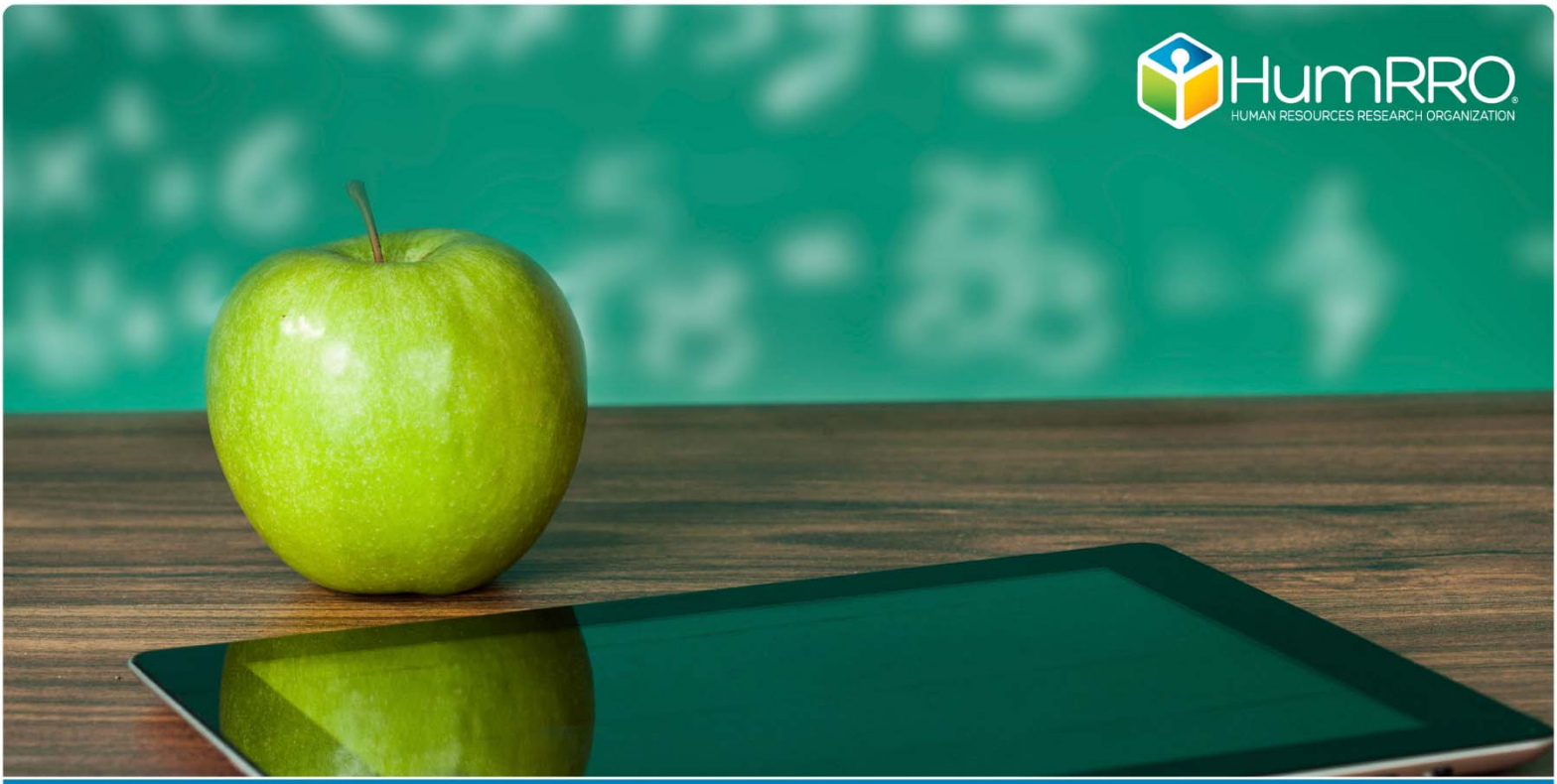

2018 No. 036

South Carolina Assessment Evaluation Report #3

Prepared for:	South Carolina Education Oversight Committee (EOC) 1205 Pendleton Street Room 502 Brown Building Columbia, SC 29201	Prepared under:	Contract Number: 4400014645
Editors:	Andrea L. Sinclair Lucy Chen Emily Dickinson Richard Deatz Arthur Thacker	Date:	June 22, 2018

South Carolina Assessment Evaluation
Report #3

Executive Summary

The South Carolina Education Oversight Committee (EOC) contracted with the Human Resources Research Organization (HumRRO) to conduct a comprehensive evaluation of its state assessments. This is the third and final report detailing the findings from the evaluation. This report serves as the final analysis of the End-of-Course Examination Program (EOCEP) for English 1, for which the text dependent analysis (TDA) item was operational for the first time in 2017-18 on the writing component of the assessment. Report #2 (Sinclair & Thacker, 2018) served as the final analysis of the South Carolina College- and Career-Ready (SC READY) assessments *and* the EOCEPs for Biology 1 and Algebra 1; Report #2 was the most comprehensive of the three reports delivered to the EOC. Report #2 included a partial evaluation of the English 1 assessment. The remaining evaluation tasks to be conducted for the English 1 assessment are reported here in this third and final report, and focus on the writing component of the English 1 assessment.

Overall, the final evaluation of the English 1 assessment indicates that the assessment adheres to industry best practices with some areas noted for improvement. We outline here the areas of strength and offer some recommendations where further improvements can be made. Each recommendation is accompanied by a priority rating. The table below presents the classification schema applied to the recommendations.

Priority Rating Codes for Recommendations

Priority Rating	Description of Priority Rating
Urgent	Definitely needs to be addressed; should be considered and addressed immediately.
High	Needs to be addressed; should be considered and addressed as soon as possible.
Medium	Should be considered and possibly addressed.
Low	Might be considered if time and resources allow.

Review of Test Administration Procedures (Task 4)

We evaluated the extent to which the evidence on test administration complies with 14 standards pertaining to industry best practices for test administration. These standards come from *The Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014; hereafter referred to as the *Test Standards*). Based on evidence collected from the observation of the English 1 writing assessment and interviews with Test Administrators (TAs) and School Test Coordinators (STCs), we increased a rating presented in the Phase 2 report on *Test Standard* 6.1 from “mostly supported” to “fully supported” (ratings were made on a 5-point Likert scale where ‘1’ indicates no evidence to support the standard and ‘5’ indicates evidence fully supports the standard). In sum, one *Test Standard* received a rating of 3, five received a rating of 4, and eight received a rating of 5. The overall mean rating across all the standards increased slightly from M = 4.43 (SD = 0.62) for Report #2 to M = 4.50 (SD = 0.65) for Report #3. This

indicates that, overall, we found the test administration procedures for the English 1 writing assessment adhere to industry best practices.

Areas of Strength

- Test preparation activities were completed as described in the *Test Administration Manual* (TAM). For example, test takers were arranged in seating assignments with ample space between students; materials with testing-related content were removed from the walls (e.g., posters displaying content pertaining grammar usage) prior to testing; student electronic devices were collected prior to the start of testing; and student Test Tickets were handled per instructions in the TAM.
- Test monitoring activities were conducted in accordance with the guidance in the TAM. For example, during testing, the TA and STC circulated around the room regularly to monitor students, and when student questions arose, those questions were addressed promptly and appropriately.
- Across the four schools visited, none of the TAs nor STCs reported concerns about technical issues or problems occurring during testing.
- Findings from the observation and interviews indicate that students navigated through the online assessment with ease and encountered little to no difficulty using the online tools. The TAs and STCs explained that the student tutorial and practice opportunities were useful for ensuring that students understood how to navigate through the online system and use the online tools during testing.
- TAs and STCs reported that they felt that the training and supports they received adequately prepared them to administer the writing assessment.
- There were no indications, either from the observation or interviews, of threats to the security of test material.

Recommendations for Improvement

- More clearly organize the TAM so that all requirements are readily highlighted and known to TAs. ***(Priority Rating: High)***
- Consider streamlining the script read by the TAs during testing to shorten the amount of time that students are listening to instructions. ***(Priority Rating: Medium)***
- Check for consistency across schools in awareness and usage of the writing resources (Writer's Checklist and TDA Scoring Guidelines) available for the writing assessment to ensure widespread use and awareness. ***(Priority Rating: Medium)***

Review of Scaling, Equating, and Scoring Processes (Task 5)

Because a new item type—the TDA item—was included on the 2017-18 English 1 writing component, we updated the Phase 2 evaluation for this task by including a review of the handscoring processes for the TDA item. We conducted a systematic document review, guided by industry standards, to evaluate the handscoring processes for the TDA item. We focused on the handscoring processes, quality control procedures, and rater qualifications for scoring the TDA item.

In Phase 2, we identified 10 *Test Standards* relevant to scaling, equating, and scoring of the EOCEP assessments and rated the extent to which the evidence adheres to those standards. We revisited those ratings in the current phase in light of additional evidence pertinent to handscoring. On the 5-point rating scale (where 1 = No evidence of the Standard found in materials and 5 = Evidence in materials fully covers the Standard), five of the standards received a rating of '5' and five standards received a rating of '4." The overall mean rating for adherence to *Test Standards* increased from *M* = 4.40 (*SD* = 0.49) in Phase 2 to *M* = 4.50 (*SD* = 0.50) in Phase 3. Thus, overall, we found that the scaling, equating, and scoring processes for the EOCEP assessments, and notably the scoring processes for the writing component of English 1, mostly or fully adhere to industry best practices. It should be noted that a limitation of this investigation is that it was based on a review of available documentation; observation of handscoring training and interviews with trainers and scorers would further strengthen the evidence supporting these findings.

Areas of Strength

- The TDA item is scored using a detailed rubric.
- Scorers are trained on examples of each possible score point.
- Scorers are required to pass training and qualifying rounds before being permitted to participate in operational scoring.
- Validity papers are interspersed among operational papers, allowing for detection of scorer drift.
- Inter-rater agreement is monitored daily through the online scoring system, and allows for identification of scorers in need of retraining/recalibration.

Recommendation for Improvement

- To further bolster the validity evidence for the handscoring procedures, DRC should keep documentation of rater agreement reports and Read Behind Logs and make those reports available for audits/independent reviews of their processes and for federal peer review. ***(Priority Rating: Medium)***

Review of Psychometric Processing and Item Parameters (Task 6)

For this task, HumRRO conducted a review of the item parameters for the fall/winter 2017-18 English 1 assessment. Our review of the item-level data indicates that, overall, items are appropriately difficult and discriminate among student ability levels.

Areas of Strength

- The English 1 items[1], and notably the TDA item, demonstrated appropriate levels of classical test theory (CTT) item difficulty and discrimination.
- Aside from one multiple-choice item, the English 1 items, and notably the TDA item, did not demonstrate differential item functioning (DIF) among gender and ethnic groups.

[1] This task included analysis of the items on the reading component of the English 1 assessment.

Recommendation for Improvement

- Rasch item statistics indicate that the TDA item is difficult for the students tested, although its difficulty parameter was only slightly above the upper bound of the acceptability range (and within range for the CTT item difficulty criterion albeit near the upper range). The higher item difficulty of the TDA item compared to other English 1 items might be due to the students' lack of familiarity with this new item type. We recommend continued monitoring of the psychometric properties of the TDA item during subsequent test administrations to ensure that the TDA item does not continue to exceed item difficulty criteria. The continued psychometric monitoring should include monitoring of Item Response Theory (IRT) step parameters in order to provide additional insight (i.e., beyond CTT item difficulty and Rasch item difficulty statistics) on the difficulty of scoring a '1,' '2,' '3,' or '4' on this polytomously-scored TDA item. ***(Priority Rating: High)***

- We recommend that DRC include additional documentation of how CTT item difficulty and item discrimination statistics are computed for the polytomously-scored TDA item. ***(Priority Rating: High)***

Conclusion

Overall, the findings indicate that the English 1 assessment adheres to sound testing practices as described in *The Standards for Educational and Psychological Testing*, and thereby supports the validity of the test scores for their intended uses and purposes. No critical concerns were identified from the evaluation of the English 1 assessment. The findings from this report should be considered in conjunction with the EOCEP results presented in Report #2 (Sinclair & Thacker, 2018).

We applaud South Carolina for securing an external evaluation of its assessments to help ensure their quality. Periodic evaluations of testing practices will help to ensure their continued technical soundness.

South Carolina Assessment Evaluation Report #3

Table of Contents

Table of Contents (Continued)

List of Tables

List of Figures

South Carolina Assessment Evaluation Report #3

Chapter 1: Introduction

The South Carolina Education Oversight Committee (EOC) contracted with the Human Resources Research Organization (HumRRO) to conduct a comprehensive evaluation of its state assessments. This is the third and final report summarizing that effort.

The EOC provides oversight of programs and expenditure of funds for the Education Accountability Act and the Education Improvement Act of 1984. As established in Section 59-6-10 of the South Carolina Code of Laws, the EOC's responsibilities include reviewing all assessments for approval as components of the state accountability system. As part of this process, assessments are evaluated for validity, including alignment with the state standards, level of difficulty, and the ability to differentiate levels of achievement. Based on the evaluation, recommendations for improvements and changes are made. The EOC shares the information and recommendations with the State Board of Education, the South Carolina Department of Education (SCDE), the Governor, the Senate Education Committee, and the House Education and Public Works Committee. The SCDE will then report to the EOC how it will address the recommendations and the EOC will decide whether to approve the assessments for accountability purposes. HumRRO's comprehensive evaluation is intended to support the EOC in meeting these legislative mandates.

The state assessment program includes the South Carolina College- and Career-Ready (SC READY) assessments and the End-of-Course Examination Program (EOCEP) for high school. Data Recognition Corporation (DRC) works in coordination with SCDE to develop, administer, and score the tests.

To meet federal accountability requirements, the SC READY is administered annually to all public school students in grades 3–8 in the content areas of English Language Arts (ELA) and math. The EOCEP is administered in ELA, math, science, and social studies to all public school students by the third year of high school. HumRRO's evaluation includes the SC READY for ELA and math at all tested grade levels, as well as the EOCEP assessments for English 1, Biology 1, and Algebra 1.

HumRRO's approach to evaluating South Carolina's assessment system included a series of separate but related tasks that focus on the key elements of assessment design and implementation. Specifically, HumRRO identified the following seven tasks that address the general requirements listed in Section III (a-f) (pgs. 15-17) in the Request for Proposals (RFP):

- Task 1: Review Item Development Processes
- Task 2: Review Items to Standards Alignment and Item Quality
- Task 3: Review Test Construction Processes
- Task 4: Review Test Administration Procedures
- Task 5: Review Scaling, Equating, and Scoring Processes
- Task 6: Review Psychometric Processing and Item Parameters
- Task 7: Review Minimum Legal Requirements of SC READY

To accomplish Tasks 1 - 7, HumRRO coordinated with DRC and SCDE to obtain the necessary documentation and data. HumRRO's primary communication was with the Project Manager at DRC, who in turn coordinated with SCDE, as needed, to address HumRRO's data requests and questions.

The seven tasks were completed over three phases (see Table 1.1). The current report is the third and final report (Phase 3), and serves as the final analysis of the EOCEP English 1 assessment. Report #1 (Phase 1) included an initial analysis of the SC READY assessments and the EOCEP Algebra 1 assessment (Dickinson, Chen, & Swain, 2017). Report #2 (Phase 2) served as the final analysis of the SC READY assessments and the EOCEP assessments for Biology 1 and Algebra 1, and a partial analysis of the EOCEP assessment for English 1 (Sinclair & Thacker, 2018). Report #2 was the most comprehensive of the three reports. Because the English 1 assessment included a new item type in 2017-18, the final analysis of the English 1 assessment was reserved for the third and final report (Phase 3).

Table 1.1. Tasks and Assessments Included in each HumRRO Report/Phase

Tasks	Report/Phase			
	SC READY	EOCEP English 1	EOCEP Biology 1	EOCEP Algebra 1
1. Review Item Development Processes	1, 2	2	2	1, 2
2. Review Item to Standards Alignment & Item Quality	2	2	2	1
3. Review Test Construction Processes	1, 2	2	2	1, 2
4. Review Test Administration Procedures	2	2, 3	2	2
5. Review Scaling, Equating, and Scoring Processes	2	2, 3	2	2
6. Review Psychometric Processing & Item Parameters	2	2, 3	2	2
7. Review Minimum Legal Requirements	2	--	--	--

The new item type included on the 2017-18 English 1 assessment is a "text-dependent analysis" (or TDA) item. This type of item requires students to read a text or passage and draw upon that text to support their written response with evidence from the text. The type of text that students read and respond to for the TDA item may be drawn from different genres (e.g., historical fiction, science fiction, non-fiction biography) or modes (e.g., narrative, expository, persuasive), but the type of writing that the students produce is not mode-specific. The TDA item is scored with a holistic rubric that has a point range of 1 (lowest) to 4 (highest). To reflect the importance of student writing, the score on the TDA item is weighted by a factor of 4 for a maximum of 16 points.

This final report includes (a) an observation of test administration, including interviews with Test Administrators and School Test Coordinators (Task 4), with a focus on the TDA item, (b) a review of handscoring procedures and score reporting for the TDA item (Task 5), and (c) a review of the psychometrics (item parameters) for the English 1 assessment (Task 6).

The remaining chapters of this report describe the evaluation method and present results and related discussion for Tasks 4 – 6 for the EOCEP English 1 assessment. The final chapter provides the conclusions for the evaluation of the EOCEP English 1 assessment.

Chapter 2: Review Test Administration Procedures (Task 4)

Introduction

The purpose of this task was to evaluate the extent to which the test administration procedures follow best practices as described in *The Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014; hereafter referred to as the *Test Standards*). The *Test Standards* provide guidelines for assessing the validity of interpretations of test scores for the intended test uses. In this chapter, we first introduce the methods we used to evaluate the test administration processes for the EOCEP English 1 assessment. Then, we describe the results. Finally, we discuss our findings and provide recommendations for improving test administration for the EOCEP English 1 assessment.

Method

For Report #2 (Sinclair & Thacker, 2018), we conducted a systematic document review to evaluate the test administration processes for the South Carolina assessments (SC READY and EOCEP assessments). We worked in cooperation with the South Carolina Education Oversight Committee (EOC), the South Carolina Department of Education (SCDE), and the Data Recognition Corporation (DRC), with primary support provided by DRC, to obtain documentation of the South Carolina test administration processes for each assessment. We also searched the SCDE website to identify additional relevant information. The documents we collected included materials such as the *Test Administration Manual* (TAM), training materials for Test Administrators, test accommodation guidelines, and test security procedures. We identified 14 standards from the *Test Standards* that are directly relevant to test administration. We then evaluated the degree to which the documents we reviewed indicated compliance with each standard.

With any review of test administration procedures, fidelity of administration and adherence to protocols is vital to reduce the impact of factors other than student achievement (i.e., construct-irrelevant variance) on a student's test performance. Consequently, for this current report (Report #3), we expanded upon the document review included in Report #2 by conducting site visits in South Carolina to observe administration of the EOCEP English 1 assessment[2] and to conduct interviews with Test Administrators (TAs) and School Test Coordinators (STCs) to further inform fidelity of administration and adherence to protocols. As such, the focus of the current report is on findings from the observation of test administration and interviews.

Participants and Procedure

HumRRO requested access to a small number of South Carolina high schools to observe the administration of the writing component of the English 1 assessment. We received support from a staff person of the Education Oversight Committee, a retired South Carolina educator, who agreed to recruit high schools on our behalf. As a result, HumRRO was able to gain the cooperation of four high schools. Because of limited testing dates (only one per week) and one administration session per school, we observed only one administration of the writing component of the English 1 assessment; however, we conducted interviews with TAs and STCs who recently administered the writing component at four high schools, including the observation

[2] The focus of this evaluation was on the English 1 writing component given that the writing component contained a new item type (i.e., the text-dependent analysis item). Thus, we observed the English 1 writing component of the assessment, but not the English 1 reading component, which was administered during a different week than the week for which the site observations were scheduled.

site. The STC and TA interview at each school was conducted in a group interview format, primarily to reduce school staff time away from their classroom. Table 2.1 displays the break out of test observation and interviews by school.

Table 2.1. Test Observation and Interviews by School

High School	Writing Test Observation (n)	STC Interviews (n)	TA Interviews (n)
A	1	1	1[a]
B	0	1	1
C	0	1	1
D	0	1	4

[a]TA was testing during STC interview; however, some feedback was obtained.

The HumRRO research scientist who conducted the school visits has been with HumRRO for 21 years, He has extensive experience collecting data through individual and group interviews and classroom observation for numerous state education clients such as Kentucky, Florida, and California in addition to other HumRRO clients such as the U.S. Army, Excelsior College, and the National Assessment for Educational Progress (NAEP). He directed and designed test observation studies using measures similar to those used in this effort for the Minnesota Department of Education and the Partnership for Assessment of Readiness for College and Careers (PARCC).

Measures

HumRRO developed data collection instruments that included an observation checklist and interview protocol. These instruments were based on previous work HumRRO has conducted for the Minnesota Department of Education school visit studies in 2017 and 2018 and historically from NAEP test administration observations (Schulz, Wiley, Buckland, Michaels, Diaz & Chen, 2017) and PARCC test administration observations (Sinclair, Deatz, & Johnston-Fisher, 2015). The observation checklist and interview protocol were developed with information from the EOCEP Spring 2018 TAM in addition to related documents, such as the 2017-18 English 1 EOCEP Blueprint, Text Dependent Analysis Scoring Guidelines, and the ELA Writers Checklist. The HumRRO Project Director conducted a quality review of both documents, providing feedback that was incorporated into the final instruments.

Observation Checklist. The observation checklist was developed first by identifying observable tasks that TAs are expected to complete when administering the EOCEPs to students and grouped under two headings: Test Observation – Prior to Testing and Test Observation - During Test Administration. A description of each task was provided, space for notes, and a checkbox if the task was successfully completed. Each task number references the related page(s) in the TAM where the task is described. Where appropriate, additional guidance was provided to the observer to target the observation on specific aspects of the tasks. This guidance was based on the information found in the source documents identified above. See Figure 2.1 for an excerpt from the checklist (see Appendix A for the complete checklist).

Interview Protocol. The primary purpose of the interview was to obtain information from those directly involved in administering the writing component to provide a school/educators'

perspective of the test administration standards from Report #2[3]. The interview questions focused on the helpfulness of the materials received to a) prepare the school to administer the test, b) train TAs and test monitors, c) prepare students for testing, d) maintain test material security, and e) conduct test administration procedures the day of testing (e.g., seating charts, script, test system access). Although the questions were written specifically for the STC, the interviews were conducted with at least one TA present to provide information from the perspective of having recently administered the writing component. The group interview is a useful format for this type of data collection because responses from one person often trigger additional input from others. Also, it minimizes the burden on school staff being away from their classrooms. See Appendix B for the complete interview questions.

Test Observations – During Test cont.			
Task #	Task	Notes	OK?
24 *p18-19, 23, 30, 59, 60*	Once testing has begun, TA and monitors actively monitor students (i.e., prohibited behaviors, item security, or cheating) by walking around the room, do not do other work or have conversations.		
25 *p59*	TA or monitors maintain order so one or more students do not distract others. (is it a good testing environment?)		
26 *Chk list*	Do the students use their scratch paper? Are they organizing ideas, thought maps? How many use? To what extent?		
27	Do students ask questions during the exam? What questions are asked? How does the TA respond?		
28	Do students access the Writer's Checklist		

Figure 2.1. Observer Checklist Example.

Results

First, we present the results from the observation of test administration. Then, we present the themes from the interviews with TAs and STCs. Finally, we present the evaluation of how well the evidence pertaining to the English 1 test administration adheres to the relevant *Test Standards,* which serves as an update to evaluation ratings included in Report #2 (Sinclair & Thacker, 2018).

Observation of Test Administration

HumRRO observed the administration of the English 1 writing component in the high school's library. The students were already in the testing area when the TA and the STC arrived. The TA immediately began seating the 17 students using her seating chart and reading the script in the TAM to get students logged into the test system, which took approximately 20 minutes. Other adults in the testing room included (a) a test monitor, who remained until the students started the exam, (b) the STC who remained in the testing room, (c) the school's principal, who was

[3] During the interview, we intended to also capture user perspectives on the English 1 reading component; however, none of the four schools had administered the reading component of the English 1 assessment at the time of the site visit.

there briefly, and (d) the District Technology Coordinator who was in the testing room for about half of the testing time. Students were released as they completed the assessment. The overall findings from the test administration observation are in Table 2.2.

Table 2.2. Observation of Test Administration Findings

Finding	Evidence
1. Test administration procedures prior to the exam starting were completed as designed.	The TA used a seating chart to place students at library tables with two students facing each other per table. This provided enough room to work and ample space between students. No test content (e.g., posters) was visible in the testing room. The TA began to read the script and students were first asked to place all electronic devices, including smart watches, in one location. The TA received the Test Tickets from the STC when directed in the script, distributed them, and collected them once students logged on the test system. The script was read with clarity and took about 15 minutes to complete.
2. Test administration procedures for monitoring the test were completed as designed.	The TA and STC moved around the room often to monitor the students. The TA remained in the testing room throughout the entire test.
3. Student questions were handled appropriately.	Three questions were asked by students. One question was asked while the TA was reading instructions, and two questions were asked during testing. The questions were: 1. How will I know if I logged out properly? 2. My screen went black, but came right back up, is it OK? 3. How do I reference the text in the passage? (students at this school were taught to list the sentence number and the passage wasn't numbered) All questions were addressed appropriately by the TA. The observer noted (and later confirmed by the TA) that no questions involved test system tool usage or navigation.
4. Test administration procedures for resolving technical issues were handled appropriately.	One technical issue occurred and was quickly resolved. A student paused the computer and when she returned she was unable to type any text. The TA called the District Technology Coordinator (who had stepped out of the testing room for a short while) who responded immediately. There was a question about shutting the computer down and losing data, so the Tech called someone (presumably a Help Desk) and was advised to shut down the computer. He did and the student was able to resume testing, with no loss of data, within 10 minutes. This issue was handled effectively and efficiently and the TA and others interviewed at the school indicated no concerns regarding technical issues or problems.
5. Test administration procedures for test security were completed as designed.	The Test Tickets are secure materials and the STC retained control of the documents after printing them that morning. He did not give them to the TA until the script indicated to do so. Both the TA and STC closely monitored students so there was no opportunity to provide or receive assistance by other students. The observer noted that there were no sidebar student conversations during testing.

Table 2.2. (Continued)

Finding	Evidence
6. The *Test Administration Manual* (TAM) may need modification.	Although the script seemed to flow reasonably well, the observer noted that some information in the script was repeated. The script could be streamlined by removing redundancies. Also, the procedure to collect Test Tickets after the students logged in did not appear to be an efficient process because the tickets are needed for students to log back in when/if a technology issue arises; this process could be clarified, and perhaps reorganized, in the TAM.
7. Most students used scratch paper.	The HumRRO observer found that 10 of 17 students used the scratch paper provided for organizing ideas for the text-dependent analysis (TDA) item. Seven students used it extensively, creating bullets with supporting text.

Interviews

HumRRO conducted group interviews (with STCs and TAs) at three additional high schools in addition to the STC interview at the first school where testing was observed. The interviews ranged between 30 and 50 minutes, depending upon the availability of school staff.

Themes that emerged from the interviews are found in Table 2.3.

Table 2.3. Interview Themes

Theme	Evidence
1. The TAM is thorough, but could benefit from some reorganization.	All staff interviewed indicated the TAM provided the information needed to prepare for and administer the English 1 writing component. However, all those interviewed agreed that it was difficult to find specific information. They recommended reorganizing key information by role or timeline (e.g., testing preparation, test administration, testing close-out) to facilitate finding information. STCs and TAs at two schools stated there should be a sequential 'day of testing checklist' of tasks for TAs complete.
2. The test administration script could benefit from streamlining.	TAs (six of seven) stated that the script was too long and repetitive. Three TAs and the observer noted that there was about 20 minutes between students logging on to the assessment and students starting the assessment. Suggestions for improvement included: having a script for each subject and highlighting the text that could be omitted if students have already taken a test in another subject.
3. The test system training prepared students for testing.	All STCs and TAs stated emphatically that the student test system tutorial and practice opportunities were helpful for preparing students for using online tools and navigating through the online assessment; none of the TAs could recall a student asking a system tool or navigational question during the assessment. TAs at one school expressed disappointment that there was not a practice TDA question to help prepare students for this type of item.
4. Supplemental testing resources are familiar to most students, but may be some variability across schools.	During the test observation (School A), it was noted that one student accessed and referred to the Writer's Checklist document. During the interview at School C, one TA stated that nearly half the students referred to the TDA Scoring Guidelines while testing. TAs at schools A, B, and C indicated that students knew and used both documents regularly in the classroom. However, the TAs at School D indicated that they were not familiar with either document.

Table 2.3. (Continued)

Theme	Evidence
5. Students should be tested as they learn.	The STC from School A indicated that students rarely write high-quality essays in one sitting; therefore, consideration should be given to allow students to complete a draft of the TDA item one day, and then revise and finalize it the next day. TAs at School D said the reading passage associated with the TDA items are too lengthy and difficult for their low performing students, particularly without access to a dictionary or audio pronunciation of unfamiliar words. TAs at Schools B and C indicated some students worked on their essays for more than four hours.
6. STCs maintained test material security and were familiar with the Test Security Violation form.	The STCs interviewed at Schools B, C, and D reported that they do not print the Test Tickets until the day of testing. Then they are handed to the TAs in the testing room. One STC (School C) said he printed them on large paper (scratch paper) and would not even hand them to the TA until the students were seated and ready to access the computers. All STCs were familiar with the Test Security Violation form and no one completed one this year.

Ratings on the Test Standards

As noted above, Report #2 included a review of the documentation related to test administration. Based on that review, two education researchers independently rated the extent to which the documentation adheres to the relevant test administration standards from the *Test Standards.* Any discrepancies in ratings were discussed until the researchers reached a consensus rating. For the current report, those ratings were re-visited by the researcher who conducted the site visits and by the project director to determine whether any of the new information gathered from the observation and interviews warranted a change in the rating assigned to each *Test Standard* for the English 1 writing assessment component. The rating scale is presented in Table 2.4.

Table 2.4. Rating Scale for Evaluating Strength of Evidence for Test Standards

Rating Level	Description
1	No evidence of the Standard found in the materials.[a] [b]
2	Little evidence of the Standard found in the materials; less than half of the Standard covered in the materials and/or evidence of key aspects of the Standard could not be found.
3	Some evidence of the Standard found in the materials; approximately half of the Standard covered in the materials, including some key aspects of the Standard.
4	Evidence in the materials mostly covers the Standard; more than half of the Standard covered in the materials, including key aspects of the Standard.
5	Evidence in the materials fully covers all aspects of the Standard.

[a] Materials include all documents and data provided, any emails or phone calls with SCDE/DRC staff, as well as information available online.
[b] For the purposes of this report, "materials" also includes evidence collected via observation of test administration and interviews with TAs and STCs.

Table 2.5 displays the rating for each relevant *Test Standard*. Ratings that changed because of the information gained from the observation and interviews are highlighted in bold and italicized text. The non-bolded, non-italicized text represent ratings that remain unchanged from the findings presented in Report #2 for the EOCEP assessments.

Table 2.5. Evaluation Results for Test Administration Procedures Based on the Test Standards

Standard Number	Standard Content	Rating
Standard 3.10	When test accommodations are permitted, test developers and/or test users are responsible for documenting standard provisions for using the accommodation and for monitoring the appropriate implementation of the accommodation.	4
Standard 4.5	If the test developer indicates that the conditions of administration are permitted to vary from one test taker or group to another, permissible variation in conditions for administration should be identified. A rationale for permitting the different conditions and any requirements for permitting the different conditions should be documented.	5
Standard 4.15	The directions for test administration should be presented with sufficient clarity so that it is possible for others to replicate the administration conditions under which the data on reliability, validity, and (where appropriate) norms were obtained. Allowable variations in administration procedures should be clearly described. The process for reviewing requests for additional testing variations should also be documented.	5
Standard 4.16	The instructions presented to test takers should contain sufficient detail so that test takers can respond to a task in the manner that the test developer intended. When appropriate, sample materials, practice or sample questions, criteria for scoring, and a representative item identified with each item format or major area in the test's classification or domain should be provided to the test takers prior to the administration of the test.	4
Standard 6.1	***Test administrators should follow carefully the standardized procedures for administration and scoring specified by the test developer and any instructions from the test user.***	***5***
Standard 6.2	When formal procedures have been established for requesting and receiving accommodations, test takers should be informed of these procedures in advance of testing.	4
Standard 6.3	Changes or disruptions to standardized test administration procedures or scoring should be documented and reported to test users.	3
Standard 6.4	The testing environment should furnish reasonable comfort with minimal distractions to avoid construct-irrelevant variance.	5
Standard 6.5	Test takers should be provided appropriate instructions, practice, and other support necessary to reduce construct-irrelevant variance.	4
Standard 6.6	Reasonable efforts should be made to ensure the integrity of test scores by eliminating opportunities for test takers to attain scores by fraudulent or deceptive means.	5
Standard 6.7	Test users have the responsibility of protecting the security of test materials at all times.	5
Standard 7.7	Test documents should specify user qualifications that are required to administer and score a test, as well as the user qualifications needed to interpret the test scores accurately.	5
Standard 7.8	Test documents should include detailed instructions on how a test is to be administered and scored.	4
Standard 7.9	If test security is critical to the interpretation of test scores, the documentation should explain the steps necessary to protect test materials and to prevent inappropriate exchange of information during the test administration session.	5

Rationale for Ratings on Test Administration Test Standards[4]

For the one *Test Standard* for which the rating increased from Report #2 to Report #3 (Standard 6.1) the rationale for that increase is provided below. Also, if additional relevant information was collected through the test observation and/or interviews, we provide that information for those *Test Standards* as well. The additional evidence collected on the English 1 assessment is presented in bold text for those *Test Standards*. The original text from Report #2 is in non-bolded text.

If there was no additional pertinent information for a given *Test Standard* that was gathered through the observation and interviews, then the rationale for the rating remains the same as provided in Report #2 and is not repeated here.

Standard 4.15 – The directions for test administration should be presented with sufficient clarity so that it is possible for others to replicate the administration conditions under which the data on reliability, validity, and (where appropriate) norms were obtained. Allowable variations in administration procedures should be clearly described. The process for reviewing requests for additional testing variations should also be documented.

The TAMs and Administrative Directions Manuals (ADMs) for online and paper/pencil testing of EOCEP and SC READY provided sufficient clarity and details. For example, the manuals included directions for both STCs and TAs, directions for administering both online and paper-and-pencil testing. In addition, the TAMs and ADMs included general test administration directions for all subjects and specific test administration directions for specific subjects for both online administration and paper/pencil administration. The verbal script in the ADMs provide enough details and clarity so that others can easily replicate the administration conditions and thereby support the reliability and validity of the assessments.

The HumRRO observer noted during the English 1 writing administration that the script flowed reasonably well; however, the time between logging on to the system and starting the test was about 15 minutes. TAs interviewed at three of the schools also stated the script is too long and repetitive. One TA reported that it took 20 minutes before students started the assessment and another TA said that while reading the scripts a couple students indicated to her that they "got it" and wanted to get going on the test.

The TAMs also describe allowable variations in administration procedures. For instance, in the TAM for SC READY, the test developers list three acceptable alternatives for ensuring that students placed in Residential Treatment Facility (RTF) are appropriately assessed (see details on p.24). The process for reviewing requests for additional testing variations is also documented. For instance, in the TAM for SC READY, it is mentioned that "testing must be conducted during the published schedule for the specific test or District Test Coordinators (DTCs) must provide the SCDE with a written request for an alternative schedule" (p.25).

Although there is sufficient documentation to replicate administration conditions across various settings, the organization of the TAMs could be improved. The overall structure flows; however, the SCDE Policies section has information regarding all phases of the test administration process and may be confusing as a Test Administrator or School Test Coordinator reads about processes that have not yet been discussed in the TAMs. For example, SC READY TAM (p. 36)

[4] For this task, we do not address elements of these standards that do not directly pertain to test administration (e.g., detecting cheating through scoring analyses).

details the timing and break procedures during administration; however, page 65 of the Test Administrator's Section only indicates that breaks should be scheduled as needed, with no reference to the details on page 36.

Another example of possible TA confusion surfaced during the school visits. The EOCEP Spring 2018 TAM policy section (page 31) and the script (pages 73 and 78) instructs the TA to pass out the Test Tickets and to tell students they will collect them once the testing begins. However, on page 79, guidance is that the TA should ensure the test tickets have been collected if your school also uses them as scratch paper. The HumRRO observer noted during the administration of the English 1 writing component that Test Tickets were collected after students logged on the test system; however, when a student's computer locked up after being paused, the TA had to retrieve the student's Test Ticket to be able to log back in after the test was restarted.

Organizing all the necessary requirements in one section would minimize the need to reference multiple sections of the document, reducing the potential to miss policies and procedures pertinent to standardization, which is particularly concerning when sections do not prompt the STC or TA to review specific sections. The current SCDE Policies section could be included as an Appendix to highlight the specific Department of Education Policies in one document. Additionally, the TAMs indicate what TAs and Monitors are permitted to answer, but do not indicate in the ADM script a specific verbal response. Including scripted responses to frequently asked questions, particularly those that TAs and Monitors are not permitted to answer could improve standardization across administrations.

Information provided during the school visit interviews echoed similar ideas of information consolidation and re-organization. Although all STCs and TAs interviewed felt the TAM was comprehensive, everyone acknowledged that some information was difficult to find. One STC stated that "…everything was in one manual, it is scattered, but all there." Six TAs requested that some type of quick reference guide be provided that contains everything they need to do, sequentially, the day of testing. They suggested a sequential bulleted list instead of long sentences.

Standard 4.16 – The instructions presented to test takers should contain sufficient detail so that test takers can respond to a task in the manner that the test developer intended. When appropriate, sample materials, practice or sample questions, criteria for scoring, and a representative item identified with each item format or major area in the test's classification or domain should be provided to the test takers prior to the administration of the test.

Evidence from the documents indicates that key aspects of Standard 4.16 are covered. The Online Tools Training (OTT) and tutorials are available to students for both EOCEP and SC READY (see files *Spring 2017 SC READY Brochure.pdf* and *Tutorials and Online Tools Training.docx*). Sufficient details are provided to test takers so that they can respond to a task in the manner that the test developer intended. There are video tutorials that provide clear instructions about how to sign in and how to use basic and advance tools of the online testing system.

Responses from all STCs and TAs during the school visits can be characterized as enthusiastically supportive of the student tutorials and practice opportunities, stating they were "excellent" and students were prepared to use the test system. There is evidence that students were prepared because there were no system related questions

from students during the test session HumRRO observed and the TAs interviewed were unable to recall any system tool or navigational questions during testing.

Information such as item types, sample items for each item type, and scoring rubrics of the writing component is available to test takers before the test date. However, practice materials may not be available in formats that can be accessed by all test takers. We did not find practice materials in a form that can be accessed by students with disabilities. Practice materials may not be suitable for students with certain disabilities (e.g., deaf or hard of hearing and sign language accommodation), but practice materials with some types of accommodations (e.g., large-print) can be provided to make the materials more accessible to test takers.

During the writing administration observation, it was noted that one student used and periodically referred to the Writer's Checklist (available as a resource on the test system). The observer asked the TA if all students knew about that document (and the TDA Scoring Guidelines) and she replied that yes, they do, "…probably better than the teachers." A TA interviewed at another school stated that almost half of the students accessed the TDA Scoring Guidelines during the test and used both documents regularly in the classroom. However, TAs at another school were not familiar with either document.

It is important to also note that STCs and TAs were asked about the practice opportunities for students with disabilities. No one indicated there were any limitations they were aware of during the practice sessions for this student population.

Four TAs at one school stated they felt the reading passages were too lengthy and the rigor was too high for their students. Also, they said there were no practice TDA items similar to what students saw when testing. TAs at two schools stated that their low performing students can use a dictionary or hear a word pronounced during instruction and other testing situations. For this test, they felt these students were disadvantaged by not, at a minimum, having an audio link to hear the pronunciation of a word they did not recognize.

Standard 6.1 – Test administrators should follow carefully the standardized procedures for administration and scoring specified by the test developer and any instructions from the test user.

Evidence from the documents indicates that key aspects of Standard 6.1 are covered. DRC provided appropriate training and documentation so that TAs understand the standardized procedures to follow. The TAMs include accepted standardized procedures for determining accommodations, minimum technology requirements, test time limits, test make-up policies, and other acceptable variations in test administration. There are training and pretest workshops for TAs, STCs, and technology coordinators.[5]

The TAs and STCs were asked during the school visit interviews about TA training. All reported that the STC provided the training and it generally consisted of reading the TAM with a face-to-face meeting to discuss key tasks and policies. TAs were asked specifically if they felt prepared to administer the exams after training and all reported that they did. Also, they stated that they encountered no problems when they

[5] We did not observe actual live training sessions and our evaluation is based on the training materials only.

administered the writing test. When observing the writing exam, no issues were evident to indicate that the TA was unprepared to administer the test.

The training materials provide instructions for TAs for when they need to make adjustments if an accommodation is required. In the SC READY training materials, some exceptions for administering the assessments in the online format are specified. For example, students who cannot take online assessments due to their disabilities, as specified in their IEPs or 504 plans, may be tested in a paper-based format. In the Training tool slides (*Spring 2017 EOCEP STC TA Training Tool.pptx, Spring 2017 SC READY_SCPASS STC TA Training Tool.pptx*), the test developers provide case scenarios related to test security to train TAs to deal with different test security issues. Similar hands-on training or concrete examples for other phases of administration could be provided to TAs as well to improve the training to ensure that TAs carefully follow the standardized procedures. Additionally, we did not find documentation about usability studies or empirical research related to topics of test administration.

Standard 6.4 – The testing environment should furnish reasonable comfort with minimal distractions to avoid construct-irrelevant variance.

The test administration processes follow this standard very well. In both the EOCEP and the SC READY TAMs, there is a section about the testing environment that specifies standards to be followed to provide a reasonably comfortable testing environment to test takers. The guidance specifies that schools must adhere to several standards to ensure that all students have an equal opportunity to perform their best on the test. Some examples of these standards include "tests should be administered in a familiar classroom or computer lab setting to reduce student test anxiety and simplify test security," "students should be tested in classrooms or computer labs that have good lighting and are well-ventilated with a reasonable temperature," and "classrooms and computer labs should be quiet and free from interruptions or distractions of any type." The technical guide documents (*DRC INSIGHT Technical Guide* and *eDirect User Guide*) provide technical instructions for using the online testing system. This helps to reduce distractions due to internet connectivity issues and technology failures and avoid construct-irrelevant variance.

The school visits also confirm that this standard is met. The observer noted during the writing assessment administration that the TA seated students using the seating chart and they had plenty of space to work. The TA maintained a positive testing environment and no sidebar student conversations were observed. One student's computer locked up after being paused. The TA asked the district technical support person for help and ultimately called the Help Desk. The computer was shut down, the student logged back in, and no data were lost. The issue was resolved in about 10 minutes with minimal disruption. Other students continued working and remained engaged throughout the exam.

At least three TAs at one school stated they were concerned about the maximum student to TA ratio, found on page 23 of the TAM. It states that when the number of students reaches 35, a test monitor must be added. The TAs agreed it would not be possible for one TA to provide a good testing environment for 35 students and suggested the number should be significantly reduced.

Standard 6.5 – Test takers should be provided appropriate instructions, practice, and other support necessary to reduce construct-irrelevant variance.

Instructions to test takers regarding how to respond and interact with the test delivery interface are clearly indicated in the TAMs, ADMs, Online Tools Training (OTT), and student tutorials.

Guidance for how to interact with and navigate the delivery platform, use the available tools, and respond to items are provided. The *eDIRECT User Guide* and the TAMs state that STCs and TAs are responsible for (a) reviewing the OTT and Tutorial prior to testing, and (b) ensuring that students practice on the device they will be taking the operational test prior to testing.

While the OTT and the Tutorial adequately address the issue of test takers being provided appropriate instructions and practice prior to operational testing, the documents we reviewed do not detail the part of the standard that addresses monitoring those practice opportunities. The documents provide little information regarding providing guidance to the STCs and TAs to ensure that the practice opportunities lead to students acceptably interacting with the testing engine (e.g., navigating, marking responses).

During the school visits, STCs and TAs were asked about students' receiving an opportunity to participate in the OTT prior to testing. The staff at all four schools indicated that although there is not a formal tracking mechanism in place to track and report which students had not attended an OTT, the students' classroom teachers were tasked to ensure their students attended the OTT. It appears that all students at these four schools attended the OTT as evidenced by the lack of student questions about system tools or navigation during the EOC writing exam. This was also noted during the test observation.

One area of importance with online testing is that students understand how to scroll through passages commonly seen on ELA tests (and sometimes in other subjects). The EOCEP English 1 and Biology 1 passage navigation (as evidenced by our review of the OTT) has a seamless transparent blue bar with white font indicating if there is more text to scroll through at the bottom and top of the passage screen. The SC READY ELA test, however, uses a pagination navigation screen at the bottom of the passage. For example, if a passage has four pages to scroll through, the bottom left of the passage will say 'Page 1 of 4.' However, clicking to the next page is not immediately made clear—to do so, one must click the right side of the passage to advance forward or the left side to go backward. The script in the ADM does include specific instructions on how to navigate, but the OTT and Tutorial does not directly address this issue. We have some concerns that younger students, in particular, may have difficulty accessing the entire passage without appropriate practice, exposure, and guidance. The scrolling passage navigation as used in the EOCEP assessments might be easier for younger students; however, consideration to which passage navigation is most intuitive and easiest for younger students should be guided by usability studies or cognitive labs.

Additionally, there were some aspects of the Tutorial that might use language that is too advanced for younger students. For example, "The ELA test will be a two-day test. For ELA Session 1, the extended response item will be a text dependent analysis or TDA item" could use simpler language or more teacher-guided direction for younger students.

Standard 6.6 – Reasonable efforts should be made to ensure the integrity of test scores by eliminating opportunities for test takers to attain scores by fraudulent or deceptive means.

This standard includes providing (a) safeguards against fraudulent activities at the local school sites and during administration, and (b) measures to detect cheating during scoring processes. This standard was only reviewed in relation to documented procedures for ensuring the integrity

of scores during test administration processes.[6] The EOCEP and the SC READY TAMs provide a separate section on test security including, state board regulations, reporting and documenting violations, and separate policies and procedures for administering online and paper tests. The TAMs also provide guidance for TAs to help reduce cheating by requiring seating charts, completion of security checklists, and providing helpful tips on how to separate students (e.g., privacy folders, space). The TA script in the ADMs includes a statement about the prohibition of electronic devices. Additionally, the training PowerPoint® files include several test security case scenario vignettes to help standardize TA understanding and implementation of test security policies and procedures.

The school visits provide evidence that test security is a high priority for school staff. HumRRO observed the TA use the seating chart and the STC provide the test tickets after students were seated for the writing assessment. The STCs and TAs who were interviewed at the three other high schools confirmed they followed the same process when administering the assessment. STCs were asked if they had completed and submitted a Test Violation Form this year. None had done so, but they said, if needed, the completed form is submitted to the District Test Coordinator.

One area that would benefit from additional specification relates to preventing breaches of accommodation policies. The TAMs identify the procedures to take should a violation occur, but there is little guidance on how to identify or minimize such breaches. It is possible, based on the criteria of who is eligible to serve as a TA, that the TA might not have sufficient knowledge of IEP/504 accommodations to be able to identify when a breach might occur.

Standard 6.7 – Test users have the responsibility of protecting the security of test materials at all times.

This standard largely means that all test users (at all levels) have the responsibility of protecting and securing test materials. Our review excludes documentation of procedures related to state agency actions (e.g., documents shown in court challenges) and focused on the responsibilities of those at the district and school level. The EOCEP and SC READY TAMs and training slides state the criteria for eligible DTCs, STCs, TAs, and Monitors and provides general requirements for ensuring test materials remain secure at all times. The TAMs include an overview of state laws regarding test security, completing required forms and checklists, and handling, storing, and returning materials.

Test Tickets are secure documents not to be distributed to students until immediately before logging into the test session (TAM page 10, 31, and more). STCs confirmed during the interviews that they print the Test Tickets and do not give them to the TA until students are seated in the testing room. One STC stated that he maintains full control of them from printing to collecting them from students at the end of the test. HumRRO observed the STC help the TA distribute the Test Tickets to students once they were seated.

[6] The scope of our Phase 2 evaluation reflects the documentation regarding test security processes, and not whether these policies and procedures are carried out with fidelity. For example, the TAM states, "the school should follow policies and procedures established by the district for investigating and documenting suspected cheating incidents (EOCEP p. 20)," but there is no specific guidance of what those district policies should include.

Other Findings

There were two additional topics expressed during the interviews that did not fit well in the Rationale for the Standards section. First, a concern was expressed by the TAs and STCs about the writing exam testing window and schedule. The interviewees reported that testing once a week for three weeks is difficult for a large school (200+ students testing at one time) to accommodate, and if any problems are encountered then school staff must wait a week to try again. To help alleviate this concern, they suggested increasing the length of the testing window or adding additional days on which testing is permitted to occur. Second, the interviewees felt that test monitors should be allowed to stay in the testing room without the TA present because there are times when a discussion between the TA and STC outside of the testing room is needed; they indicated that it was inefficient to pass information through an intermediary.

Discussion

Our evaluation of the EOCEP English 1 assessment for the current report focused on the observation of the writing test administration and interviews with Test Administrators (TAs) and School Test Coordinators (STCs). We generally found that the test administration processes for the English 1 writing component reflect the *Test Standards* pertinent to test administration. Based on the additional evidence we collected via the test administration observation and interviews, we increased a rating presented in the Phase 2 report on *Test Standard* 6.1 from a '4' to a '5' ('1' indicates no evidence to support the standard and '5' indicates evidence fully supports the standard). Thus, the overall mean rating across all the standards increased slightly from M = 4.43 (SD = 0.62) to M = 4.50 (SD = 0.65). With the exception of one standard (Standard 6.3), we found that the policies and procedures mostly or fully address the key aspects of industry standards pertaining to test administration (the rating for *Test Standard 6.3* remains a '3' given lack of guidance in the test administration materials on documenting and reporting changes or disruptions to standardized test administration procedures).

Based on the observation of test administration and interviews, we found that the test administration procedures—both prior to the start of testing and during testing—were completed as described in the TAM. For example, test takers were arranged in seating assignments with ample space between students; materials with testing-related content had been removed from the walls (e.g., posters displaying content pertaining grammar usage) prior to testing; student electronic devices were collected prior to the start of testing; and student Test Tickets were handled per instructions in the TAM. During testing, the TA and STC circulated around the room regularly to monitor students. When student questions arose, those questions were addressed promptly and appropriately by the TA. Only one technical issue arose during test administration at the site observed, and it was promptly addressed and resolved. Across the four schools visited, none of the TAs nor STCs reported concerns about technical issues or problems occurring during testing. Moreover, findings from the observation and interviews indicate that students navigated through the online assessment with ease and encountered little to no difficulty using the online tools. During the interviews, the TAs and STCs reported that they believed that the student tutorial and practice opportunities were useful for ensuring that students understood how to navigate through the online system and use the online tools during testing. The TAs and STCs also reported that they felt that the training and supports they had received adequately prepared them to administer the writing assessment. Finally, there were no indications, either from the observation or interviews, of threats to the security of test material.

Overall, the evidence collected suggests that test administration practices are consistent with industry standards. Nonetheless, some opportunities to further strengthen and improve test administration were found. First, the script read by TAs to test takers is lengthy. During the observation of test administration, the observer noted that it took approximately 15 minutes to read the script. During interviews, the TAs and STCs also reported that the script is too lengthy and repetitive. They reported that students become impatient during the period of time that TAs read the script. They recommended that some of the repetition in the script could be removed, for example, by highlighting text that could be skipped over if students have already taken a test in another subject. They also recommended presenting information in the script in concise bullets, rather than lengthy sentences. The TAs also expressed that the TAM could be better organized to make it easier to find information. For example, they recommended organizing key information by time periods—i.e., test preparation activities, test administration activities, and test close-out activities.

Regarding the new item type—the TDA item, the interviewees indicated that students do not typically complete high quality writing (e.g. essays) in a single sitting. Rather, students are typically taught to write a draft, and then later go back and revise and finalize their draft. Interviewees also reported that some students worked on their essays for more than four hours. A suggestion was offered to consider allowing students to break up the writing component of the English 1 assessment so that students could write a draft during an initial testing session, and then revise and finalize their draft in a subsequent testing session. Finally, the SCDE may want to check for consistency across schools regarding awareness and usage of the writing resources (Writer's Checklist and TDA Scoring Guidelines) available for the writing assessment. Three of the four sampled schools reported awareness and use of these resources, but one school did not.

Finally, it is important to note that the observation and interviews were based on a very small sample, which limits the generalizability of these findings. It is important to keep this caveat in mind when interpreting results from the observation and interviews.

Chapter 3: Review Scaling, Equating, and Scoring Processes (Task 5)

Introduction

Because a new item type—the text-dependent analysis (TDA) item—was included on the 2017-18 English 1 assessment (i.e., the Writing component), we revisited Task 5 to include a review of the handscoring processes for the TDA item. Handscoring refers to the scoring of student responses by experienced, human scorers as opposed to the scoring of student responses by machines. Thus, the purpose of this task is to document the extent to which the handscoring processes for the TDA item follow industry best practices as described in *The Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014; hereafter *Test Standards)*. In the current task, we focus on the handscoring processes, quality control procedures, and rater qualifications for the TDA item. This chapter presents updates to our findings provided in the Task 5 chapter from our Phase 2 report (Sinclair & Thacker, 2018).

Methods[7]

We conducted a systematic document review, guided by industry standards, to evaluate the handscoring processes for the TDA item on the EOCEP 2017-18 English 1 assessment. We worked in cooperation with the Data Recognition Corporation (DRC) to obtain the necessary documentation. Table 3.1 lists the documents we collected and reviewed. The additional documents reviewed during Phase 3 (current phase) are highlighted in bold in the far-right column. These documents provided information about the steps and procedures related to handscoring for the TDA item.

The evaluation of the documentation in the last column of Table 3.1 was guided by the *Test Standards*. In Phase 2 of this evaluation, we identified 16 standards from the *Test Standards* that are directly relevant to scoring. During Phase 2, we rated the extent to which the evidence adheres to these *Test Standards.* For the current phase, we revisited our Phase 2 ratings to reflect the documentation relevant to handscoring. The scale on which our ratings were based is presented in Table 3.2. This is the same rating scale that was used in the earlier phase of our evaluation. Two HumRRO researchers independently assigned a rating based on evidence reviewed and then compared and discussed their initial ratings and rationales to reach a final consensus rating for each relevant standard.

[7] The process for reviewing materials for adherence to relevant *Test Standards* is the same process as used in Task 4 (Review of Test Administration).

Table 3.1. Documents Reviewed for Task 5 – Equating, Scaling, and Scoring

Document Focus	Document/Folder File Name	Relevant Assessment(s)	
		EOCEP (Algebra 1, English 1, Biology 1) Phase 2	EOCEP (English 1 - TDA) Phase 3
Technical specifications for item calibration, equating, and scoring. Technical reports and special studies.	[a]024F_EOCEP Reports_Technical_Standard Setting_Special Studies	X	
	[a]025F_SC READY Reports_Technical_Standard Setting_Special Studies		
	[a]029F_Reading PLDs		
	SC-MAP-Linking-Study		
Documentation of item scoring procedures; Quality assurance processes for automated scoring	[a]028F_Phase I_Item Development _Forms Construction Document	X	
	043_Item Scoring and Quality Control	X	
Scorer training materials (TDA only).	[a]015F_SC READY Scorer Training Materials		
	[a]033F_EOCEP Handscoring Materials		**X**
Criteria for scorer qualification (TDA only)	039_SC READY Scorer Qualification		
	033F_EOCEP English 1 Scorer Qualifications		**X**
Processes for monitoring scorer accuracy and consistency (TDA only)	040_SC READY Scorer Accuracy and Consistency		
	033F_EOCEP English 1 Scorer Accuracy and Consistency		**X**
Documentation related to creation of vertical scales (SC READY only)	[a]027_2017 SC READY Vertical Equating		
	042_SC READY Creation of Vertical Scales		
	047_SC READY Horizontal_Vertical Linking Process		
Sample 2016-2017 student and school score reports	041_EOCEP Score Report Users Guide	X	
	045_Spring 2017 SC READY Score Report Users Guide		
Sample 2017-2018 student and school score reports	**036F_5.3_EOCEP ISR Mockup**		**X**
	036F_5.3_School Roster Mockup		**X**

Note. [a]Indicates a folder including multiple files.
New documents and folders added in Phase 3 are highlighted in **bold**.

Table 3.2. Rating Scale for Evaluating Strength of Evidence for Test Standards

Rating Level	Description
1	No evidence of the Standard found in the materials.[a]
2	Little evidence of the Standard found in the materials; less than half of the Standard covered in the materials and/or evidence of key aspects of the Standard could not be found.
3	Some evidence of the Standard found in the materials; approximately half of the Standard covered in the materials, including some key aspects of the Standard.
4	Evidence in the materials mostly covers the Standard; more than half of the Standard covered in the materials, including key aspects of the Standard.
5	Evidence in the materials fully covers all aspects of the Standard.

[a] Materials include all documents and data provided, any emails or phone calls with SCDE/DRC staff, as well as what could be found online.

Results

Results are organized around the relevant *Test Standards* and include details from our documentation review to support judgments about the extent to which industry standards are met. Table 3.3 provides the ratings from Phase 2 with the Phase 3 updates presented in bold text (n = 2). Standards 6.8 and 6.9 were updated from "not applicable" ratings in Phase 2 to '5s' in Phase 3, based additional evidence pertinent to handscoring of the English 1 TDA item. Thus, of the 10 relevant standards, five received a rating of '4' and five received a rating of '5' (*M* = 4.50; *SD* = 0.50).

Table 3.3. Evaluation Results Based on the Test Standards

Standard Number	Standard Content	EOCEP Rating
Standard 5.1	Test users should be provided with clear explanations of the characteristics, meaning, and intended interpretation of scale scores, as well as their limitations.	4
Standard 5.2	The procedures for constructing scales used for reporting scores and the rationale for these procedures should be described clearly.	4
Standard 5.5	When raw scores or scale scores are designed for criterion-referenced interpretation, including the classification of examinees into separate categories, the rationale for recommended score interpretation should be explained clearly.	5
Standard 5.6	Testing programs that attempt to maintain a common scale over time should conduct periodic checks of the stability of the scale on which the scores are reported.	4
Standard 5.12	A clear rationale and supporting evidence should be provided for any claim that scale scores earned on alternate forms of a test may be used inter-changeably.	4
Standard 5.13	When claims of form-to-form score equivalence is based on equating procedures, detailed technical information should be provided on the method by which equating functions were established and on the accuracy of the equating functions.	4

Table 3.3. (Continued)

Standard Number	Standard Content	EOCEP Rating
Standard 5.21	When proposed score interpretations involve one or more cut scores, the rationale and procedures used for establishing cut scores should be documented clearly.	5
Standard 5.22	When cut scores defining pass-fail or proficiency levels are based on direct judgments about the adequacy of item or test performances, the judgmental process should be designed so that the participants providing the judgements can bring their knowledge and experience to bear in a reasonable way.	5
Standard 5.23	When feasible and appropriate, cut scores defining categories with distinct substantive interpretations should be informed by sound empirical data concerning the relation of test performance to the relevant criteria.	5
Standard 6.8	**Those responsible for test scoring should establish scoring protocols. Test scoring that involves human judgement should include rubrics, procedures, and criteria for scoring. When scoring of complex responses is done by computer, the accuracy of the algorithm and processes should be documented.**	**5**
Standard 6.9	**Those responsible for test scoring should establish and document quality control processes and criteria. Adequate training should be provided. The quality of scoring should be monitored and documented. Any systematic source of scoring errors should be documented and corrected.**	**5**
Standard 6.10	When test score information is released, those responsible for testing programs should provide interpretations appropriate to the audience. The interpretations should describe in simple language what the test covers, what scores represent, the precision/reliability of the scores, and how score are intended to be used.	4

Note. Updated EOCEP ratings based on the handscoring documentation are highlighted in bold.

Ratings on two of the *Test Standards* were updated to reflect the new documentation on handscoring. The following section provides rationales for our two updated ratings. Then we provide detailed findings for the three areas that were the focus of our review: handscoring processes, quality control procedures, and rater qualifications.

Standard 6.8 -Those responsible for test scoring should establish scoring protocols. Test scoring that involves human judgement should include rubrics, procedures, and criteria for scoring. When scoring of complex responses is done by computer, the accuracy of the algorithm and processes should be documented.

Evidence in the materials thoroughly covers all relevant aspects of this Standard. Training materials, which include scoring protocols and procedures, are provided in the folder *033F_EOCEP English 1 Handscoring Materials/EOCEP English 1 Scorer Training Materials*. Adequate training was provided for each TDA item. Each TDA item requires item-specific training materials, including

a scoring guide composed of a rubric, a passage, and three annotated anchor responses per score point. Following rangefinding, Scoring Directors compose anchor and training sets of committee-scored responses for each item to be scored. For each TDA item, DRC develops two training sets and two qualifying sets of 10 student responses each. Anchor responses are selected to illustrate scoring concepts. These training and qualifying materials were used to further indicate each rater's ability to discern the different score-point levels accurately and consistently.

Scoring guidelines and rubrics are clear and comprehensive. The *EOCEP English 1 Scoring Guidelines_Rubric* document contains detailed scoring criteria. According to the rubric, two sets of skills were measured—the capabilities of analysis and the skills of writing. Four anchor scores were defined in the rubric: demonstrates effective analysis of text and skillful writing (score=4), demonstrates adequate analysis of text and appropriate writing (score=3), demonstrates limited analysis of text and inconsistent writing (score=2), and demonstrates minimal analysis of text and inadequate writing (score=1). Per the *EOCEP English 1 Scoring Guidelines_Rubric*, raters score the TDA item based on the following seven criteria:

- how effectively all parts of the task are addressed
- depth of analysis based on explicit and implicit meanings from the text to support claims, opinions and ideas;
- reference to the main ideas and relevant details using details, examples, quotes, and/or facts;
- organizational structure of the writing;
- use of transition to link ideas;
- use of precise language and domain-specific vocabulary, and the
- amount of error in sentence formation, grammar, usage, spelling, capitalization, and punctuation.

Detailed step-by-step instructions for using the online scoring system (online TQR[8] application) can be found in the *Introduction of the Scoring System* document. The Online TQR application provides an automated method of training, qualifying, and recalibrating readers on handscoring items. The Online TQR application also allows for the administration of sets, reader scoring of sets, and the collection of the results in the form of reports.

Standard 6.9 - Those responsible for test scoring should establish and document quality control processes and criteria. Adequate training should be provided. The quality of scoring should be monitored and documented. Any systematic source of scoring errors should be documented and corrected.

Evidence in the materials thoroughly covers all aspects of this Standard. Rater training procedures and anchor sets are provided in the folder *033F_EOCEP English 1 Handscoring Materials/EOCEP English 1 Scorer Training Materials*. Criteria for rater qualification and training procedures are detailed in the document *EOCEP English 1 Scorer Qualifications.* As described in the *EOCEP English 1 Scorer Accuracy and Consistency* document, quality control processes include the distribution of validity responses to scorers, inter-rater agreement monitoring, and supervisor spot-checks.

[8] The TQR acronym was not defined in the documentation provided.

Pre-scored validity papers were interspersed among live student responses. These validity papers were not identified to raters, thereby allowing scoring supervisors to compare raters' scores to the reference score. Throughout TDA item scoring, a rater must maintain at least 70 percent exact agreement on validity checks. Validity reports are produced daily to detect rater drift at the individual, team, or room level. If a drifting trend is detected, raters are re-trained before resuming scoring, and all responses scored by these raters since the last acceptable validity check are rescored. Throughout handscoring, daily and cumulative reports of inter-rater agreement are produced and reviewed. Inter-rater agreement monitors how often raters are in exact, adjacent, and nonadjacent agreement with each other. A rater must maintain at least 70 percent exact agreement on daily and cumulative inter-rater agreement checks, a criterion that is considered to reflect substantial inter-rater consistency (Landis & Koch, 1977). However, no inter-rater agreement report or computer log was provided for the current documentation review.

The Image Handscoring System randomly selects responses that have been scored by raters and forwards those selected to supervisors for spot checks. Typically, one out of five records are monitored. If there is disagreement between the supervisor and the rater, the supervisor corrects the score and then uses the response to retrain the rater. A Read-Behind Log was used by the team leader/scoring directors to monitor inter-rater agreement. However, we were not provided with an example Read-Behind Log for review.

Discussion

We evaluated the handscoring documentation for the EOCEP English 1 2017-18 assessment. Our document review focused on components of the *Test Standards* that specifically address scoring by human scorers. We found that the TDA item is scored using a detailed rubric. Scorers are trained on examples of each possible score point, and are required to pass training and qualifying rounds before being able to participate in operational scoring. Validity papers are interspersed among operational papers, allowing for detection of scorer drift. Inter-rater agreement is monitored daily through the online scoring system, and uses appropriate criteria for triggering retraining/recalibration. Overall, the handscoring processes of the EOCEP English 1 assessment adhere to industry best practices based on the available documentation. To further bolster their validity evidence for the handscoring procedures, DRC should keep documentation of rater agreement reports and Read Behind Logs and make those reports available for audits/independent reviews of their processes. We have no other recommendations for improving English 1 handscoring processes based on the results of this task. It should be noted that a limitation of this investigation is that it was based on a review of available documentation. Observation of handscoring training and interviews with trainers and scorers would further strengthen the evidence supporting these findings.

Chapter 4: Review of Psychometric Processing and Item Parameters (Task 6)

Introduction

HumRRO conducted a review of item parameters for the English 1 EOCEP fall/winter 2017-18 assessment[9]. This task addresses the RFP's request for a specific evaluation of psychometric validity. It replicates analyses conducted in Phase 2 of the evaluation, providing updated results for English 1 following the addition of the new TDA item type in 2017-18. The review of item parameters addresses the following elements of psychometric validity outlined in the RFP:

- Is the difficulty level of the item appropriate?
- Are the item discrimination statistics acceptable?
- Do the item characteristics support that the items were written in such a way as to reduce the likelihood that a student could get the item correct by guessing?

Method

HumRRO received an item-level data file from the English 1 EOCEP fall/winter 2017-18 administration. For each item, indexes of classical test theory (CTT)—item difficulty (p-values) and item discrimination (item-total correlation) were provided. For multiple-choice items, the percentage of students selecting each response option and point-biserial correlations were also provided. Also, for each item, the Rasch item difficulty was provided. Finally, differential item functioning (DIF)[10] categories were provided with 'A' indicating little or no difference between groups (male/female and white/other), 'B' indicating small to moderate differences, and 'C' indicating substantial differences.

We first calculated the distribution of CTT item difficulty and discrimination statistics for each item type. Next, we flagged items with CTT item difficulty and discrimination statistics that failed to fall within an acceptable range of values (i.e., $p < .10$, $p > .95$, and item-total correlation < .10). These flags are based on work HumRRO has done previously for another assessment program and were selected because they reflect more stringent criteria than the key check criteria provided by DRC. While DRC has documented key check criteria, the DRC criteria were not employed in the current review as they are intended to identify items for potential mis-key issues, not items that may not belong in the item bank. The number of items flagged for differential item functioning (DIF) in category C among gender and ethnicity subgroups was also analyzed. Finally, Rasch item difficulty was analyzed and items were flagged for high (difficult) or low (easy) values.

Results

Table 4.1 presents a summary of CTT item difficulty statistics for the operational English 1 EOCEP items from the fall/winter 2017-18 administration. Items with p-values greater than .95 were very easy for this group of examinees, while items with p-values less than .10 were very difficult. Items that are very easy or very difficult provide little information on student achievement, and so ideally item p-values should fall between .10 and .95. As Table 4.1 shows, no item on the English 1 fall /winter 2017-2018 assessment was flagged for a p-value falling outside of this acceptable range, which suggests that all items have appropriate difficulty levels.

[9] Data from the 2018 spring administration was not available in time to include in this report.
[10] DIF occurs when students of approximately equal ability in different groups perform in substantially different ways on a test question.

The CTT item difficulty statistics in Table 4.1 indicate that the TDA item difficulty was in the appropriate range. The TDA item was roughly equivalent in difficulty to the harder of the two evidence-based selected response (ESBR) and technology enhanced (TE) items (based on the minimum p-values), and somewhat harder than the easier of the two ESBR and TE items (based on the maximum p-values). The TDA item difficulty was lower than the mean level item difficulty of the other item types, indicating that the TDA item was somewhat more difficult relative to the other item types, on average. However, the item data file from DRC did not include the calculation or supplemental documentation of how the p-value was calculated for the TDA item.

Table 4.1 Item Difficulty Analysis: English 1 EOCEP (fall/winter 2017-18)

Item Type	Item p-values					Item Difficulty Flags % (N)	
	N	Min	Max	Mean	SD	p-value above .95	p-value below .10
EBSR	2	.448	.575	.512	.090	0	0
MC	53	.295	.844	.591	.129	0	0
[a]TDA	1	.451	.451	.451	NA	0	0
TE	2	.447	.530	.489	.059	0	0

Note. MC= Multiple choice; EBSR = Evidence-based selected response; TDA= Text-dependent analysis; TE = Technology enhanced; N = number of items; Min = Minimum; Max = Maximum; SD = Standard Deviation; NA = Not applicable.
[a]Value for Min, Max and Mean are identical because only one TDA item was administered; standard deviation is not applicable when only one item is analyzed.

Table 4.2 presents a summary of CTT item discrimination statistics for the operational English 1 EOCEP items from the fall/winter 2017-18 administration. Items with item-total correlations less than .10 do not help differentiate between students who are low performing and students who are high performing in English 1. As Table 4.2 shows, no item was flagged for a low item-total correlation. The TDA's item discrimination level was higher than the other item items at .624, indicating that the item discriminates very well between low and high performing students. However, again, the item data file from DRC did not include the calculation or supplemental documentation of how the item-total correlation was calculated for the TDA item.

Table 4.2 Item Discrimination Analysis: English 1 EOCEP (fall/winter 2017-18)

Item Type	Item-Total Correlations					Item Discrimination Flags % (N)
	N	Min	Max	Mean	SD	Item-total correlation below .10
EBSR	2	.360	.534	.447	.090	0
MC	53	.191	.496	.383	.077	0
[a]TDA	1	.624	.624	.624	NA	0
TE	2	.403	.476	.440	.052	0

Note. MC= Multiple choice; EBSR = Evidence based selected response; TDA= Text-dependent analysis; TE = Technology enhanced; N = number of items; Min = Minimum; Max = Maximum; SD = Standard Deviation; NA = Not applicable.
[a]Value for Min, Max, and Mean are identical because only one TDA item was administered; standard deviation is not applicable when only one item is analyzed.

Table 4.3 presents a summary of results from DIF analyses for operational English 1 items from the fall/winter 2017-18 administration. As Table 4.3 demonstrates, only one multiple-choice item (MC) was flagged for C DIF for the female/male comparison. This suggests that females with the same ability level as males have a lower probability of getting this item right. The TDA item was not flagged for DIF among either the gender or ethnicity subgroups. The presence of DIF is a necessary, but insufficient indicator of bias. It should be used to trigger further scrutiny of an item. Furthermore, given that only one item was flagged for DIF, this suggests that there were no systematic fairness issues with the operational English 1 items from the fall/winter 2017-18 administration.

Table 4.3 Differential Item Functioning (DIF) Analysis: English 1 (fall/winter 2017-2018)

Item Type	N	C DIF Flags % (N)	
		Female/Male	Black/White
EBSR	2	0	0
MC	53	1.89 (1)	0
TDA	1	0	0
TE	2	0	0

Note. MC= Multiple choice; EBSR = Evidence based selected response; TDA= Text-dependent analysis; TE = Technology enhanced.

Table 4.4 summarizes Rasch item statistics from the English 1 fall/winter 2017-18 administration. Items with Rasch difficulty below -2 were very easy for this group of examines, while items with Rasch difficulty above 2 were very difficult for this group of examinees. Items that are very easy or very difficult contribute little information to our understanding of student achievement, and so ideally Rasch difficulty will fall between -2 and 2. In contrast to the CTT item difficulty results, the TDA item was flagged for a difficulty level that falls outside of the acceptable range (Rasch difficulty above 2). A closer look at this item indicates that its difficulty level is very close to the acceptable range, with the Rasch difficulty parameter equal to 2.075. However, the item data file from DRC did not include IRT step parameters for the various score points (1, 2, 3, or 4) on the polytomously-scored TDA item. This item may be contributing considerable information, for example, at the '1' and '2' score levels; however, this is unknown based on the overall Rasch difficulty statistic reported in the item data file. It is not surprising that the Rasch difficulty parameter for the TDA item was above 2 given that constructed-response items generally show higher difficulty compared to selected-response formats in writing and reading assessments, despite measuring the same latent trait (Downing, 2009; Hohensinn & Kubinger, 2009). Finally, no item was flagged for low item difficulty (Rasch difficulty below -2). Overall, the available Rasch item statistics indicate that the fall/winter 2017-18 operational English 1 EOCEP items measured student achievement in English 1 at appropriate levels of difficulty and that items functioned as intended.

Table 4.4 Rasch Item Statistics: English 1 (fall/winter 2017-18)

Item Type	N	Rasch Empirical Item Difficulty					
		Min	Max	Mean	SD	Rasch difficulty above 2 % (N)	Rasch difficulty below -2 % (N)
EBSR	2	.479	.806	.642	.231	0	0
MC	53	-1.587	1.865	.069	.699	0	0
[a]TDA	1	2.075	2.075	2.075	NA	100 (1)	0
TE	2	.415	.863	.639	.316	0	0

Note. MC= Multiple choice; EBSR = Evidence based selected response; TDA= Text-dependent analysis; TE = Technology enhanced; N = number of items; Min = Minimum; Max = Maximum; SD = Standard Deviation; NA = Not applicable.
[a]Value for Min, Max, and Mean are identical because only one TDA item was administered; standard deviation is not applicable when only one item is analyzed.

Discussion

Our review of the item-level data from the fall/winter 2017-18 administration of the English 1 assessment indicates that overall, items are appropriately difficult and discriminate among student ability levels. Of special interest in Phase 3 was the newly added text-dependent analysis (TDA) item. The TDA item demonstrated appropriate levels of CTT item difficulty and discrimination and did not show DIF among gender and ethnic groups. Although Rasch item statistics indicate that the TDA item was difficult for the students tested, its difficulty level was only slightly above the upper bound of the acceptability range. The slightly high item difficulty of the TDA item compared to other English 1 items might be due to the constructed response nature of the item and students' lack of familiarity with this particular item format. Familiarity with item content can make an item more relevant, engaging, and more easily understood, and can then lead to differential performance, even for examinees of the same ability level (Alonzo, 2012). Because this is a new item type, and because only one TDA item was administered, we are unable to draw strong conclusions about observed differences in student performance across the item types. Combining multiple item formats may maximize the impact of positive item features, while at the same time minimizing their limitations (Messick, 1993). Therefore, we find that the inclusion of TDA and other item types to be in line with industry best practices. We recommend continued monitoring of the psychometric properties of the TDA (in particular) and other item formats during subsequent test administrations. Regarding continued psychometric monitoring, we recommend that DRC provide additional detail on how CTT statistics (p-values and item-total correlations) are computed for the polytomously-scored TDA item, and include IRT step parameters for the TDA to provide greater insight about the information being contributed by each score level of the TDA item.

Chapter 5: Summary and Conclusions

This third and final report completed a comprehensive, external evaluation of the EOCEP English 1 assessment. The evaluation entailed six tasks related to the design, administration, scoring, and reporting of the assessment:

- Task 1: Review Item Development Processes
- Task 2: Review Items to Standards Alignment and Item Quality
- Task 3: Review Test Construction Processes
- Task 4: Review Test Administration Procedures
- Task 5: Review Scaling, Equating, and Scoring Processes
- Task 6: Review Psychometric Processing and Item Parameters

The results from Tasks 4 – 6 are presented in the current report for the 2017-18 English 1 assessment. The findings from Tasks 1-3 for English 1 are presented in Report #2 (Sinclair & Thacker, 2018). Thus, findings from this report should be considered in combination with the findings presented in Report #2.

Overall, the findings from these tasks indicate that the test administration practices, the handscoring processes, and the item parameters for the English 1 assessment support industry best practices as described in *The Standards for Educational and Psychological Testing*, and thereby support the validity of the test scores for their intended uses and purposes. No critical concerns were identified from the technical evaluation of the English 1 assessment. Some recommendations were offered for further improvement. We applaud South Carolina for securing an external evaluation of its assessments to help ensure their quality. Periodic evaluations of testing practices will help to ensure their continued technical soundness.

References

Alonzo, A. C. (2012). Eliciting student responses relative to a learning progression. In Learning progressions in science (pp.241-254). SensePublishers, Rotterdam.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Dickinson, E.R., Chen, J., & Swain, M. (2017). *South Carolina Assessment Evaluation Report #1.* (2017 No. 019). Alexandria, VA: Human Resources Research Organization.

Downing, S. M. (2009). Written tests: Constructed-response and selected-response formats. Assessment in health professions education. New York: Routledge, 149-84.

Hohensinn, C., & Kubinger, K. D. (2009). On varying item difficulty by changing the response format for a mathematical competence test. *Austrian Journal of Statistics*, 38(4), 231-239.

Messick, S. (1993). Trait equivalence as construct validity of score interpretation across multiple methods of measurement. In R. E. Bennett & W. C. Ward (Eds.), Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment (pp. 61-73). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Schulz, S. R., Wiley, C. R. H., Buckland, W. W., Michaels, H. R., Diaz, T. E., & Chen, J. (2017). *NAEP quality assurance option year 2: End-of-year site visit report* (2017 No. 071). Alexandria, VA: Human Resources Research Organization

Sinclair, A., Deatz, R., Johnson-Fisher, J. (2015). *Findings from the PARCC quality of test administration investigations: Year 1 of operational administration* (2015 No. 046). Alexandria, VA: Human Resources Research Organization.

Sinclair, A., & Thacker, A. (2018). *South Carolina Assessment Evaluation Report #2, Part I: Technical Evaluation & Part II: Legal Evaluation.* (2018 No. 007). Alexandria, VA: Human Resources Research Organization.

Appendix A. School Visit Observation Checklist

SC EOCEP ENG1 ASSESSMENT – OBSERVATION CHECKLIST

School Name:_____ City: _____ District Name: _____

School/District Test Coordinators:_____Test Monitor Name: _____

Observer's L Name): _____ Date of Observation: _____ Assessment (subject/grade): ENG1 Writing

Test Observation – Prior to Testing			
Task #	Item	Notes	OK?
1 *p39, 59*	TA receives Testing Tickets/scratch paper (lined for ENG1 exam) from STC. (monitors-not authorized)		
2 *p23, 30, 57*	TA and/or monitor set up testing room with adequate spacing or testing screens and use their seating chart with student names.		
3 *p23*	TA or monitor covers any testing content on posters or whiteboard and posts a Do Not Disturb sign on entrance doors.		
4 *p39*	Does the TA have someone available for monitoring, troubleshooting, and answering questions? (Ask TA to identify who: the DAC, SAC, or technical staff)		

Test Observation – During Test			
Task #	Item	Notes	OK?
5 *p58*	TA asks students to turn off and turn in all electronic devices (cell phones, pagers, iPods, MP3s, PDAs, music players, etc.) until they are dismissed after the test. Observer: what about Apple watches?		
6 *p58*	Only materials authorized for use during assessments are on students' desks during the assessment.		
7 *p38,58*	TA uses the online testing roster to verify correct Test Ticket (ENG1-reading and writing exams) and track distribution of the Test Tickets.		
8 *p59*	TA does not distribute Test Tickets until prompted by the script.		
9 *p59*	Student questions regarding directions are answered before assessment begins. (what did the students ask?)		
10 *p59*	TA collects the Test Tickets after students have successfully logged on.		
11 *p26, 29, 59*	Directions are presented or read clearly, loudly, and exactly as printed in TAM. (does it appear TA ad libs at all and does the TA have the TAM?)		
12 *p18-19, 23, 30, 59, 60*	Once testing has begun, TA and monitors actively monitor students (i.e., for prohibited behaviors, item security, or cheating) by walking around the room; TA does not do other work or have conversations.		
13 *p59*	TA or monitors maintain order so one or more students do not distract others. (is it a good testing environment?)		
14 *Chk list*	Do the students use their scratch paper? Are they organizing ideas, thought maps? How many use? To what extent?		

Test Observation – During Test, cont.			
Task #	Item	Notes	OK?
15	Do students ask questions during the exam? What questions are asked? How does the TA respond?		
16 *p77*	Do students access the Writer's Checklist and TDA Scoring Guidelines while testing?		
17 *p30, 31*	TA does not leave the assessment room at any time. (Observers or monitors are not authorized to watch materials)		
18 *p60*	If a student needs to use the restroom, the TA follows school policy. There may be a group restroom break. (Note how breaks are handled, are screens covered/closed?)		
19 *p17*	Students allowed to work at own pace; allowed to finish each part of the assessment without being pressured to finish.		
20 *p18*	Students remain seated until all online assessments are exited or accommodated test materials are collected. They may read books that are not content related to the test. (note how this is handled).		
21 *P40, 41, 60, 59*	Student testing tickets and any materials used as scratch paper are collected at the end of the testing session and returned to STC.		

Additional Notes:

Appendix B. School Visit Interview Questions

SC EOCEP ENG1 ASSESSMENT – INTERVIEW QUESTIONS

School Name:_____ City: _____ District Name: _____

School/District Test Coordinators:_____Test Monitor Name: _____

Observer's L Name): _____ Date of Observation: _____ Assessment (subject/grade): ENG1 Writing

Interview with School Test Coordinator (STC) and/or Test Administrator (TA)		
General Information for ENG1		
Q #	Question	Notes
1	Did you receive the training and support materials in time to be prepared for administering the exam? Were the materials sufficient? What was missing, unclear, redundant/unnecessary? What was particularly helpful?	
2	Did you (and staff) feel sufficiently trained to successfully administer the reading and writing exam to students on testing day? What was missing or unclear? Any possible unnecessary or cumbersome tasks? Are the qualifications for Test Administrators (TAs) and monitors clear?	
3	Did you feel the training and support materials sufficient for you to complete all close-out activities such as any record retention requirements? What is missing? What is unclear? Are there unnecessary or cumbersome tasks?	
4	Do you feel the TA and Monitor training to be sufficient? How were they trained (e.g., in-person, online DRC system training modules)? Were there any practice opportunities for logging students on or troubleshooting?	

Interview with School Test Coordinator (STC) and/or Test Administrator (TA)		
General Information for ENG1		
Q #	Question	Notes
5 *p28*	Do all students receive Online Tools Training and tutorials prior to testing? (how is that tracked?) Reference is made in the TA script to students' access to the Writer's Checklist and TDA Scoring Guidelines on the DRC system. Are students taught to use these tools for ENG1 and throughout the year? Have TAs or monitors provided any student with help reading the guidelines or checklist while testing?	
6	Even though the exams are not timed, do students complete the exams in a reasonable time? (no estimates yet in TAM). Any difference between reading and writing)? Any impact with having the TDA item? Have you had students use the full day, how does that work? (when do other students get released)	
7 *p30, 57*	Are seating charts created and used for each test session? (how created? Are they retained?) Are they useful? Students with the same test forms are not seated together?	
8 *p10, 37, 38*	Are test materials (e.g., printed Testing Tickets) secured in locked storage area with limited access? Follow-up: who has key besides STC; custodians or principal? Who prints them and when? How distributed and returned on testing day? When are they destroyed?	
9	Do the TA scripts work well? Is revision needed and in what way? Is there sufficient guidance as to any possible variations in test administration (1st test session vs. 2nd or 3rd)? If not clear, in what way?	
10 *p5*	Are *Test Security Violation (TSV) Action Forms* completed by DTC or STC? How do TAs report a violation (i.e., immediately)? Have you had one? What was the violation?	

Interview with School Test Coordinator (STC) and/or Test Administrator (TA)		
General Information for ENG1		
Q #	Question	Notes
11 *Apx c & d*	For those students with IEPs or designated EL who take the regular ENG1 assessment with accommodations, do the training and support materials adequately prepare staff to administer the exam? Do all students receive Online Tools training and tutorials prior to testing? Is their training modified in any way to prepare them for the exam? Any differences between these accommodations and ones students receive during instruction? Any other concerns?	
12	Is there anything else you would like to share about the English 1 assessment, encompassing both the Reading and Writing components?	