Equating Student Scores Across Ohio's State Test Administration Modes Fall 2019 Grade 3 ELA

Introduction

Senate Bill 216, 132nd General Assembly, allows districts the option of paper or online test administration for the grade 3 Ohio's State Test (OST), beginning in the 2019-2020 school year. For the third-grade English Language Arts (ELA) test administered in fall 2019, about 98,400 students (77.1%) took the test online, and 29,176 (22.9%) students took the test on paper. Prior to reporting test scores, an equating study was performed to evaluate differences in item and test performance between the online and paper test administrations and to identify the linking constants that may be needed to place item parameter estimates across modes on a common scale for test scoring and reporting.

A single, fixed operational test form that included 26 reading items and one writing prompt was used to administer grade 3 OST fall 2019 assessment online. In addition, an equivalent test form was constructed for the paper-based test administration. The paper form was designed to be as similar as possible to the online form, differing only in how student responses are captured. Thus, all operational items, including the three writing dimension scores, are considered common between the online and paper test administration modes. These common items between the online and paper forms provide the basis for an equating study to examine the performance of items between the online and paper modes of test administration.

A matched samples design (Way, Davis, and Fitzpatrick, 2006) was used to equate ability estimates between the online and paper test administrations. A covariate regression approach was implemented to construct equivalent groups of students taking the grade 3 OST ELA assessment in both modes of test administration.

Matched Samples

Typically, the regression analysis would identify for each student a predicted score on the online OST assessment from previous achievement on the online administered Ohio State Test, covarying individual and school-level demographic variables in the development of the prediction equation. This predicted raw score distribution would be used to identify two matched samples for each assessment to conduct the mode equating study. For grade 3, however, there is no prior OST achievement score. Although one could identify matched samples based only on demographic variables, demographic variables are weak predictors of student achievement and would not result in samples sufficiently matched on achievement to reliably detect mode differences. Therefore, we used student performance-level classifications based on the grade 2 reading diagnostic assessment, as well as student test scores on the Kindergarten Readiness Assessment (KRA) to predict student performance in grade 3. Although the performance classification for the grade 2 diagnostic assessment is dichotomous, limiting the predictive utility of the assessment result, it does provide an index of reading achievement that is likely correlated with the grade 3 OST ELA assessment. Likewise, although the KRA assesses prereading skills several years prior to

the grade 3 OST assessment, prereading skills assessed in the KRA are likely correlated with subsequent OST ELA achievement scores. Because the grade 2 reading diagnostic assessments are administered on paper, we built the regression using students participating in the fall 2019 paper test administration so that mode of test administration was constant in the prediction of grade 3 ELA achievement. In addition, the grade 2 attendance rates for students participating in the grade 3 OST test administrations as well as Ohio district typology classification for the district in which each student was enrolled were also included in the regression model.

Procedure

The following procedures were used to define the matched samples between the online and paper test administration modes.

For students participating in the fall 2019 paper test administration, fall 2019 OST raw scores were regressed on their grade 2 reading diagnostic performance-level classifications, the kindergarten readiness assessment scores, and the individual- and school-level demographic variables. The individual demographic variables include ethnicity, gender, English learner (EL) status, special education (SPED) status, and students' grade 2 attendance rate. School-level variables include the ratio of African American students, ratio of Hispanic students, ratio of multiethnic students, ratio of ELs, ratio of SPED, school average achievement as indexed by the previous year grade 3 OST scores and the Ohio district typology code.

All variables were entered into the regression equation simultaneously in the prediction of grade 3 OST test performance:

$$\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n$$

where $\hat{Y}$ is the predicted OST raw score on the current assessment, β_n refers to the estimated regression weight for covariate X_n. The bidirectional stepwise selection algorithm was used to identify the final predictors. The adjusted R-square for the final best model is 0.42. The estimated regression weights are provided in the Appendix A.

With the obtained regression weights, the prediction equation was applied to all students participating in OST across test administration modes, yielding a predicted OST raw score for each student. The scatter plots of observed and predicted raw scores for both online and paper samples are provided in Appendix B.

Using the predicted raw score distribution, the paper sample, which has the smaller number of students, was divided in 20 equal-sized groups. Each group included 911 students. The predicted raw score distribution cut points determined by the equal-sized groups were used to divide students in the larger online test sample into each of the 20 ability-level groups. The 19 cut points are listed in Appendix C. Within each of the 20 ability groups in the larger sample, a random sample of students was drawn, equal in size to the number of paper test students in each of the predicted ability-level groups. The table in Appendix D provides a comparison of the demographic and achievement characteristics between the one of matched online and paper samples drawn for the mode equating study. The table presents the proportion of students classified in each demographic category, the mean and standard deviation of test

score on the previous assessment, as well as the average predicted raw score on grade 3 OST fall 2019 assessment. The comparison indicates that the demographic composition and prior achievement of the matched samples is quite similar and that the matching procedure was effective.

Analyses

Item Response Theory (IRT) parameter estimates were calibrated independently for the matched online and paper test administration mode samples, centering ability to zero for each of the matched samples. Because the samples were matched on ability, the resulting item parameter estimates can be used to identify differential item difficulty between test administration modes. In addition, obtained parameter estimates for the common items can be used to identify the linking constant necessary to bring the matched sample paper item parameters onto the matched sample online scale. Mean linking was taken as the difference between the average item difficulty estimates from the matched sample paper calibration and the average item difficulty estimates from the matched sample online item parameter estimates.

Item calibrations were performed in *Winsteps®*. The difference between the identified group means provides the linking constant necessary to place paper test scores onto the OST online scale. The online scale is the reference scale for the OSTs, since estimation of all IRT parameters, adoption of performance standards, and all previous reporting have been conducted with respect to the item parameters estimated from online test administrations.

The standard error of the group means provides an index of the reliability of the difference between the group means. ODE determined that any mean differences beyond two standard errors of the mean would be considered significant and require application of a mode linking constant to place the paper assessments onto the OST online scale. The bootstrap method was used to produce the standard error of the mode difference constant.

Results

The sample size for the paper sample and each of the 100 online matching samples is 18,220. The average item difficulty across the 100 online calibrations was taken as the average online difficulty. The difference between the average online difficulty and the average paper difficulty is the mode linking constant. The standard deviation of the average difficulties across the 100 online samples represents the conditional bootstrap standard error of the linking constant. Table 1 shows the average Rasch difficulty of the matched online and paper samples, the mode linking constant and the standard error of mode linking constant. All the values are on the theta metric.

Table 1: Linking Constant Resulting from the Matched Samples Equating

Test	Mean Item Difficulties		Mode Linking Constant	Standard Error of Mode Linking Constant
	Online	Paper		
G3 ELA	0.345	0.206	0.139	0.006

As noted above, ODE determined that any mean difference beyond two standard errors of the mean would be considered significant and would require application of the mode linking constant to place the paper test results on the OST online scale.

The mode linking constant is applied to the ability estimates for students participating in paper test administrations. The linking constant is applied to all subject area and subscale scores. The linking constant is applied to paper ability estimates in the underlying theta scale, prior to transformation of ability estimates to the appropriate scale score. For grade 3 ELA, this includes both the overall ELA scale score and the Reading scale score used to support Ohio's Third Grade Reading Guarantee. In addition, the linking constant is applied to subscale ability estimates prior to computing subscale performance-level classifications. Table 2 presents the raw to scale score conversion table for the online and adjusted paper scale. Table 3 presents the reading guarantee raw to scale score conversion table for the online and adjusted paper scale.

Table 2. Raw to Scale Score Conversion Table for Online and Adjusted Paper Scale

Raw Score	Online Scale		Adjusted Paper Scale	
	Scale Score	Performance Level	Scale Score	Performance Level
0	545	Limited	545	Limited
1	545	Limited	545	Limited
2	570	Limited	564	Limited
3	591	Limited	585	Limited
4	606	Limited	600	Limited
5	619	Limited	612	Limited
6	629	Limited	623	Limited
7	638	Limited	632	Limited
8	646	Limited	640	Limited
9	654	Limited	648	Limited
10	661	Limited	655	Limited
11	667	Limited	661	Limited
12	673	Basic	667	Limited
13	679	Basic	673	Basic
14	685	Basic	678	Basic
15	690	Basic	684	Basic
16	695	Basic	689	Basic
17	700	Proficient	694	Basic
18	706	Proficient	700	Proficient
19	711	Proficient	704	Proficient
20	716	Proficient	709	Proficient
21	721	Proficient	714	Proficient
22	726	Accelerated	719	Proficient

Raw Score	Online Scale		Adjusted Paper Scale	
	Scale Score	Performance Level	Scale Score	Performance Level
23	731	Accelerated	725	Accelerated
24	736	Accelerated	730	Accelerated
25	742	Accelerated	736	Accelerated
26	748	Accelerated	742	Accelerated
27	754	Advanced	748	Accelerated
28	761	Advanced	754	Advanced
29	768	Advanced	761	Advanced
30	775	Advanced	769	Advanced
31	783	Advanced	777	Advanced
32	792	Advanced	786	Advanced
33	802	Advanced	795	Advanced
34	813	Advanced	806	Advanced
35	825	Advanced	818	Advanced
36	838	Advanced	832	Advanced
37	854	Advanced	848	Advanced
38	863	Advanced	863	Advanced
39	863	Advanced	863	Advanced
40	863	Advanced	863	Advanced
41	863	Advanced	863	Advanced

Table 3. Third Grade Reading Guarantee Raw to Scale Score Conversion Table for Online and Adjusted Paper Scale

Reading Raw Score	Online Scale		Adjusted Paper Scale	
	Reading Scale Score	Reading Guarantee	Reading Scale Score	Reading Guarantee
0	16	No	16	No
1	17	No	16	No
2	24	No	23	No
3	29	No	27	No
4	32	No	31	No
5	35	No	34	No
6	37	No	36	No
7	39	No	38	No
8	41	No	40	No
9	43	No	41	No
10	44	No	43	No
11	46	Yes	44	No
12	47	Yes	46	Yes
13	49	Yes	47	Yes
14	50	Yes	49	Yes
15	51	Yes	50	Yes
16	52	Yes	51	Yes
17	54	Yes	52	Yes
18	55	Yes	54	Yes
19	56	Yes	55	Yes
20	58	Yes	56	Yes
21	59	Yes	58	Yes
22	61	Yes	59	Yes
23	63	Yes	61	Yes
24	64	Yes	63	Yes
25	67	Yes	65	Yes
26	69	Yes	68	Yes
27	72	Yes	70	Yes
28	75	Yes	74	Yes
29	80	Yes	79	Yes
30	86	Yes	86	Yes
31	86	Yes	86	Yes

Table 4. Raw to Scale Score Conversion Table for Online and Adjusted Paper Scale: Reading Informational Text Reporting Category

Raw Score	Online Scale		Adjusted Paper Scale	
	Scale Score	Performance Level	Scale Score	Performance Level
0	545	Below Mastery	545	Below Mastery
1	577	Below Mastery	571	Below Mastery
2	613	Below Mastery	607	Below Mastery
3	637	Below Mastery	630	Below Mastery
4	655	Below Mastery	649	Below Mastery
5	671	Below Mastery	664	Below Mastery
6	685	At/Near Mastery	679	At/Near Mastery
7	699	At/Near Mastery	692	At/Near Mastery
8	712	At/Near Mastery	706	At/Near Mastery
9	726	Above Mastery	719	At/Near Mastery
10	740	Above Mastery	734	Above Mastery
11	755	Above Mastery	749	Above Mastery
12	773	Above Mastery	766	Above Mastery
13	793	Above Mastery	787	Above Mastery
14	819	Above Mastery	813	Above Mastery
15	859	Above Mastery	853	Above Mastery
16	863	Above Mastery	863	Above Mastery

Table 5. Raw to Scale Score Conversion Table for Online and Adjusted Paper Scale: Reading Literary Text Reporting Category

Raw Score	Online Scale		Adjusted Paper Scale	
	Scale Score	Performance Level	Scale Score	Performance Level
0	551	Below Mastery	545	Below Mastery
1	586	Below Mastery	579	Below Mastery
2	623	Below Mastery	617	Below Mastery
3	646	Below Mastery	640	Below Mastery
4	663	Below Mastery	657	Below Mastery
5	677	Below Mastery	671	Below Mastery
6	689	At/Near Mastery	683	At/Near Mastery
7	700	At/Near Mastery	694	At/Near Mastery
8	711	At/Near Mastery	705	At/Near Mastery
9	721	At/Near Mastery	715	At/Near Mastery
10	733	Above Mastery	726	Above Mastery
11	745	Above Mastery	739	Above Mastery
12	760	Above Mastery	754	Above Mastery
13	780	Above Mastery	773	Above Mastery
14	812	Above Mastery	806	Above Mastery
15	844	Above Mastery	838	Above Mastery

Table 6. Raw to Scale Score Conversion Table for Online and Adjusted Paper Scale: Writing Reporting Category

Raw Score	Online Scale		Adjusted Paper Scale	
	Scale Score	Performance Level	Scale Score	Performance Level
0	583	Below Mastery	576	Below Mastery
1	621	Below Mastery	615	Below Mastery
2	669	At/Near Mastery	663	At/Near Mastery
3	706	At/Near Mastery	699	At/Near Mastery
4	739	At/Near Mastery	733	At/Near Mastery
5	772	Above Mastery	766	Above Mastery
6	805	Above Mastery	799	Above Mastery
7	838	Above Mastery	831	Above Mastery
8	863	Above Mastery	863	Above Mastery
9	863	Above Mastery	863	Above Mastery
10	863	Above Mastery	863	Above Mastery

Reference

Way, W. D., Davis, L. L., & Fitzpatrick, S. (2006, April). Score comparability of online and paper administrations of the Texas Assessment of Knowledge and Skills. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

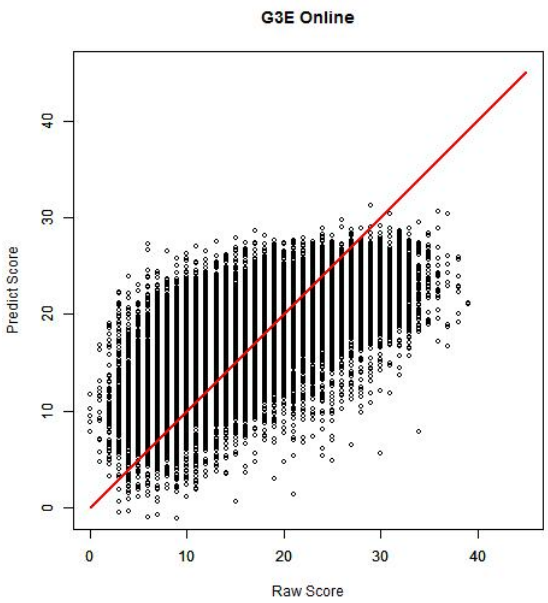
Appendix A – Regression Models Used to Produce the Predicted Scores

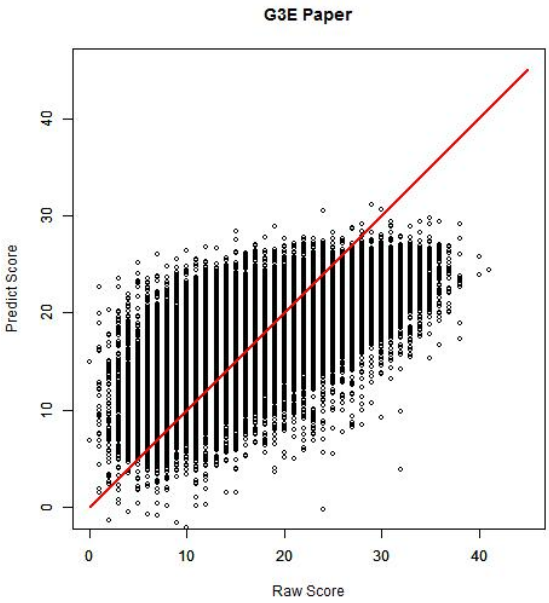
Test	Predictor	Regression Coefficients	Significance (p-value)
Grade 3 ELA (Adjusted R-Squared=0.42)	read_no	-5.75	0.000
	k_score	0.17	0.000
	school_ss_mean	0.07	0.000
	SPED	-1.95	0.000
	AfriAm	-2.39	0.000
	Gender	0.76	0.000
	Asian	2.29	0.000
	attd_rate	5.59	0.000
	AfriAm_ratio	2.16	0.000
	Multi	-0.71	0.000
	ELL_ratio	4.43	0.000
	ELL	-1.57	0.000
	read_expt	-4.72	0.003
	typo_2	-1.16	0.000
	SPED_ratio	1.46	0.006
	Hisp_ratio	-1.74	0.062
	typo_8	-1.10	0.000
	typo_1	-0.82	0.000
	typo_4	-0.74	0.000
	typo_7	-0.82	0.000
	typo_0	-1.08	0.000
	typo_6	-0.39	0.030

Note: The table below provides a description of the variables in the regression models.

Variable	Description
read_no	Reading diagnostic dummy code group: "NO" (Assessed, not on track)
k_score	Kindergarten Readiness Assessment score
school_ss_mean	School average achievement as indexed by the Fall 2018 grade 3 ELA OST scale scores
SPED	Special education (SPED) status
AfriAm	Ethnicity dummy code group: African American
Gender	Gender (with the male as the reference group)
Asian	Ethnicity dummy code group: Asian
attd_rate	Student attendance rate
AfriAm_ratio	Ratio of African American students at school
Multi	Ethnicity dummy code group: Multiracial
ELL_ratio	Ratio of English Language Learner (ELL) students at school
ELL	English Language Learner (ELL) status
read_expt	Reading diagnostic dummy code group: "EX" (Exempt from Diagnostic Assessment)
typo_2	District typology dummy code group: 2 (Rural - Average Student Poverty & Very Small Student Population)
SPED_ratio	Ratio of special education (SPED) status students at school
Hisp_ratio	Ratio of Hispanic students at school
typo_8	District typology dummy code group: 8 (Urban (Ohio8) - Very High Student Poverty & Very Large Student Population)
typo_1	District typology dummy code group: 1 (Rural - High Student Poverty & Small Student Population)
typo_4	District typology dummy code group: 4 (Small Town -High Student Poverty & Average Student Population Size)
typo_7	District typology dummy code group: 7 (Urban - High Student Poverty & Average Student Population)
typo_0	District typology dummy code group: 0 (Community School)
typo_6	District typology dummy code group: 6 (Rural - High Student Poverty & Small Student Population)

Appendix B – Scatter Plots of Observed and Predicted Raw Scores: Fall 2019





Appendix C –Cut Scores Used to Divide the Paper Sample into 20 Ability Groups

Cut	Predicted Paper Raw Score
1	8.073551
2	9.857688
3	11.26508
4	12.40247
5	13.54734
6	14.76611
7	15.91637
8	16.85645
9	17.64478
10	18.33454
11	18.96879
12	19.60224
13	20.14701
14	20.69854
15	21.26383
16	21.86833
17	22.51944
18	23.32191
19	24.49366

Appendix D – Comparison of Matched Samples of Grade 3 ELA: Fall 2019

Demographic and Achievement Variables	Online Sample	Paper Sample
Male	0.52	0.50
Female	0.48	0.50
African American	0.14	0.15
Asian	0.02	0.01
American Indian	0.00	0.00
Hispanic	0.04	0.03
Pacific/Hawaiian Islander	0.00	0.00
Multiple Ethnicities	0.08	0.08
Limited English Proficiency (LEP)	0.03	0.02
Individualized Education Program (IEP)	0.13	0.12
Kindergarten Readiness Assessment Score Mean	266.81	268.14
Average School Fall 2018 Score Mean	688.49	686.72
Average Student Attendance Rate	0.95	0.95
Predicted Fall 2019 Score Mean	17.37	17.37
Predicted Fall 2019 Score Standard Deviation	5.15	5.13
Predicted Fall 2019 Score Minimum	-3.68	-2.07
Predicted Fall 2019 Score Maximum	31.23	31.07
Predicted Fall 2019 Score Skewness	-0.49	-0.49
Predicted Fall 2019 Score Kurtosis	2.58	2.56