Florida
Standards Assessments

Florida Standards Assessments

2018–2019

Volume 7
Special Studies

FLORIDA DEPARTMENT OF EDUCATION
fldoe.org

TABLE OF CONTENTS

This volume consists of a series of independent special studies conducted by AIR.

Florida
Standards Assessments

Trend analysis of Scores on FSA

2019 Paper Tests

October 2019

FLORIDA DEPARTMENT OF EDUCATION
fldoe.org

CONTENTS

FIGURES

TABLES

TREND ANALYSIS OF SCORES ON FSA 2019 PAPER TESTS

Florida *House Bill 7069* required a transition from a census computer-based test to a paper-pencil format in grades 4–6 English language arts (ELA) and grades 3–6 mathematics beginning in spring 2019. AIR and the Florida Department of Education (FDOE) developed procedures to minimize any potential mode effect beginning with item development strategies and test form construction methods. This document summarizes the score trends observed as a component of our continued monitoring of the change in mode.

STUDY BACKGROUND

This analysis uses historical data prior to the change in mode as well as observed test scores collected after the change in mode to evaluate the degree to which aggregate patterns of score trends remain stable or change over time and across modes. We have no formal control group and instead have only observed test scores from students under different testing conditions and modes. However, monitoring trends in test scores before and after the change in mode can illustrate any possible effects that may be related to the differences in test administration.

Specifically, we can consider this a quasi-experimental design where trends in student scores prior to the change in mode serve as the control and trends in scores affected by the mode change can be compared to the baseline data from the control. This most closely resembles an interrupted time series design where trend data are available prior to any change in mode and then data are available for a group affected by the mode.

The analyses presented here are descriptive only, and inferences can be made in a limited way given that a formal experimental design is not used. Our primary criterion for understanding impact is whether score trends appear to be similar to or different than historical score trends. That is, if score trends for students affected by the change in mode differ markedly from score trends for students unaffected by the change, then we might have reason to further consider mode as a possible factor. On the other hand, if score trends are comparable to historical trends, then we may have reason to believe change in mode had minimal impact, if any.

For this analysis, we form the following research question and evaluate this question using observed score correlations and longitudinal trends in student outcomes:

- Are trends in aggregate student performance in 2019 comparable to trends in student performance prior to 2019?

METHODS

In this study, we investigated the possible impact of the change in testing mode by using the following three key indicators:

1. Within-grade score distributions should be consistent with historical trends observed.
2. Cross-grade (quasi-longitudinal) changes in performance level should be consistent with historical trends.

3. Cross-year correlations in scores should be comparable to historical cross-year correlations.

We analyzed the historical trends in scale score distributions and proficiency rates[1] by following two different cohorts across years. The spring 2019 cohort referred to the cohort of students who were in the grade of interest in spring 2019, and the spring 2018 cohort was the cohort of students who were in the same grade in spring 2018. These two cohorts were chosen because the former had experienced the testing mode change, whereas the latter did not. Therefore, the spring 2018 cohort was the control group and served as a baseline measure to compare with the spring 2019 cohort. If there were systematic differences in the historical trends for these cohorts, then this would suggest that students' scores and proficiency classifications are affected by the mode difference. On the other hand, if the trends were similar, this would show that the scores and classifications between different testing modes are comparable.

In addtion to score distributions and proficiency classifications, we compared the correlations of students' performance from multiple cohorts of examinees. The spring 2019 cohort was the focal group, which was the group of students who experienced the testing mode change from spring 2018 to spring 2019. The cohorts from earlier years (i.e., spring 2018, spring 2017, and spring 2016) were the control groups, who experienced the same testing mode between two consecutive years. The correlations of scale scores between the two consecutive years of test administrations for each group were computed and compared to see if they were considerably different. Our hypothesis was that if there was no mode effect, the correlations from these two groups would be very similar for any given two consecutive years.

The data used in this study come from the State Student Results (SSR) files for the population from spring 2019. The data for prior years were also used for this same group of students. The scale scores from FSA tests were used to compute the Pearson correlations.

[1] Proficiency rate is defined as the percentage of students who are classified as Level 3 (Satisfactory) and above.

RESULTS

The data used for this study are summarized in Table 1. This table provides the descriptive summary of scale scores for students with reported score status.

Table 1: Descriptive Summary of Scale Scores in ELA and Mathematics

Administration	Subject	Grade	N	Mean SS	SD SS
SP19	ELA	6	211,645	326.15	24.05
		5	218,868	322.04	23.27
		4	211,410	313.40	20.88
	Math	6	202,154	325.06	24.24
		5	219,274	324.15	25.43
		4	210,445	316.10	23.77
SP18	ELA	6	211,279	324.86	24.70
		5	211,086	321.89	23.19
		4	215,827	312.11	20.68
	Math	6	203,162	323.68	24.38
		5	213,499	324.46	24.39
		4	217,434	314.79	23.46

SCALE SCORE DISTRIBUTIONS ACROSS YEARS

Figures 1 and 2 provide the historical trends of scale scores in ELA and mathematics, respectively. We considered the variablity of the scale scores and compared two cohorts, one of which (spring 2019) experienced the testing mode change whereas the other (spring 2018) did not experience the change in testing mode. Within the panel for each cohort, score distributions were plotted for the same cohort of students from earlier grade levels to show historical trends.

Figures 1 and 2 show that the two cohorts were comparable in the historical trends of scale score distributions. If the trend of scale score distribution for the spring 2019 cohort were very different from that for the spring 2018 cohort we could suspect the existence of mode effect.

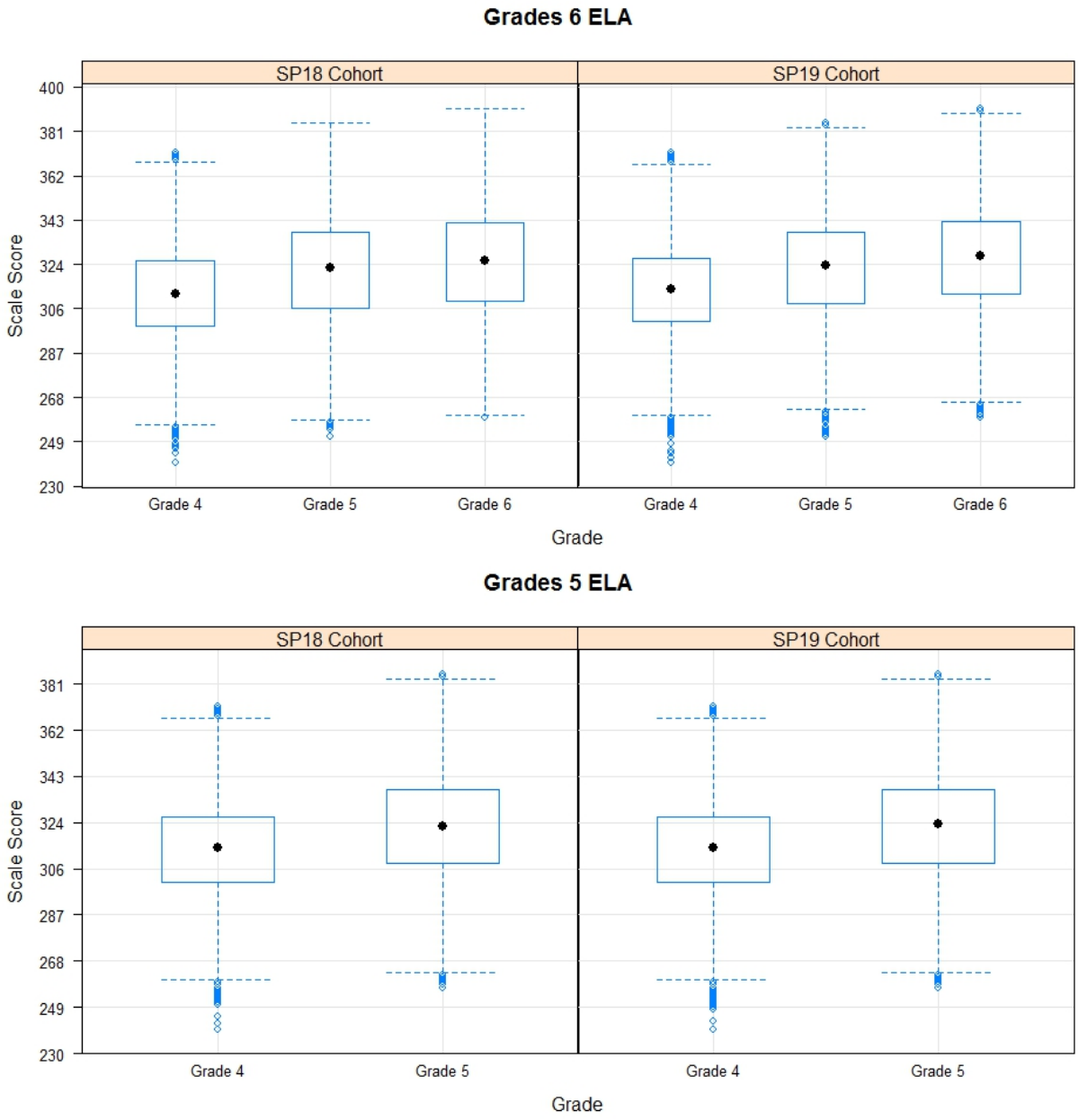

Figure 1: The Scale Score Distribution for Grades 5 and 6 ELA

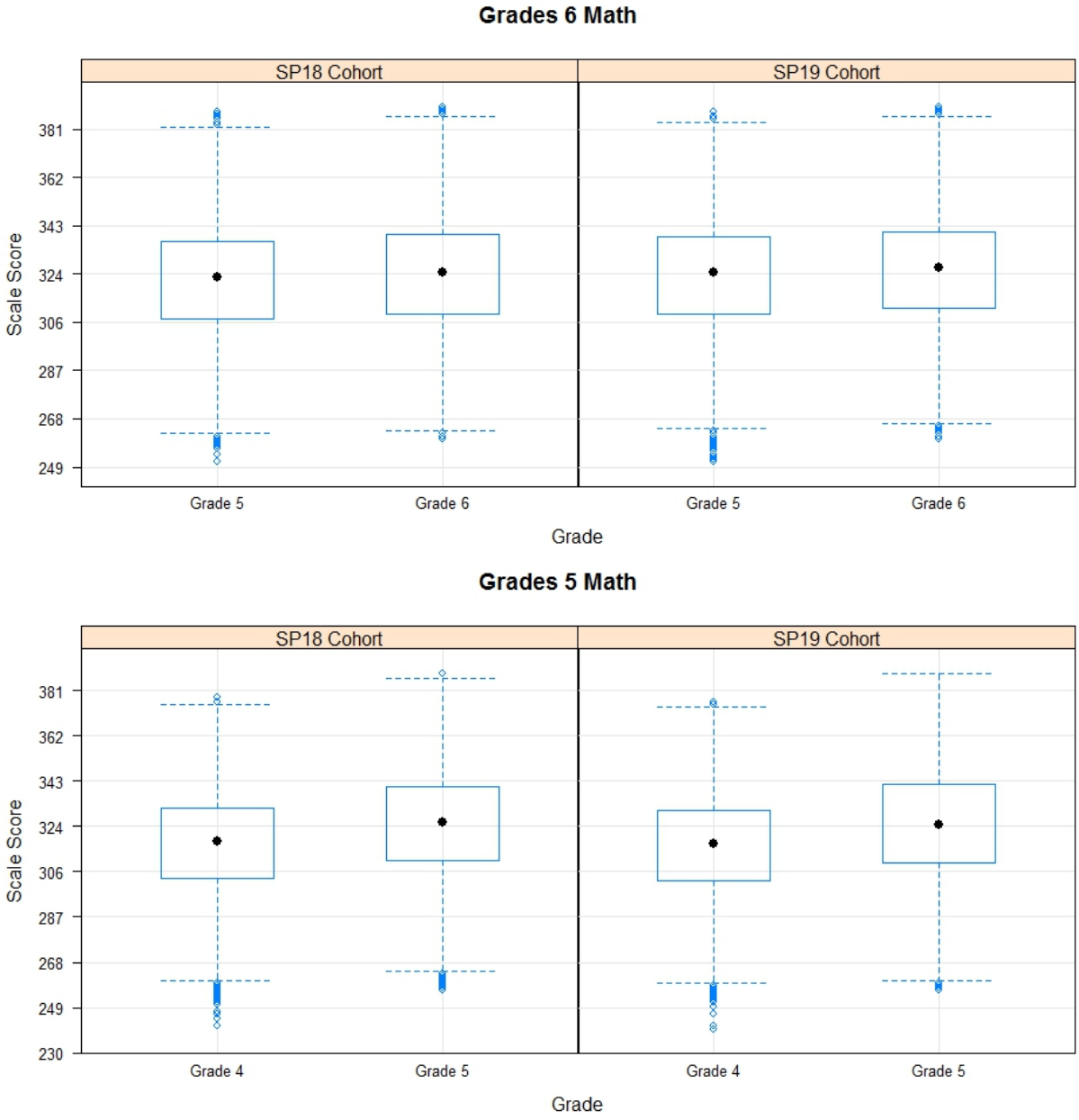

Figure 2: The Scale Score Distribution for Grades 5 and 6 Mathematics

PROFICIENCY RATES ACROSS YEARS

Figures 3 and 4 depict the historical trends of proficiency rates in ELA and mathematics, respectively. Table 2 provides such data in a tabular format. As with the scale score distribution analysis, we compared two cohorts (the spring 2019 cohort and the spring 2018 cohort) and, for each cohort, proficency rates were plotted for the same cohort of students from earlier grade levels to show historical trends.

Figures 3 and 4 show that the two cohorts were, in general, comparable in the historical trends of proficiency rates, where small decreases in proficiency rates were found as grade levels increased. There was only one anomaly with the grade 6 ELA spring 2018 cohort that saw a slight increase in percent proficient for grade 5 relative to grade 4. However, this small anomaly was not associated with any change in testing mode; both grade 4 and grade 5 for this cohort were tested online. That said, as far as the transition from online to paper in spring 2019 was concerned, the two cohorts displayed comparable trends. If the trend of proficiency rates for the spring 2019 cohort was very different from that for the spring 2018 cohort, we could suspect the existence of mode effect.

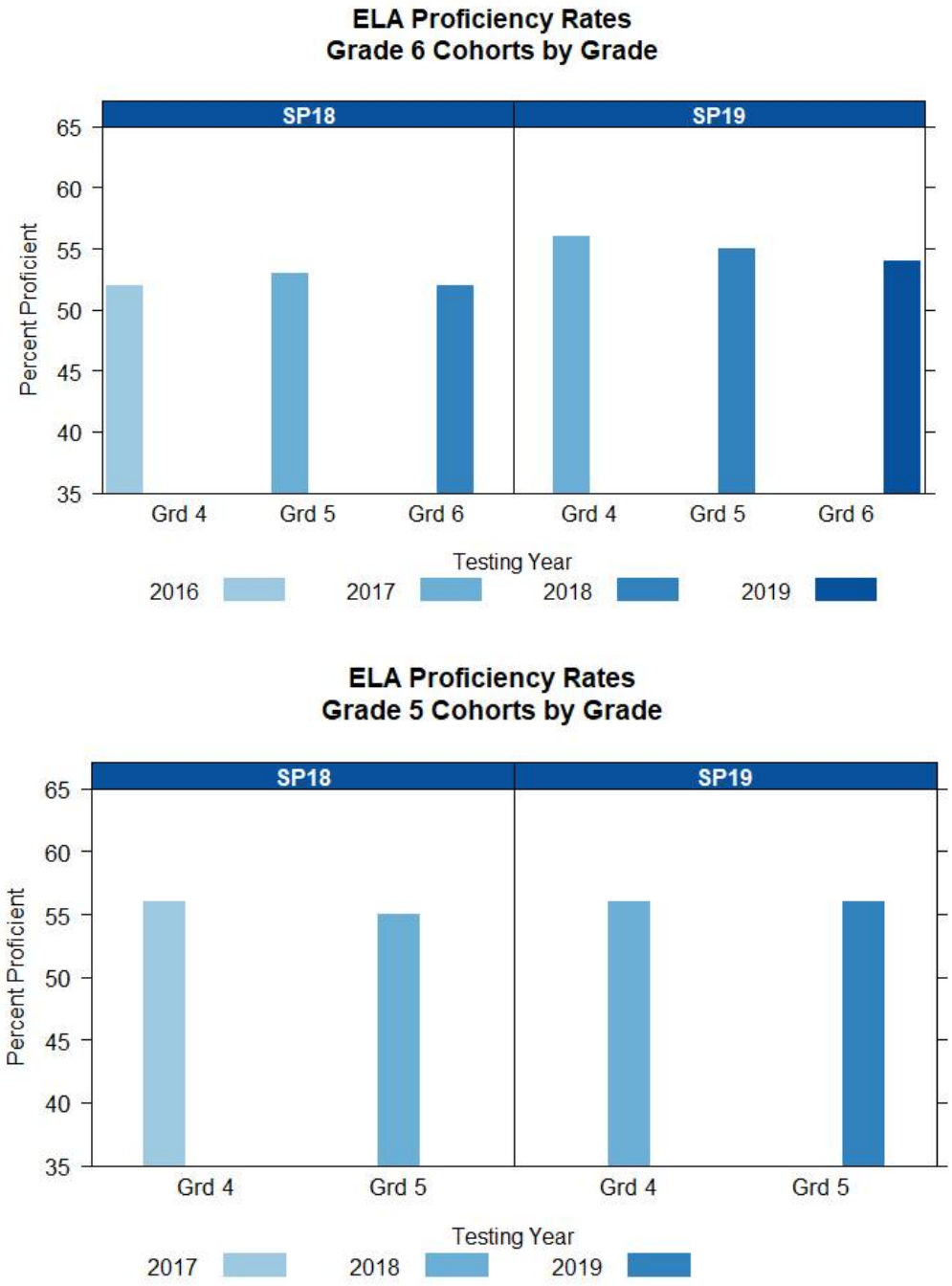

Figure 3: Proficiency Rates for Grades 5 and 6 ELA

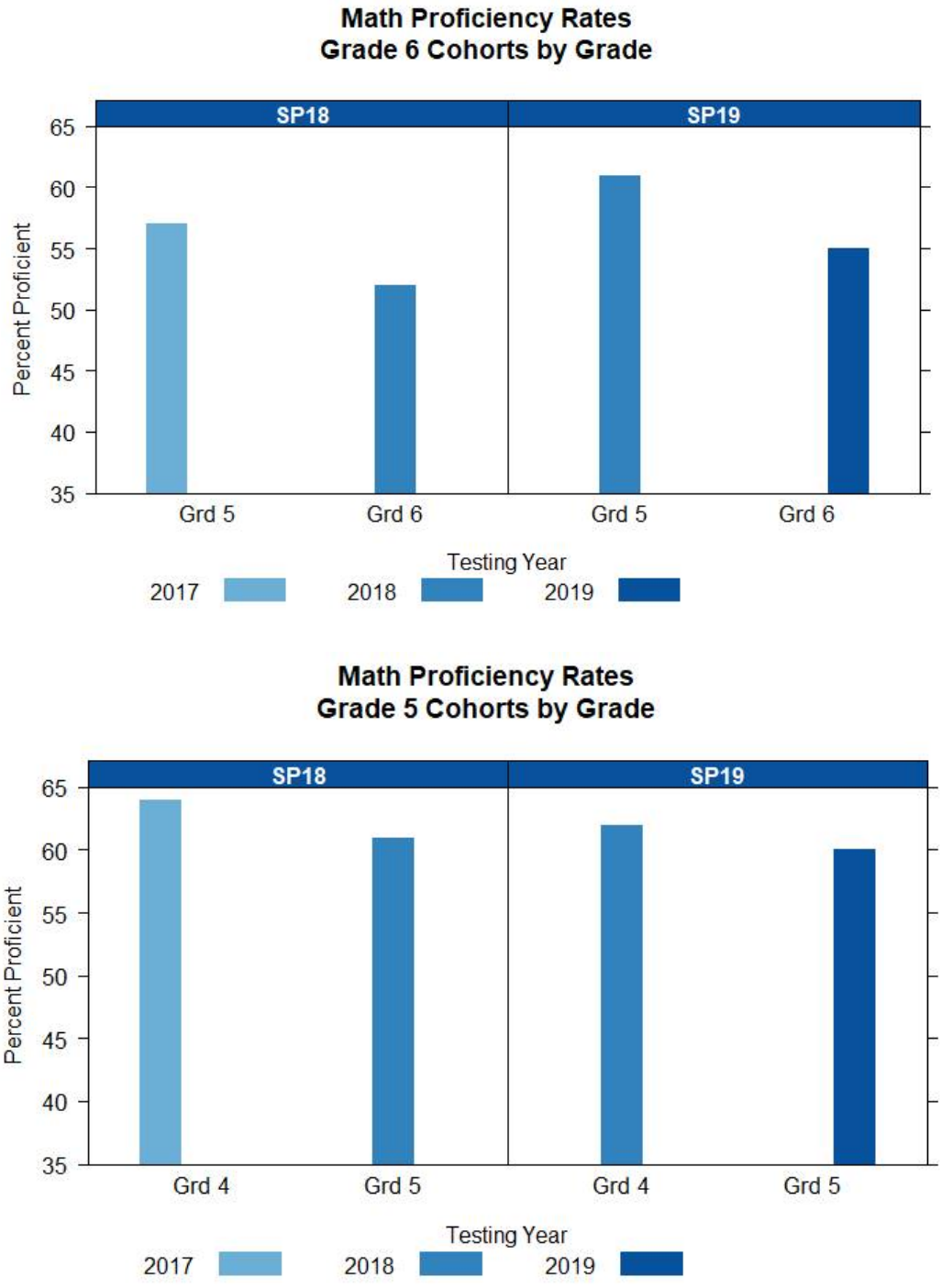

Figure 4: Proficiency Rates for Grades 5 and 6 Mathematics

Table 2: Proficiency Rates for ELA and Mathematics

Subject	Grade	2015	2016	2017	2018	2019
ELA	G4	54	52	56	56	58
	G5	52	52	53	55	56
	G6	51	52	52	52	54
Math	G4	59	59	64	62	64
	G5	55	55	57	61	60
	G6	50	50	51	52	55

Note: School year 2018–2019, in which tests changed from online to paper, is highlighted in yellow.

CROSS-YEAR CORRELATIONS

Figures 5 and 6 compare cross-sectional correlations of scale scores between two consecutive years for the spring 2019 cohort and earlier cohorts. Table 3 provides these data in a tabular format. The correlations of scale scores for the spring 2019 cohort (the group who experienced mode change) were compared to those for the prior cohorts (the control groups who did not experience testing mode change). We used as many control groups as data allowed to provide more reliable baseline measures. For example, for both grade 6 ELA and grade 6 mathematics, we used grade 6 from spring 2018, spring 2017, and spring 2016 to serve as control groups for grade 6 from spring 2019. For some other tests, however, we did not include all the years because there was a change in mode involved in the year of interest. For example, grade 5 mathematics did not include cohorts from spring 2017 or spring 2016, because both cohorts experienced a mode change (paper to online) moving from grade 4 to grade 5.

The results from cross-sectional analyses showed that the correlations between scale scores in two consecutive years were very similar for the spring 2019 cohorts and earlier cohorts. If we had observed very different correlations between these two cohorts, we would have suspected the possibility of mode effect on students' performance.

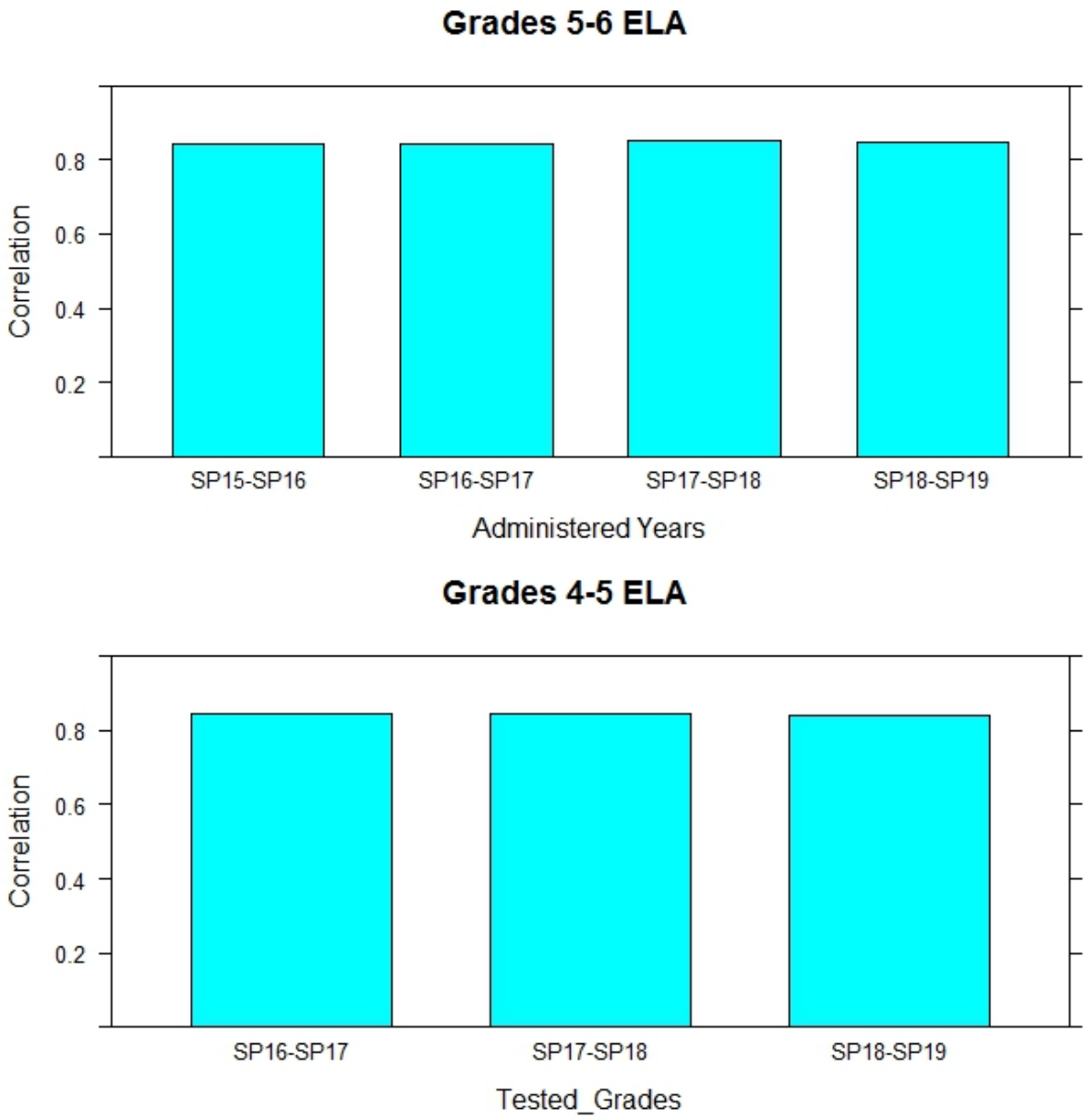

Figure 5: Cross-Sectional Correlations in ELA

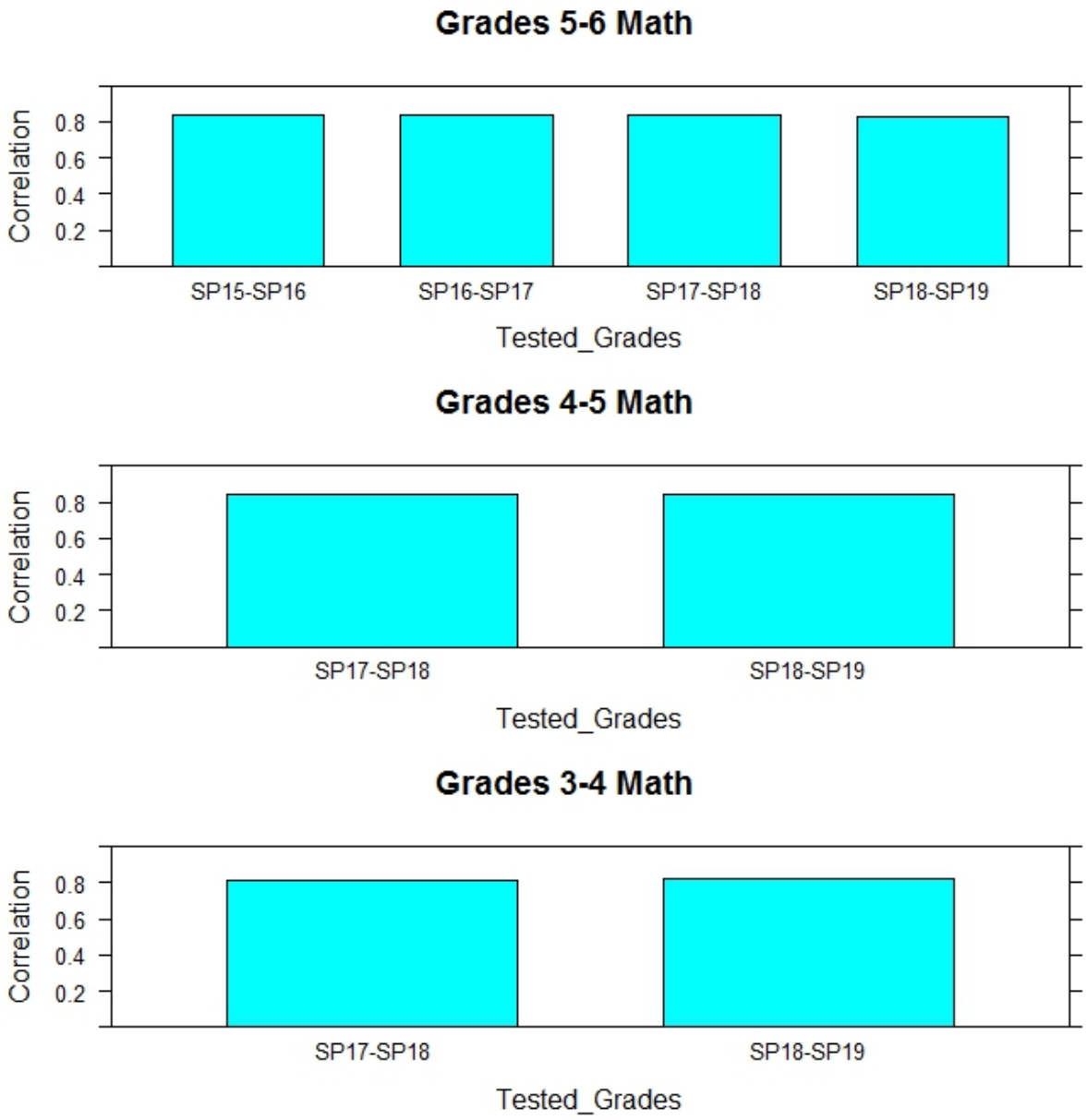

Figure 6: Cross-Sectional Correlations in Mathematics

Table 3: The Cross-Sectional Correlations in ELA and Mathematics

Subject	Tested Grades	Administered Year	Correlation	Mode	Mode Change
ELA	4–5	SP16–SP17	0.846	Online-Online	No
	4–5	SP17–SP18	0.843	Online-Online	No
	4–5	SP18–SP19	0.838	Online-Paper	Yes
	5–6	SP15–SP16	0.842	Online-Online	No
	5–6	SP16–SP17	0.843	Online-Online	No
	5–6	SP17–SP18	0.853	Online-Online	No
	5–6	SP18–SP19	0.846	Online-Paper	Yes
Math	3–4	SP17–SP18	0.816	Online-Online	No
	3–4	SP18–SP19	0.820	Online-Paper	Yes
	4–5	SP17–SP18	0.847	Online-Online	No
	4–5	SP18–SP19	0.845	Online-Paper	Yes
	5–6	SP15–SP16	0.840	Online-Online	No
	5–6	SP16–SP17	0.833	Online-Online	No
	5–6	SP17–SP18	0.834	Online-Online	No
	5–6	SP18–SP19	0.830	Online-Paper	Yes

SUMMARY

Data suggest that score trends observed for student groups after the mode change are similar to the score trends observed prior to the mode change. Rather, the collection of data holistically forms a picture suggesting that student test scores and the general patterns observed historically are intact even after the FSA was administered on paper for some groups of students.

Florida
Standards Assessments

Device Comparability

October 2018

Florida Department of Education
fldoe.org

CONTENTS

TABLES

Device Comparability

The Florida Department of Education (FDOE) submitted evidence related to the required Critical Elements necessary for the Federal peer review process. The U.S. Department of Education (USDOE) responded with requests for some additional information and evidence to more fully satisfy requirements needed on a subset of the critical elements. This document provides additional evidence related to Critical Element 4.6. The USDOE requested "Evidence of the comparability of the FSA tests across the most frequently used platforms (e.g., computers, tablets) for at least one grade level test."

The American Institutes for Research (AIR) serves as the test vendor for the Florida Standards Assessment (FSA) and maintains an indicator in the data showing the type of device used by a student when participating in the FSA. This indicator, along with the observed test scores, can be used to assess the degree to which FSA scores for students on different devices are comparable.

Study Background

The question of score comparability across different devices can be examined to assess whether student performance on the FSA differs between students given the prior year test scores. For example, the data can be used to examine if students who take the FSA test on a tablet tend to have higher or lower scores than students who take the test on a Chromebook or a Windows PC. If there is a device effect (e.g., systematically lower or higher scores on a certain device relative to other devices), it may suggest that students taking the test on that device have a disadvantage or advantage causing for their scores to be affected.

This naturally lends itself to a research question which can be stated simply as "are scores for students participating in the FSA comparable from any device used in the administration?" Simply examining current year scores and disaggregating by device would be insufficient. Students are not randomly assigned to different devices and so we must control for the potential effects of any preexisting differences that would possibly confound the outcomes.

This study analyzes FSA data for students participating in the English Language Arts (ELA) grade 10 retake administration and evaluates the degree to which scores for students taking the test across devices are comparable. The approach controls for preexisting differences between students to control for the non-random assignment of students to different devices.

Methods

The device comparability study implemented uses the grade 10 ELA scaled scores and controls for preexisting differences between students using the grade 9 ELA test score as a covariate. An indicator for device is available at the student level and so we can implement a regression model with the following form

$$y_{tig} = \mu + \beta_1(y_{t-1,i}) + \sum_{j=2}^{5} \beta_j(D_j) + \delta_g + \varepsilon_{ig}$$

Where y_{tig} is the grade 10 ELA scaled score for the ith student who is in school g at time t. The coefficient β_1 is the effect of the prior year grade 9 ELA score and is used to control for preexisting differences between individuals. This variable is measured with error and so the model accounts for that error to avoid bias as described by Doran (2014) and Greene (2000). The coefficients β_j for $j = \{2, \ldots, 5\}$ represent the effect of device D which is a binary coded variable indicating that

$$D_j = \begin{cases} 1 \text{ if student was administered device } j \\ 0 \text{ otherwise} \end{cases}$$

Students are clustered within common groups and so the random effect δ_g is used to account for clustering at the school level to return model-based standard error consistent with the clustered nature of the data.

We can examine the device effects marginally via the regression coefficients. However, that doesn't answer the overall research question. To broadly determine if any device leads to scores that are significantly higher or lower than any other devices, we can use a likelihood ratio test (LRT) to compare the model expressed above (now referred to as the fully specified model) to a baseline model that has the simple form

$$y_{tig} = \mu + \beta_1(y_{t-1,i}) + \delta_g + \varepsilon_{ig}$$

This equation represents a reduced form of the model above where the only difference is that the set of device predictors are not included. The deviance between the fully specified model and the reduced model can be compared using the likelihood ratio test (LRT) as the overall omnibus test to assess whether any device is significantly different from any other device overall.

The LRT test used is

$$Deviance = 2[LL_{fs} - LL_r]$$

The deviance is a χ^2 distributed variable with degrees of freedom equal to the difference in the number of parameters. The *p*-value of the deviance serves as an indicator on the degree to which the fully specified model is significantly different from the reduced model. If the *p*-value of the difference between the two models is significant, then it suggests that student scores on at least one of the devices is significantly different from scores on one of the other devices. If the *p*-value on the deviance is not significant, then it indicates scores between devices are comparable.

Results

The data used for this study are summarized in Table 1 showing the number of students used in this analysis. Table 2 provides the n-sizes by device and the means and standard deviations of the scores disaggregated by device. Table 3 provides the results of the regression models showing the model coefficients and their standard errors for the fully specified and reduced models. Last, Table 4 provides the results of the LRT showing the difference between the effects of the two models.

Table 1: Sample Size by Administration

Administration	N
Spring 2018 Grade 10 ELA Retake	124618
Spring 2017 Grade 9 ELA	201784
Common between two administrations	65562

Table 2: Sample Size by Device (Spring 2018 Grade 10 ELA Retake)

Device	N	Scale Score	
		Mean	SD
Windows	55956	336.377	17.679
Chrome	7046	335.402	17.859
Mac	2134	338.134	16.253
iPad	76	344.579	12.887
Other	350	328.663	19.100
Total	65562	336.298	17.673

Table 3: Regression Coefficients

Model	Variable	Coefficient	Std.Error
Baseline Model	(Intercept)	52.207	1.200
	Prior Year Grade 9 ELA Scale Score	0.866	0.004
Fully Specified Model	(Intercept)	51.416	1.225
	Prior Year Grade 9 ELA Scale Score	0.866	0.004
	iPad	2.542	2.678
	Mac	0.512	0.573
	Other	1.467	0.742
	Windows	0.796	0.261

Table 4: Log Likelihood and Deviance Statistics

Model	Log Likelihood	Deviance	DF	*p*-value
Baseline Model	-2.522580e+05	3.8	4	0.434
Fully Specified Model	-2.522599e+05			

Summary

The *p*-value on the LRT test ($p = .434$) is non-significant indicating that the fully specified model adds no predictors that are significantly different from the reduced model. These results indicate that all regression coefficients on the devices are statistically equivalent, meaning there is no statistically significant difference in the scores for students participating in the FSA on the different types of devices. The data support the notion that there are no systematic differences in the scores for students when administered the FSA on different devices.

References

Doran, H.C. (2014), Methods for Incorporating Measurement Error in Value-Added Models and Teacher Classifications, *Statistics and Public Policy 1*(1), 114-119.

Greene, W.H. (2000), Econometric Analysis *(4th ed.)*, Saddle River, NJ: Prentice Hall.