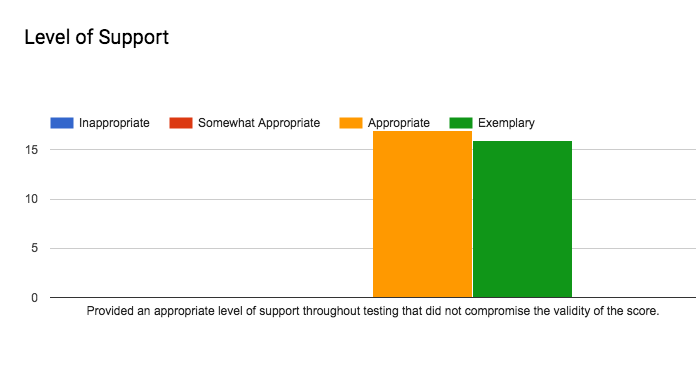
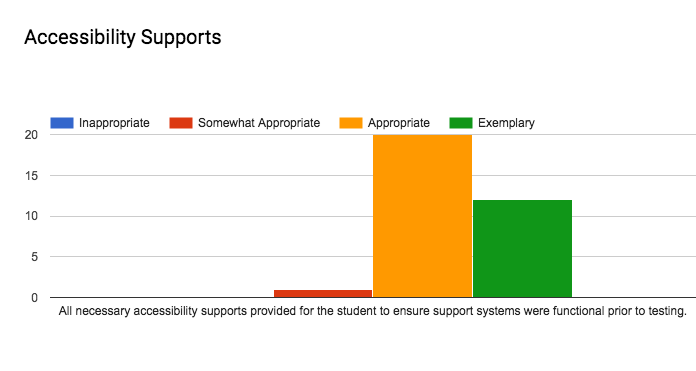
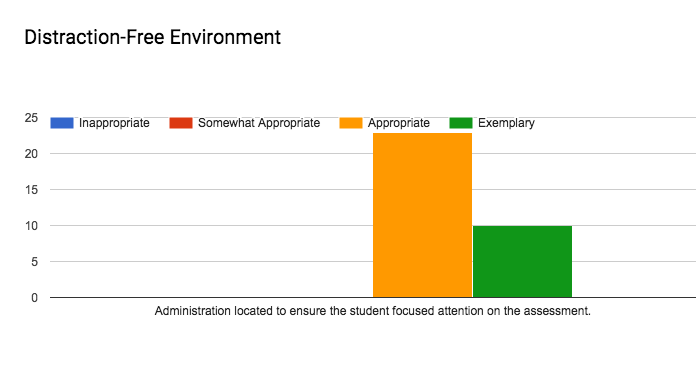
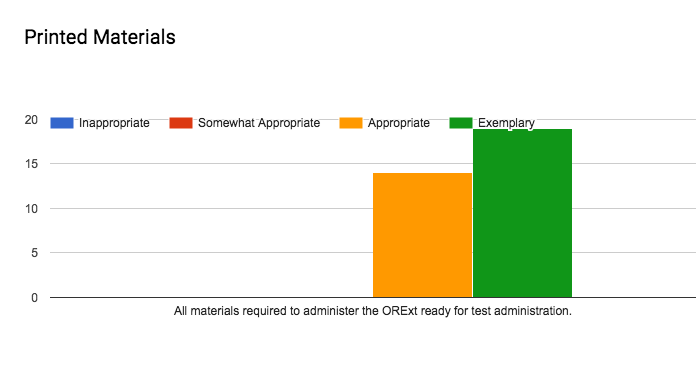
*(Record an “X” in the cell that corresponds to your rating)*

**Domain Definitions**

1. ***Test Security*** – The QA utilized a system to ensure that all test materials were stored in a secure location,. The QA also had a district Assurance of Test Security form on file.
2. ***Printed Materials*** – the QA had all materials required to administer the ORExt ready for test administration
3. ***Distraction-Free Environment*** – the QA arranged to provide the ORExt in a one-on-one test administration in a location that ensured that the student focused attention on the assessment.
4. ***Accessibility Supports*** – the QA provided all necessary accessibility supports for the student and ensured that all support systems were functional prior to testing.
5. ***Level of Support***– The QA provided an appropriate level of support throughout testing that did not compromise the validity of the score.
6. ***Praise***– The QA utilized praise appropriately to support student involvement without leading the student to the correct answer.
7. ***Motivation***– The QA appropriately maintained the student’s motivation during the assessment using relevant strategies, such as token systems.
8. ***Score Interpretation***– The QA demonstrated an appropriate understanding of how to use the cut scores and achievement level descriptors to interpret scores (i.e., ask the QA to describe how they interpret scores for parents).
9. ***Minimum Participation Rule*** - The QA demonstrated an appropriate understanding of the minimum participation rule (i.e., ask the QA to define the rule if it is not used).





****

**Interpretation**

Provide an interpretation of the results above here. Make sure to mention that the score

**Next Steps and Recommendations**

Give ODE action steps to pursue in any areas of concern (writing scoring accuracy/exemplars and score interpretation training)