Comparability of ISIP™ Reading Scores Across Alternate Backgrounds

August 2021

Chalie Patarapichayatham, PhD
Southern Methodist University

Victoria Locke, PhD
Istation

Sean Lewis, MA
Istation

Executive Summary

Formative assessment is used to help inform instruction and provide information on students across an academic year. This research evaluates whether student performance is comparable on a formative assessment when the background changes color while keeping the content consistent. Students in second and third grade who regularly take the Istation's Indicators of Progress – Early Reading assessment (ISIP™-ER) participated in this study. Results show that after controlling for prior achievement and ability level, having a new background slightly increased student performance. Effects varied by grade and subtest; however, all effect sizes were small. The most likely explanation is a novelty effect as the students were exposed to a background that was different than what they had been used to seeing on their monthly assessment.

Introduction

With the prevalence of computer-adaptive testing (CAT) in formative assessment, there is an opportunity for assessments to give students greater choice and agency while maintaining valid and reliable results. Typically, formative assessment is not considered a high-stakes assessment. Rather, formative assessment is used to help inform instruction and provide information on student growth over the course of an academic year so that teachers can give differentiated, time-sensitive instruction (Klute et al., 2017). Taking the same assessment consecutively may lead to student fatigue. The purpose of this research is to evaluate if giving a different visual background, and thus a different look and feel to an assessment, has an impact on student scores and if the norms are applicable for the alternative backgrounds.

In a standardization of an assessment, the purpose is to maintain a fixed delivery mode and keep all testing conditions the same, unless it impacts the student's performance and thus the norms (Way et al., 2016). However, with the advent of digital technology in assessment, test publishers conducted studies to determine if these commonly accepted truths about standardization and norms were still applicable as they made the transition from paper to digital assessment. There is value in maintaining norms if appropriate, as they provide valuable longitudinal information, and they are difficult to compose (Daniel & Wahlstrom, 2019). Researchers conducted equivalency studies to determine if the norms were reliable for the new mode of administration, and they determined that the norms were still applicable if the effect sizes between digital and paper were less than .20 (Daniel & Wahlstrom, 2019; Wright, 2019; Drozdick et al., 2016).

Since most of the research has centered on digital and paper equivalence or comparability, less is known about how differences in backgrounds for a digital assessment may influence a student's scores. This research will evaluate if different backgrounds in the Istation's Indicators of Progress (ISIP) Early Reading (ER) assessment) are comparable.

Background and Objectives

The ISIP ER is a formative assessment and reading screener that is used by millions of students. Based on the science of reading, ISIP ER was authored by reading specialists including Drs. Joe Torgesen, Jeannine Herron, and Patricia Mathes as a way of providing assessment results to teachers that can be used to inform instruction and provide differentiated lessons and intensive intervention for students who are falling behind in reading (Mathes et al., 2016). The ISIP ER subtests include Listening Comprehension, Letter Knowledge, Phonemic Awareness, Vocabulary, Reading Comprehension, Fluency, and Spelling. The ISIP ER is also an approved dyslexia screener in several states including Washington, Texas, Kansas, Oklahoma, and Arkansas. The assessment is administered in a game-like environment with a character that introduces the subtests, tells students to "show what you know," and encourages them to work as fast as they can and do their best. The assessment is engaging and fun for younger students.

The ISIP ER is a CAT assessment that uses a two-parameter item response theory model (Mathes et al., 2016). The first time the student takes the assessment in an academic year, they are given an item of medium difficulty for the grade. If the student answers correctly, they are given a slightly more difficult item. If they answer incorrectly, the next item is less difficult. The next time the student takes the assessment, the CAT algorithm uses a Bayesian estimator of their ability level, or *theta,* from the previous administration and

starts them with an item at their previous theta. In this way, the computer adapts within and across administrations to assess a student's current level of reading ability (Mathes et al., 2016). The ISIP ER also features two separate item banks that rotate each assessment month to limit item exposure month to month.

Many school systems, states, and districts have used ISIP ER for several years, and some students have taken it monthly from prekindergarten through third grade. While the items and subtests may change each year, the basic format of the assessment has stayed the same. We wanted to know if changing the background of the assessment would make a difference in student's scores. The ideal is to provide a student choice of background, as agency may help students achieve their highest ability level (Zeiser et al., 2018).

The goals for the different backgrounds was to make them engaging for students without impacting the item content or how it is delivered, similar to guidelines recommended by Dadey et al. (2018). While the guidelines were specific for studies that look at how an assessment performs across devices, the recommendations were pertinent for this research. These recommendations were conducting a functionality review, holding cognitive laboratories, keeping the content display consistent, and having students familiar with the device they are using. We incorporated all of these suggestions into the design and execution of the forms and the study, which are explained below. We will discuss device familiarity in the Data and Methods section.

Content Display

The team of designers were given instructions to minimize the threats to score comparability. The verbal directions were kept the same, although they are not presented by a character. The test questions are the same, with no differences in how much scrolling or paging a student would need to do to complete the items. The scoring rules and stopping

rules are the same, and the testing conditions are also the same. The only difference is the background of the assessment.

Cognitive Laboratories

The designers created several different backgrounds. Cognitive labs were held with students to ask them to evaluate the backgrounds and identify which they liked best. The designers were able to ask questions regarding the students' choices and make modifications based on the students' recommendations. Two finalists for the backgrounds were selected: Skyline and Night. Skyline has a pastel background and offers a horizon with clouds. The text is presented on a white square with black and white type. Night features a black and teal color scheme. The three backgrounds are presented in Figure 1.

Functionality Review

After the themes were selected, the assessment screens were composed and the designers and quality assurance team went through the screens and items to assess whether or not the items were different from a content perspective or if they performed differently with additional vertical or horizontal scrolling, font sizes, or other issues that may affect an item. The items were deemed to function similarly across several different devices and backgrounds.

Introductory screens for the Original, Skyline, and Night backgrounds

Reading Comprehension screens for the Original, Skyline, and Night backgrounds

Spelling screens for the Original, Skyline, and Night backgrounds

Figure 1. Depiction of the three different backgrounds for ISIP ER for the introductory screens and sample items

Research Questions

In a report sponsored by the Council of Chief State School Officers (CCSSO), framing how the comparability question is determined is important when reviewing claims of comparability. We can ask whether or not the student would receive the same score across the backgrounds, or we can ask which is most likely to produce the most accurate estimates

of student achievement. The first question places importance on standardization, and the second question gives more flexibility with regards to student performance (DePascale et al., 2016).

Since the backgrounds were designed to minimize the differences in the actual assessment — including the content of the questions, the amount of scrolling, the presentation of items, and how students respond to the content — we wanted to know if the alternate backgrounds, which give variety to the student, may help them achieve their highest ability scores. We anticipate that there may be some novelty effects from receiving something new, but we would not expect it to impact a student's true ability level. A student's reading ability is their reading ability, and it is unlikely they would be able to achieve a score that is dramatically divergent from that ability simply by having a different background. Therefore, the research questions guiding us throughout this project were:

1) Is there comparability between the three versions of the ISIP ER assessment (Original, Skyline, and Night) after controlling for prior achievement?
2) Does changing the background of the ISIP ER assessment have an impact on students' scores?

Data and Methods

This research consisted of a pilot study in February 2021 for students in second and third grade, and it was repeated again in May 2021. All participants were recruited from within the Istation customer base and consisted of second and third grade students who were already familiar with Istation and the ISIP ER assessment. Device familiarity was established by having the student assess on the same device that they used in previous administrations in the same school year.

February Pilot Study

In February 2021, Istation customers were invited to participate in the background study, and students in participating schools received the study after logging in to the Istation system. Students were enrolled in 39 different schools in 9 states. Students were randomly assigned to the Original, Skyline, or Night background via a random number generator based on the time and day of the log-in. This randomization feature produced an unbalanced design, with more students being assigned the Original and Night than were assigned to Skyline, as shown in Table 1. Since this assessment was administered during the COVID-19 pandemic, some of the assessments were administered at school, while others were administered at home. We evaluated the sample by looking at the scores from January when there was no exposure to a new form. Results in Table 1 shows that there were not equivalent groups for the February scores. Students in second grade who were administered Skyline had scores in January that were 4-5 points lower than students who received Night or Original in both the home and school conditions. Students in third grade who received Night had slightly lower scores in January than those that received Original or Skyline.

Since there were mean differences between the groups for the January score, we composed a second analytic sample that was stratified based on the student's performance level on the January assessment. Istation offers tier and level groupings for students based on their percentile rank. Level 1 consists of students at or below the 20th percentile, level 2 consists of students between the 21st and 40th percentiles, level 3 consists of students between the 41st and 60th percentiles, level 4 consists of students between the 61st and 80th percentiles, and level 5 consists of students at or above the 81st percentile. Using proprietary code in *R* statistical software, we randomly selected 39 students from each of the five levels. There were 195 students in each grade, leading to a sample size of 390.

Table 1. Means and N counts for Original, Night, and Skyline for January and February Overall ISIP Scores

Grade 2	Score Month	Original	Night	Skyline
Home	January Score	228.89 (N=365)	229.95 (N=368)	225.51 (N=160)
Home	February Score	228.97 (N=365)	233.78 (N=368)	228.46 (N=160)
	Difference	-0.08	3.83	2.95
School	January Score	226.65 (N=402)	226.75 (N=442)	222.55 (N=148)
School	February Score	226.90 (N=402)	230.45 (N=442)	226.24 (N=148)
	Difference	-0.35	4.35	4.31

Grade 3	Score Month	Original	Night	Skyline
Home	January Score	241.61 (N=383)	242.33 (N=442)	241.55 (N=148)
Home	February Score	241.50 (N=383)	244.18 (N=442)	242.92 (N=148)
	Difference	-0.11	1.85	1.37
School	January Score	241.52 (N=407)	239.74 (N=527)	241.35 (N=163)
School	February Score	240.66 (N=407)	243.15 (N=527)	244.44 (N=163)
	Difference	-0.86	3.41	3.09

We used an analysis of variance (ANCOVA) with effects coding to determine if there were differences in scores between the three backgrounds, including the student's January scores to control for prior achievement in both analytic samples. ANCOVA is an appropriate

method because it allows us to control for prior achievement as well as the home or school testing environment.

May Validation Study

We also conducted a study in May 2021. Again, we recruited the participants for the study from the Istation user base. Some schools were new to the background study, while other students had participated in February. There were 1,802 second grade students and 1,801 third grade students, totalling 3,603 students. Of these, 1,373 also participated in the February study. Including these students is appropriate if they do not have the same item exposure (Diao & Keller, 2020). All students were familiar with ISIP ER, and since the assessment is a CAT assessment, there were no specific anchor items used in this test, since it was a regular administration of the assessment. The backgrounds were randomly assigned using a systematic method based on the last digit of a randomly assigned unique student identifier.

In second grade, 777 (43.1%) students took Original, 513 (28.5%) students took Night, and 512 (28.4%) students took Skyline. In third grade, 775 (43.0%) students took Original, 538 (29.9%) students took Night, and 488 (27.1%) students took Skyline. In second grade, 1,139 (63.2%) took the assessment at home and 663 (36.8%) took the assessment at school. In third grade, 1,185 (65.8%) took the assessment at home and 616 (34.2%) took the assessment at school. Half of the sample was female and the other half of the sample was male. There were 47 participating schools in 5 different states. We then removed outliers from the data set by running a regression analysis with the January score and the May score, using Cook's D to identify outliers.

To obtain a representative sample and control for any school effects, we stratified the May sample of each grade level by randomly selecting students equally from the five levels

based on the students' overall scores in January, with 200 students in each level. The final sample consisted of 2,000 students, 1,000 from each grade. In second grade, 431 (43.1%) students took Original, 284 (28.4%) students took Night, and 285 (28.5%) students took Skyline. In third grade, 428 (42.8%) students took Original, 295 (29.5%) students took Night, and 377 (27.7%) students took Skyline. Small samples are often identified as having 100 or fewer examinees per form (Furter & Dwyer, 2000). These sample sizes are sufficient for our research as they are large enough to provide representation across all ability levels. In second grade, 619 (61.9%) took the assessment at home and 381 (38.1%) took the assessment at school. In third grade, 658 (65.8%) took the assessment at home and 342 (34.2%) took the assessment at school.

Next, multiple regression analysis was used in this study. The analysis was completed by grade level for each outcome available. We controlled for prior student achievement by including the January score as a predictor. Testing location was controlled for by creating a dummy variable to indicate whether the assessment was taken at home or at school. We controlled for student variability by including the standard error of the May assessment, and we also included a variable in the regression if they saw a new theme in the February pilot study. Finally, we created a dummy variable for the student's level in Istation in January to control for variability of the slope for higher or lower achieving students. By controlling for all of these elements we were then able to see if the background made a difference. In sum, the predictors in the multiple regression analyses were: 1) May background (Original, Night, and Skyline); 2) May assessment location (Home vs School); 3) January score; 4) May Standard Error (SE); 5) February background (whether a student got a new background in February), and 6) January performance level (Levels 1 to 5). The outcome variables were the

May scores (Comprehension, Overall, Spelling, and Vocabulary). The analyses were completed using *R* statistical software under the lm function.

Results

February Pilot Study

The pilot study in February showed some differences in outcomes on ISIP ER; however, these differences did not remain consistent when accounting for uneven group sizes or prior achievement. Using the complete sample, we ran a simple regression with effects contrast coding, comparing the Night and Skyline backgrounds to the Original background. Using this approach, we found that there were differences in outcomes by background in grade 2. For second grade students who took the assessment in the Night background, scores were significantly higher for students taking the assessment at school (b = 2.48, p = 0.009, N = 360) or at home (b = 2.51, p = 0.031, N = 401).

Table 2. Results of Regression with Contrast Coding on Full Sample

Group	Theme	Coeff.	*p*
G2 Home	Night	2.506	.031*
	Skyline	-1.301	.368
G2 School	Night	2.475	.009**
	Skyline	1.910	.101
G3 Home	Night	1.213	.256
	Skyline	-0.145	.918
G3 School	Night	-0.036	.970
	Skyline	1.987	.123

*p < .05, ** p<.01

However, this did not account for prior achievement or the substantially unequal group sizes. To account for this, we used ANCOVA with January scores as a covariate to account for prior achievement. This model still showed the presence of some differences, though results were inconsistent with the previous model. Our analysis using ANCOVA revealed significant differences between scores in grade 2, regardless of location of the assessment. However, there were now also significant differences in grade 3. These results are summarized in Table 3 below.

Table 3. ANCOVA Model with Contrast Coding Using January Score as Covariate on Full Sample

Group	Variable	Coeff.	SE	*t*	*p*
G2 Home	January Score	0.83	0.02	43.40	0.000
	Night	-6.48	5.86	-1.11	0.269
	Skyline	23.30	6.67	3.50	0.000
G2 School	January Score	0.85	0.02	49.46	0.000
	Night	-8.68	5.10	-1.702	0.089
	Skyline	21.06	5.93	3.55	0.000
G3 Home	January Score	0.82	0.02	40.79	0.000
	Night	14.42	6.09	2.37	0.018
	Skyline	-16.21	8.18	-1.98	0.048
G3 School	January Score	0.83	0.02	43.80	0.000

	Night	8.02	5.56	1.44	0.150
	Skyline	-4.80	7.68	-0.63	0.532

As a result of the inconsistencies between the initial model and the ANCOVA, we ran the models again using the stratified sample. This meant that in the final sample there were equal numbers of students at each achievement level across the three backgrounds as well as equal numbers of students overall. Results are available in Table 4. These results revealed no statistically significant differences across backgrounds for students in grade 2, regardless of location of test administration. However, there were significant differences between students in grade 3, but only for students who tested at school and those with the Skyline background. Specifically, students who were tested in the Night background had higher scores (b = 5.29, p = .069) as did students who tested in the Skyline background (b = 6.01, p = 0.047).

The results from the February pilot study were promising, as while there was some statistical significance, the results varied across the samples. We did not observe large effects if students took a different background, indicating that the background would not impact a student's overall reading ability. We conducted another study in May to see if the results would change with a larger sample size.

Table 4. Regression with Contrast Coding on Stratified Sample Outcomes

Group	Theme	Coeff.	p
G2 Home	Night	-1.498	0.694
	Skyline	1.511	0.695
G2 School	Night	3.011	0.302

	Skyline	1.968	0.497
G3 Home	Night	0.862	0.811
	Skyline	2.177	0.523
G3 School	Night	5.293	0.069†
	Skyline	6.014	0.047*

May Validation Study

We ran an additional study in May to again assess whether there were differences in student performance across the three backgrounds. Next, we looked at the mean scores of the Overall score, Reading Comprehension (CMP), Spelling (SPL), and Vocabulary (VOC) in January and May by grade and by the background shown in Tables 5 and 6. Even though there was better randomization of forms, students who had the original background in May had lower scores in January than students who received either the Night or the Skyline background. There were also differences in the May scores. Therefore, any mean differences in May could perhaps be due to student level ability, rather than the background of the assessment.

Table 5: Mean Scores of CMP, Overall, SPL, and VOC in January and May by Grade and by May Background of Second Grade

May Background	Subtest	January		May	
		Mean	SD	Mean	SD
Original	Overall	219.8	25.3	221.5	27.3
	CMP	223.1	26.7	224.5	29.3
	SPL	221.1	22.3	224.5	24.4
	VOC	226.8	26.8	229.2	28.3
Night	Overall	224.7	25.5	228.7	25.1
	CMP	229.1	25.9	237.2	24.4
	SPL	226.8	23.4	229.6	23.1
	VOC	232.2	28.7	235.0	28.3
Skyline	Overall	225.8	24.9	232.8	24.8
	CMP	231.3	22.3	240.7	24.2
	SPL	226.5	21.6	231.8	21.2
	VOC	232.3	31.4	236.8	28.2

Table 6: Mean Scores of CMP, Overall, SPL, and VOC in January and May by Grade and by May Background of Third Grade

May Background	Subtest	January		May	
		Mean	SD	Mean	SD
Original	Overall	236.7	24.8	237.5	28.6
	CMP	238.3	24.7	241.4	28.1
	SPL	236.7	22.8	237.4	25.3
	VOC	242.3	29.9	246.2	31.6
Night	Overall	239.6	24.4	244.3	25.9
	CMP	244.5	25.7	252.9	25.8
	SPL	239.3	21.2	244.8	20.1
	VOC	247.1	27.7	250.8	32.0
Skyline	Overall	239.1	23.8	242.2	25.3
	CMP	244.5	25.6	250.1	26.5
	SPL	238.4	23.1	242.3	21.8
	VOC	245.7	29.0	250.2	30.7

The mean scores of Overall, CMP, SPL, and VOC in May by grade, by assessment location, and by May background are shown in Table 7. Students who took the assessment from home scored higher than students who took the assessment from school in both second

and third grades across all backgrounds. This is consistent with prior research that indicates the scores from assessments taken at home are slightly higher than those taken at school (Huff, 2020; Kuhfeld, M. et al., 2020; Locke, Patarapichayatham, & Lewis, 2021). The differences range from 2 to 15 points depending on grade and subtest. On average, students in second grade who took the assessment from home with Original background scored 6 points higher than students who took the assessment from school with Original background. Students who took the assessment from home with the Night background scored 5 points higher, and students who took the assessment from home with Skyline background scored 11 points higher than those students who took the assessment at school. In third grade, students who took the assessment from home with Original background scored 3 points higher on average than students who took the assessment from school with Original background. Students who took the assessment from home with the Night background or the Skyline background scored 9 points higher than those students who took the assessment at school. The standard deviations (SD) of scores were consistent but slightly higher for students who took the assessment from home.

Table 7: Mean Scores of CMP, Overall, SPL, and VOC in May by Grade, by Assessment Taking Location, and by May Background

May Theme	Assessment Location	Subtest	Grade 2		Grade 3	
			Mean	SD	Mean	SD
	Home	Overall	223.1	28.0	238.3	30.4
		CMP	226.6	30.2	242.4	29.0
		SPL	227.2	24.5	238.4	26.2

Original		VOC	231.1	28.8	248.1	33.0
	School	Overall	218.6	25.7	235.9	24.8
		CMP	220.6	27.2	239.4	26.3
		SPL	219.2	23.5	235.6	23.5
		VOC	225.7	27.1	242.5	28.4
Night	Home	Overall	231.0	24.2	247.4	26.2
		CMP	238.8	22.2	257.0	24.7
		SPL	232.6	23.3	247.5	19.2
		VOC	235.6	28.5	254.6	33.1
	School	Overall	225.5	26.1	239	24.5
		CMP	234.9	27.1	245.7	26.3
		SPL	225.2	22.1	240.2	21.0
		VOC	234.1	28.3	244.6	29.4
Skyline	Home	Overall	237.2	24.9	246.2	25.4
		CMP	245.7	20.7	253	26.1
		SPL	236.7	21.0	245.6	22.2
		VOC	242.1	29.3	255.1	31.4
		Overall	226.4	23.4	234.0	23.1

	School	CMP	236.3	27.7	243.9	26.4
		SPL	225.3	19.8	235.4	19.4
		VOC	229.2	24.6	240.0	26.6

Text Fluency is another subtest for students in second and third grades. This particular subtest is not included in the Overall score calculation, and it has a different approach of calculating the scores. We computed the word correct per minute (WCPM) for this subset by grade and by background. We then computed predicted WCPM by background and by grade and results are in Table 8. The background does not seem to have an impact on students' performance in May. In second grade, students with Original had 19 WCPM, 21 for Night, and 20 for Skyline. In third grade, students with Original and Night had 23 WCPM, and students with Skyline had 24. Theme does not appear to have an impact on students' WCPM on the Text Fluency subtest.

Table 8. Word Correct Per Minute for Text Fluency

Grade	Theme	WCPM	p-value
2	Original	19	
	Night	21	0.101
	Skyline	20	0.558
3	Original	23	
	Night	23	0.770
	Skyline	24	0.396

Because there were mean differences at the outset as determined by the differences in the January scores, a multiple regression model was calculated to predict each outcome variable and results are shown in Tables 9 and 10. This analysis was conducted with the stratified sample. Students without a January score or who did not have an Overall or May subtest score were dropped from the analysis. The table includes the effects for the background, the p value, effect size, and model R^2.

Table 9: Multiple Regression Analyses Results of Second Grade

May Score	May Theme	Estimate	Standard Error	*p*	Effect Size	Model R^2	F-statistic/ p-value
Overall	Original	229.51	1.18	<0.001		0.74	236.0(10, 845) <0.001
	Night	231.66	1.06	0.042	0.03		
	Skyline	230.61	1.05	0.248	0.02		
CMP	Original	228.51	1.42	<0.001		0.67	158.7(10, 784) <0.001
	Night	234.95	1.30	<0.001	0.06		
	Skyline	232.11	1.27	0.003	0.04		
SPL	Original	228.64	1.63	<0.001		0.55	77.9(10, 788) <0.001
	Night	231.95	1.46	0.011	0.03		
	Skyline	231.21	1.44	0.070	0.02		
VOC	Original	235.24	0.83	<0.001		0.89	664.2(10, 845) <0.001
	Night	237.70	0.74	<0.001	0.01		
	Skyline	237.07	0.73	0.012	0.01		

Table 10: Multiple Regression Analyses Results of Third Grade

May Score	May Theme	Estimate	Standard Error	*p*	Effect Size	Model R^2	F-statistic/ p-value
Overall	Original	242.23	1.30	<0.001		0.68	187.6(10, 880) <0.001
	Night	244.98	1.18	0.011	0.01		
	Skyline	243.64	1.20	0.048	0.01		
CMP	Original	239.50	1.28	<0.001		0.70	195.3(10, 821) <0.001
	Night	244.99	1.16	<0.001	0.03		
	Skyline	244.53	1.22	<0.001	0.03		
SPL	Original	237.17	1.53	<0.001		0.57	73.7(10, 807) <0.001
	Night	241.59	1.34	0.001	0.03		
	Skyline	238.89	1.42	0.217	0.01		
VOC	Original	248.47	0.82	<0.001		0.91	843.2(10, 879) <0.001
	Night	249.18	0.83	0.337	0.01		
	Skyline	248.67	0.89	0.079	0.01		

At the subtest level, the difference in scores varies by background. For the CMP subtest in second grade, compared to students who took the Original background, students who took Night scored approximately 5 points higher, and students who took Skyline scored approximately 3 points higher. Eta Squared is 0.06 and 0.04, showing a small effect size. In

SPL and VOC, students who took Skyline or Night scored approximately 2 points higher than students who took the Original background. Skyline was not significantly different from Original for SPL and VOC, and Night was not significant for VOC. Effect sizes continue to be small. Eta squared was 0.03 for Night and 0.02 for Skyline in SPL, and in VOC eta squared is 0.01 for both Night and Skyline.

In third grade, students who took Night scored approximately 2 points higher on the Overall score, and students who took Skyline scored approximately 1 point higher than students who took the Original background. Eta squared is 0.01 for both backgrounds. Students who took Night or Skyline scored approximately 5 points higher than students who took the Original background on the CMP subtest, and eta squared is 0.03 for both backgrounds. In SPL, there were no statistically significant differences on scores among students who took the Original or Skyline backgrounds. Students who took the Night background, however, scored 4 points higher than students who took the Original background. Eta squared was 0.03 for Night and 0.01 for Skyline. In VOC, there were no statistically significant differences on scores among students who took the Original, Night, or Skyline backgrounds. Eta-Squared was 0.01 for both Night and Skyline.

Discussion

This research evaluated whether changing the background of an assessment that is familiar to students helps them better demonstrate their reading ability. The students in this study were enrolled in schools that had used Istation and ISIP ER for at least two years, and the students had all taken the ISIP ER assessment previously, some for three years. In ISIP ER, the student plays a game with characters that introduce a task, and the student is a contestant on the Show What You Know game. However, the novelty may wear off after a few

months, and students may become overly familiar with how the material is presented. There is value in using the same assessment over several years, as school districts can watch students grow on the same scale, compared to a comparable norm set. The purpose of a formative assessment is to have students assess periodically so that teachers can use the data to inform instruction and address the students' needs quickly (Klute et al., 2017). Therefore, it is important that an assessment be novel enough that a student is engaged and can do their best, and familiar enough that the student can do their best without forgetting how to take the assessment, losing valuable time.

The results from this study demonstrate that having something that is both new and familiar may help students stay engaged in the testing process and help them achieve their reading ability. No matter which background a student receives, if they do not know a vocabulary word, for example, they will not get the question correct except by guessing. The same goes for the Spelling, Text Fluency, and Reading Comprehension subtests. The most that the background can do is give them a screen that scrolls well and does not detract from the content, while providing something a little new.

Scores are typically considered comparable if the effect sizes are less than 0.20 (Daniel & Wahlstrom, 2019). Although significant effects may be detected in the data, throughout this research the effect sizes as measured by eta squared are all well below 0.20, with the highest at 0.04 to 0.06 for the CMP subtest in second grade.

The most likely explanation for the results is the novelty effect of having something new after several testing sessions across multiple years of seeing the same background. When we included a variable for whether or not the student had also seen a new background in the February pilot study, the variable was not significant, indicating that the second time of

seeing a new background still had an appeal for the student. Having something slightly different may help students stay engaged.

Recommendations

The CCSSO presented two statements regarding score comparability (DePascale et al., 2016). The first statement is concerned about forms producing the same score. This statement is favored by those who are interested in standardization of forms, and everyone taking the same assessment under the same conditions. The second statement concerns whether different forms helped students to best demonstrate their academic ability, and this statement allows a degree of flexibility accepting that the construct being measured is not altered, and it is consistent with the interpretation and use of the testing results. Our research questions were written with these statements in mind.

In evaluating the first research question, we found that there were small yet positive effects for using the assessment with a different background. Districts that are focused on standardization and using ISIP ER for higher-stakes decisions may want to give students access only to the ISIP ER Original theme.

In answering the second question, the results from this research shows that having something that is new yet still familiar may help students better demonstrate their reading ability. We determined that the new forms had slightly higher scores, but the effect sizes were small. The effect sizes indicate that the differences in scores were not enough to demonstrate huge gains in reading, but rather the novelty of having a different background helped students reach their ability level. Districts that are using ISIP ER for formative assessment to drive instruction may decide to give students access to all three themes and allow the students to choose their background.

Limitations and Future Research

There are some limitations to this research that may limit the generalizability of the results. First, the sample came entirely from Istation users that were familiar with the assessment. The backgrounds may produce somewhat different results if they were administered to students with no familiarity with the assessment. Second, since Istation does not require student-level demographics in our system, we do not have reliable information on student race/ethnicity and poverty status. We accounted for this by stratifying the final sample by student ability, thus ensuring a representative sample of students of all ability levels. Last, we do not know if these results will remain consistent after students have become habituated to having access to more than one background. When we accounted for whether or not students had seen either Skyline or Night in the February pilot, the variable was not significant, indicating that having students who took the assessment with Skyline or Night continued to have slightly increased scores. Future research should investigate if, over time, these results are mediated by familiarity with the different backgrounds.

We also do not know yet whether having a choice of background may also help students stay engaged in the testing process. Student agency is the ability to own and manage one's own learning, and it can have a significant effect on academic performance (Zeiser et al., 2018). In this research, students received one of three backgrounds by randomization, and they were not able to select a background. Future research should evaluate whether being able to select a background, thus increasing student agency, would have an impact on a student's scores and help them better demonstrate and achieve their highest score on a reading assessment.

REFERENCES

DePascale, C., Dadey, N., Lyons, S., & Council of Chief State School Officers (CCSSO). (2016). Score Comparability across Computerized Assessment Delivery Devices: Defining Comparability, Reviewing the Literature, and Providing Recommendations for States When Submitting to Title 1 Peer Review. In *Council of Chief State School Officers*. Council of Chief State School Officers. Retrieved from https://files.eric.ed.gov/fulltext/ED610777.pdf.

Dadey, N., Lyons, S., & DePascale, C. (2018). The comparability of scores from different digital devices: A literature review and synthesis with recommendations for practice. Applied Measurement in Education, 31, 1, 30-50. Retrieved from https://doi.org/10.1080/08957347.2017.1391262

Daniel, M., & Wahlstrom, D. (2019). Raw-score equivalence of computer-assisted and paper versions of WISC–V. *Psychological Services, 16*(2), *213–220*. Retrieved from https://doi.org/10.1037/ser0000295

Diao, H. & Keller, L. (2020). Investigating repeater effects on small sample equating: Include or exclude? *Applied Measurement in Education, 33, 1, 54-66*. Retrieved from https://doi.org/10.1080/08957347.2019.1674302

Drozdick, L. W., Getz, K., Raiford, S. E., Zhang, O. (2016). WPPSI-IV: Equivalence of Q-interactive and paper formats. San Antonio: Pearson. Retrieved from https://images.pearsonclinical.com/images/assets/WPPSI-IV/WPPSI-Qi-Tech-Report-14_FNL.pdf.

Furter, R. T., & Dwyer, A. C. (2020). Investigating the classification accuracy of Rasch and

nominal weights mean equating with very small samples. *Applied Measurement in Education, 33,1, 44-53*. Retrieved from https://doi.org/10.1080/08957347.2019.1674307.

Huff, K. (2019). National Data Quantifies Impact of COVID Learning Loss. Retrieved from https://www.curriculumassociates.com/-/media/mainsite/files/i-ready/ca-impact-of-covid-learning-loss-fall-2020.pdf.

Klute, M., Apthorp, H., Harlacher, J., Reale, M., Regional Educational Laboratory Central (ED), National Center for Education Evaluation and Regional Assistance (ED), & Marzano Research Laboratory. (2017). Formative Assessment and Elementary School Student Academic Achievement: A Review of the Evidence. REL 2017-259. In *Regional Educational Laboratory Central*. Regional Educational Laboratory Central.

Kuhfeld, M., Lewis, K., Meyer, P. & Tarasawa, B. Comparability analysis of remote and in-person MAP Growth testing in fall 2020 (2020). Retrieved from https://www.nwea.org/content/uploads/2020/11/Technical-brief-Comparability-analysis-of-remote-and-inperson-MAP-Growth-testing-in-fall-2020-NOV2020.pdf

Locke, V. N., Patarapichayatham, C., & Lewis, S. (2021). Learning Loss in Reading and Math in U.S. Schools Due to the COVID-19 Pandemic. Retrieved from https://www.istation.com/Content/downloads/studies/COVID-19_Learning_Loss_USA.pdf.

Way, W. D., Davis, L. L., Keng, L., & Strain-Seymour, E. (2016). From standardization to personalization: The comparability of scores based on different testing conditions, modes, and devices. In E. Drasgow (Ed.), *Technology and Testing: Improving Educational and Psychological Measurement*. (pp. 260-284). New York: Routledge.

Wright, A. J. (2018). Equivalence of remote, online administration and traditional,

face-to-face administration of Woodcock-Johnson IV Cognitive and Achievement Tests. *Archives of Assessment Psychology, 8, 1, 23-35*. Retrieved from https://www.assessmentpsychologyboard.org/journal/index.php/AAP/article/view/122/78.

Zeiser, K., Scholz, C. & Cirks, V. (2018). Maximizing Student Agency: Implementing and Measuring Student-Centered Learning Practices. American Institutes of Research. Retrieved from ERIC, August 9, 2021 https://files.eric.ed.gov/fulltext/ED592084.pdf.