**Inter-Rater-Reliability**

Background

ODE’s technical documentation plan (see page 136 in the 2016-17 Technical Report), included an inter-rater reliability study for the 2017-18 school year, pursuant to IRR study here (small lit synth). The results of the study will also 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 Assessment (ORora). The study was conducted in two manners:

1. QTs in districts across the state observed a sample of their respective QAs using the observation protocol and enter their data online.
2. Expert reviewers from Behavioral Research & Teaching (BRT) observed district-level QTs and those QAs who give the assessment in more than one school/district.

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:

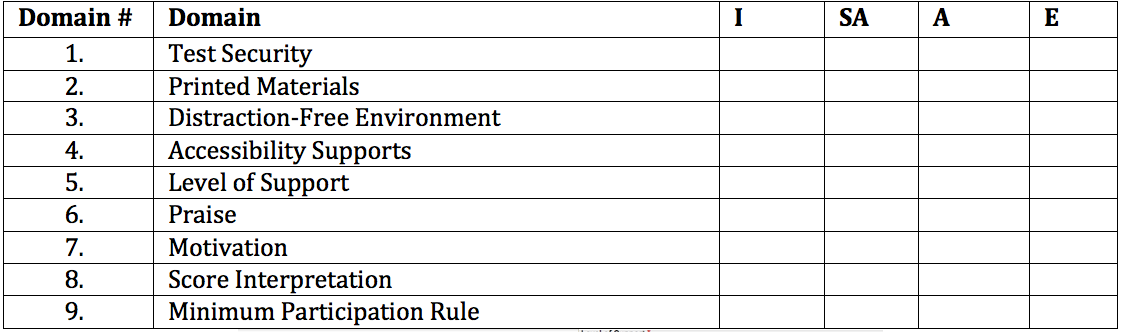
* **First**, QT’s reviewed ORExt paper/pencil test preparation and administration using the rubric (Appendix 4.1 InterRater\_Observation\_Form). Test preparation/administration domains were rated on a four-point scale from *Inappropriate (I)* to *Exemplary (E)*:
  + ***Inappropriate (I)*** denotes a 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).

**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).



Methods

[Describe the methodology employed HERE - Pearson correlation, likely. Make sure to mention and connect the state demographics to the districts that participated (large, medium, small, regional, etc.]

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. Observations were done at individual student’s typical testing location. The study found a 99.3% Inter-Rater Reliability agreement between the test observers and test administrators on student item (answer) selection.

[Disaggregate all of the agreement figures by content area (ELA, Reading, Writing, and Math) HERE. A table should work]

Test Administration Results

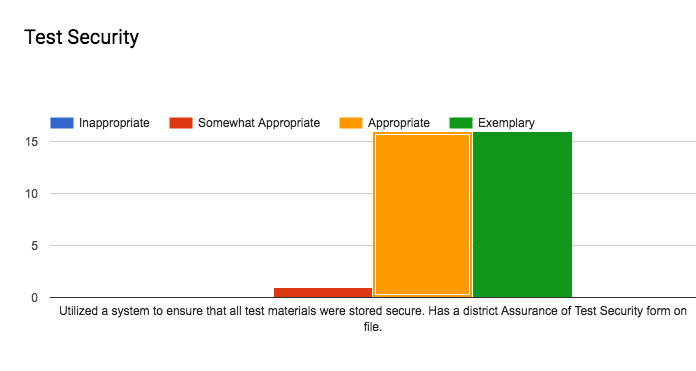
[Introduce the areas observed, etc. HERE]

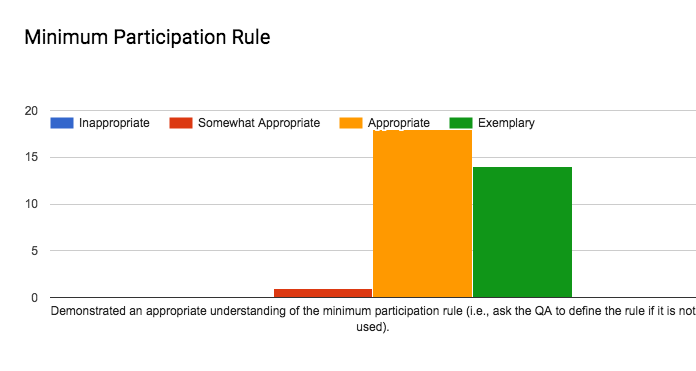
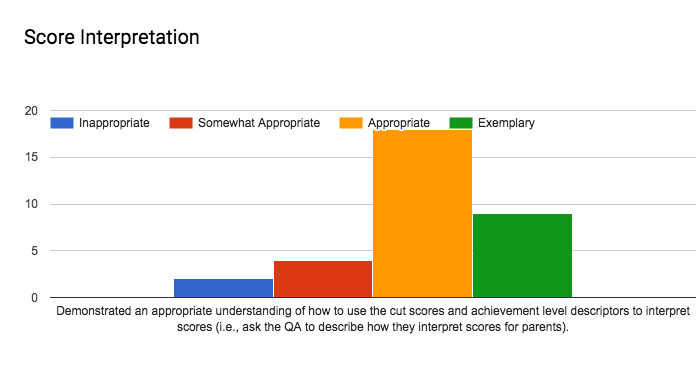
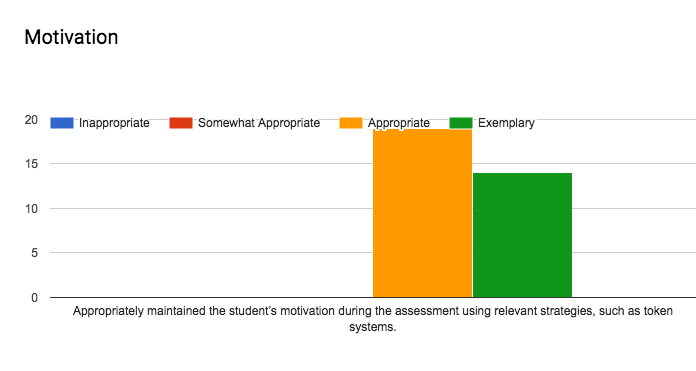
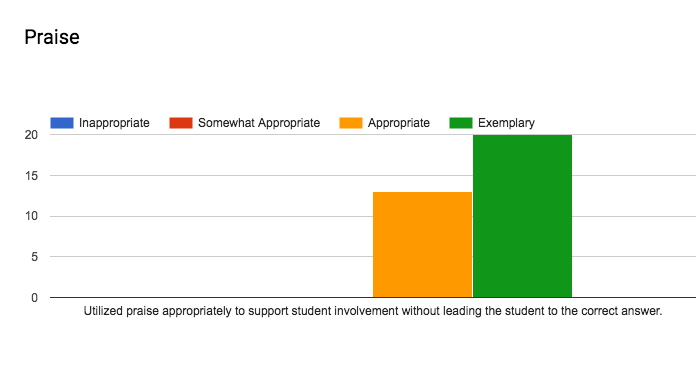
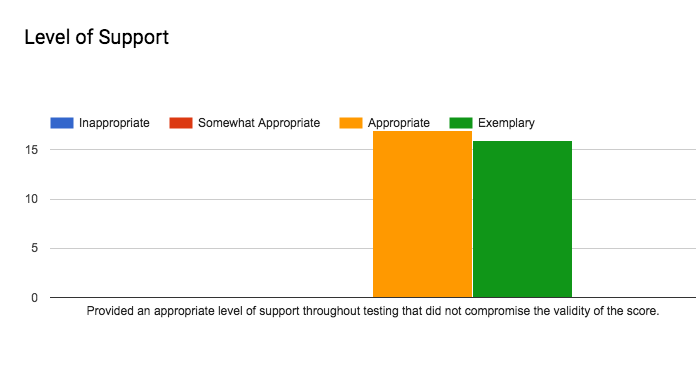
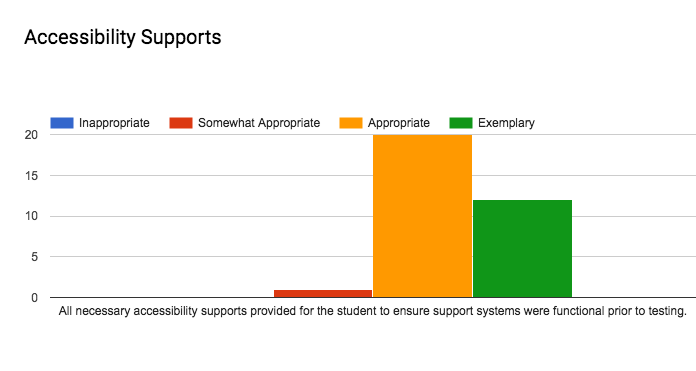
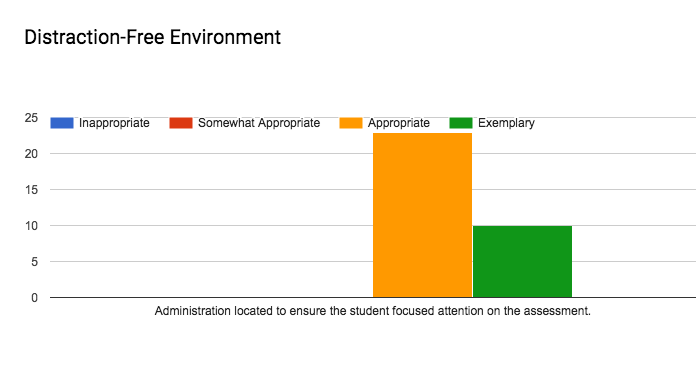
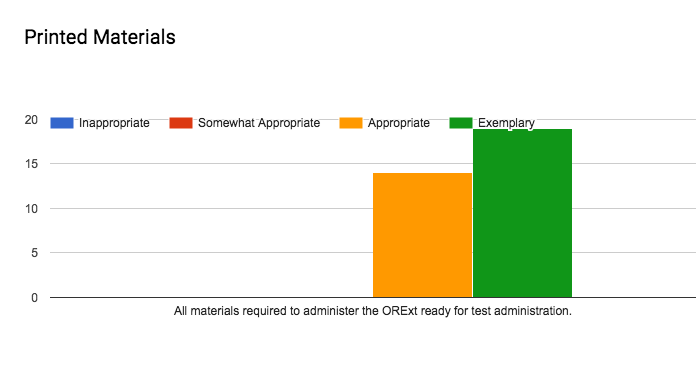
**Percentage of observations by subject area:**

English Language Arts = 48.5%

Mathematics = 33.3%

Science = 18.2%



****

**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)