“COVID Slide” Not Evident in Individually Administered Clinical Test Scores Obtained From a Large, Referred Sample

Susan Engi Raiford, PhD
Kristina Breaux, PhD
Siqi Chen, PhD
Tyler Matta, PhD

February 2021

Abstract

Simulation studies in spring 2020 forecasted that classroom reading and math test scores would be lower than typical fall scores in the fall of 2020 due to widespread school interruption related to the COVID-19 pandemic. Subsequent data provided by the authors of the simulation studies, however, suggested that classroom test scores obtained in fall 2020 did not show the expected differences for reading, and that the size of the differences for math test scores was much smaller than predicted.

Clinical test users, concerned about the impact of school disruption on diagnostic tools following the aforementioned projections about classroom tests, requested information about the consistency and continued applicability of the norms for individually administered achievement and cognitive tests used for evaluations in referred populations. Given these concerns about the applicability of norms following the pandemic's onset, we sought to use data available to us to begin answering the question of how test scores may have shifted between 2019 and 2020. Because COVID-19 was an unexpected event, the data here are not the result of an a priori study design, but rather, convenience samples that offered the best available preliminary insights.

Survey data were obtained from practitioners who used Pearson's Q-interactive® to engage in performance-based testing during May through August of 2020 to gain more information about decisions to assess during the COVID-19 outbreak and the conditions under which testing sessions were conducted. Practitioners stated that during these months they used personal protective equipment (PPE) and remote assessment for a large proportion of the evaluations and that they prioritized testing for the examinees with greatest urgent need of assessment.

Scores of referred examinees in the U.S. aged 6–16 who were administered clinical achievement or cognitive ability tests using Pearson's Q-interactive in May–August 2020 were deidentified (with the exception of the examinee's age) and compared with the analogous scores of examinees tested during May–August 2019. Despite school disruption and remote learning as well as widespread use of PPE and remote assessment, the data indicate that the scores and composite score distributions obtained by a referred population on individually administered clinical achievement and cognitive ability tests were highly consistent across May–August 2019 and May–August 2020. The samples are not randomly selected or demographically matched, and for these reasons, the results should be considered preliminary. However, the May–August 2019 and May–August 2020 composite score distributions and means very closely map onto one another, which suggests that the normative data for these clinical achievement and cognitive ability tests continue to provide a valid and appropriate reference point for score interpretation. These preliminary results should not be used to infer expected performance on large-scale classroom-based tests used with the general population, but should be considered when interpreting clinical, individually administered tests with referred populations only.

Introduction

The COVID-19 pandemic resulted in an almost complete closure of physical school buildings in spring 2020; 27 states mandated or recommended building closures beginning March 16, with the remainder doing so by March 23 (Education Week, 2020). Estimates indicate that school closures impacted at least 55.1 million students in 124,000 U.S. public and private schools (Education Week, 2020), which represents almost all children in the U.S. aged 5–17 (U.S. Census Bureau, 2020).

Simulation studies conducted using large classroom test datasets forecasted lower fall 2020 learning test scores due to COVID-19-related educational disruption in spring 2020 (Kuhfeld et al., 2020a). These projections were based upon estimates from prior literature and analyses of typical effects of summer break, weather-related school closure (e.g., Hurricane Katrina), and chronic absenteeism. The projections for the 2019–20 school year suggested learning gains in reading could be as little as 63–68% of the typical school year, and as little as 37–50% in math. It was further projected that the loss of progress wouldn't necessarily affect all students equally and that perhaps the top third of students may make gains in reading. Subsequent analyses on fall 2020 data (Kuhfeld et al., 2020b) indicate that the simulation studies overprojected reading score reductions; in fact, students in fall 2020 performed similarly in reading relative to students from the same grade in fall 2019. However, fall 2020 math scores were 5–10 percentile points lower than those of same-grade students in fall 2019.

Clinical test users, concerned about the impact of school disruption on diagnostic tools, requested information about the scores produced following pandemic-related school disruption on individually administered, norm-referenced achievement and cognitive tests used for evaluations in referred populations. One national organization questioned whether norms for standardized clinical tests would apply to children being evaluated during the 2020–21 school year (National Association of School Psychologists, 2020).

Present Study

The goal of the current study was to determine if differences are observed in achievement and cognitive ability score distributions in a large, referred sample of examinees following educational disruption that occurred nationwide in spring 2020. The focus of the present investigation was on scores of school-aged children on individually administered, norm-referenced, diagnostic academic achievement and cognitive ability tests.

Method

Study Design

Given the concerns about norms applicability following the disruptive impact of the pandemic in educational settings, we sought to use data available to us to begin answering the question of how test scores may have shifted between 2019 and 2020. Because COVID-19 was an unexpected event, the data here are not the result of an a priori study design. Rather, they are based upon convenience samples that offered the best available preliminary insight into the impact of the pandemic on standardized clinical test scores.

For the present study, academic achievement and cognitive ability test scores were obtained for children aged 6–16 tested in the U.S. on Pearson's Q-interactive platform during May through August of 2019 and May through August of 2020. Test scores obtained by these very large, naturally occurring, referred samples were examined for differences observed between the two time periods.

The nature of the data is such that only the examinees' ages and the months and years of administration are available. For this reason, a survey was conducted with all Q-interactive users to gather information about any modifications to their current practices. The results of this survey are discussed in the Procedure section.

Measures

WIAT–III

The *Wechsler Individual Achievement Test–Third Edition* (WIAT–III; Pearson, 2009) is an individually administered clinical instrument designed to measure the academic achievement of examinees ages 4 through 50 and students in prekindergarten (PK) through Grade 12. The WIAT–III has 14 subtests and 4 core composite scores.

KTEA–3

The *Kaufman Test of Educational Achievement–Third Edition* (KTEA–3; Kaufman & Kaufman, 2014) is an individually administered, comprehensive measure of educational achievement for children, adolescents, and young adults ages 4 through 25 and students in PK through Grade 12. The KTEA–3 has 14 subtests and 13 composite scores. The 4 core composite scores were examined for this study.

WISC–V

The *Wechsler Intelligence Scale for Children–Fifth Edition* (WISC–V; Wechsler, 2014a) is an individually administered, comprehensive clinical instrument for assessing the cognitive ability of children ages 6–16. It has five primary index scores and a composite score that represents general intellectual ability (i.e., Full Scale IQ. ***Due to issues with the WISC–V Coding subtest in digital format in 2019, an a priori decision was made to exclude the data associated with that format from the 2019 analyses for the Coding subtest scaled scores and for the Processing Speed Index***. Hence, only data associated with presentation of the test in paper format (i.e., response booklet) is included.

Participants

Tables 1–4 present detailed information about the ages of children tested on the two measures of achievement, the WIAT–III and the KTEA–3, and the measure of cognitive ability, the WISC–V.

Table 1. Sample Size and Age-Related Characteristics of the May–August Samples, by Test and Year

	WIAT–III		KTEA–3		WISC–V	
	2019	**2020**	**2019**	**2020**	**2019**	**2020**
N	9,587	4,480	9,334	3,030	25,193	11,094
Age						
Mean	10.5	10.5	10.0	10.1	10.2	10.2
SD	3.0	3.0	2.8	2.9	2.9	2.9

Table 2. Percentages of the May–August WIAT–III Samples by Age and Year

	2019			2020		
Age group	*N*	Percent	Cum. percent	*N*	Percent	Cum. percent
6	701	7.3	7.3	333	7.4	7.4
7	1,148	12.0	19.3	542	12.1	19.5
8	1,222	12.7	32.0	599	13.4	32.9
9	1,106	11.5	43.6	490	10.9	43.8
10	982	10.2	53.8	449	10.0	53.9
11	895	9.3	63.1	417	9.3	63.2
12	797	8.3	71.5	346	7.7	70.9
13	766	8.0	79.5	314	7.0	77.9
14	713	7.4	86.9	369	8.2	86.1
15	650	6.8	93.7	322	7.2	93.3
16	607	6.3	100.0	299	6.7	100.0

Table 3. Percentages of the May–August KTEA–3 Samples by Age and Year

	2019			2020		
Age group	*N*	Percent	Cum. percent	*N*	Percent	Cum. percent
6	773	8.3	8.3	280	9.2	9.2
7	1253	13.4	21.7	414	13.7	22.9
8	1397	15.0	36.7	389	12.8	35.7
9	1265	13.6	50.2	376	12.4	48.2
10	1041	11.2	61.4	343	11.3	59.5
11	912	9.8	71.1	293	9.7	69.1
12	710	7.6	78.8	204	6.7	75.9
13	586	6.3	85.0	209	6.9	82.8
14	514	5.5	90.5	212	7.0	89.8
15	492	5.3	95.8	162	5.3	95.1
16	391	4.2	100.0	148	4.9	100.0

Table 4. Percentages of the May–August WISC–V Samples by Age and Year

	2019			2020		
Age group	*N*	Percent	Cum. percent	*N*	Percent	Cum. percent
6	1,955	7.8	7.8	943	8.5	8.5
7	3,395	13.5	21.2	1,450	13.1	21.6
8	3,665	14.5	35.8	1,663	15.0	36.6
9	3,037	12.1	47.8	1,263	11.4	47.9
10	2,567	10.2	58.0	1,154	10.4	58.3
11	2,296	9.1	67.1	992	8.9	67.3
12	2,011	8.0	75.1	847	7.6	74.9
13	1,999	7.9	83.1	810	7.3	82.2
14	1,863	7.4	90.5	868	7.8	90.0
15	1,661	6.6	97.0	781	7.0	97.1
16	744	3.0	100.0	323	2.9	100.0

As shown, fewer children were assessed with these tests in May–August 2020 than during May–August 2019. The numbers of children tested in May–August 2020 with the WIAT–III, the KTEA–3, and the WISC–V were 47%, 33%, and 44% respectively of the numbers of children tested with these measures during May–August 2019. The age-related values (i.e., means and standard deviations) of examinees assessed during these time periods, as well as the distribution of examinees by age, are highly similar across the two years, however.

Procedure

Survey

A survey was sent to Q-interactive practitioners to gather information about their backgrounds, primary work setting, and approach to assessment during this range of months. A primary interest was whether the clinical conditions of the assessments conducted in May–August 2020 differed relative to the assessments conducted from May–August 2019.

Information about respondents (practitioners) who engaged in performance-based testing on Q-interactive with school-age children from May–August 2020 appears in Tables 5–7.

Table 5. May–August 2020 Q-interactive Practitioners With Various Job Titles

Title	Percent	*N*
Clinical psychologist	15.6	128
Neuropsychologist	7.2	59
School psychologist	47.3	359
Psychometrist/Psychological technician/Educational diagnostician	8.6	71
Speech-language pathologist	10.7	88
Other	10.7	88
Total	**100.0**	**823**

Table 6. May–August 2020 Q-interactive Practitioners from Various Primary Work Settings

Setting	Percent	*N*
School–Public	64.5	531
Private practice/clinic/group practice	25.3	208
Hospital	3.9	32
University	2.2	18
Other	1.7	14
School–Private	1.3	11
Mental health center	0.9	7
Early childhood facility	0.2	2

Table 7. May–August 2020 Q-interactive Practitioners and U.S. Population by State

State	Percent of practitioners	U.S. population	State	Percent of practitioners	U.S. population
Alabama	0.7	1.5	Nebraska	0.8	0.6
Alaska	0.5	0.2	Nevada	0.4	0.9
Arizona	1.8	2.2	New Hampshire	0.4	0.4
Arkansas	0.5	0.9	New Jersey	2.2	2.7
California	7.7	11.9	New Mexico	0.5	0.6
Colorado	2.5	1.7	New York	4.2	5.9
Connecticut	0.8	1.1	North Carolina	2.3	3.2
Delaware	0.5	0.3	North Dakota	0.1	0.2
Florida	6.2	6.5	Ohio	4.8	3.5
Georgia	5.0	3.2	Oklahoma	1.0	1.2
Hawaii	0.7	0.4	Oregon	1.2	1.3
Idaho	0.2	0.5	Pennsylvania	4.8	3.8
Illinois	4.8	3.9	Rhode Island	0.2	0.3
Indiana	3.4	2.0	South Carolina	1.6	1.6
Iowa	0.8	1.0	South Dakota	0.4	0.3
Kansas	0.2	0.9	Tennessee	1.8	2.1
Kentucky	0.7	1.4	Texas	13.3	8.7
Louisiana	0.1	1.4	Utah	2.4	1.0
Maine	0.7	0.4	Vermont	0.2	0.2
Maryland	2.2	1.8	Virginia	3.4	2.6
Massachusetts	3.3	2.1	Washington	1.4	2.3
Michigan	2.2	3.0	Washington, DC	0.1	0.2
Minnesota	1.9	1.7	West Virginia	0.8	0.5
Mississippi	0.6	0.9	Wisconsin	1.0	1.8
Missouri	1.7	1.9	Wyoming	0.2	0.2
Montana	0.5	0.3			

As shown, despite school closures, practitioners from the public schools and school psychologists were well represented. Additionally, practitioners who conducted testing during this time period were distributed across all states and in roughly similar proportions to the U.S. general population.

Practitioners were asked to rate their agreement with statements about how they prioritized assessment and whether they limited performance-based testing during this time period in 2020. Their responses are summarized in Table 8.

Table 8. Percentages of Q-interactive Practitioners Agreeing with Statements About Testing Practices in May–August 2020

Statement	Strongly disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Strongly agree	*N*
I conducted less performance-based testing than typical for that time of year.	12.5	11.7	6.5	32.4	36.9	802
I prioritized assessment as I usually do.	7.1	16.4	7.6	30.3	38.6	803
I limited performance-based testing to only the cases with the most serious need.	22.0	17.6	13.2	30.5	16.6	794

As expected, a large majority of respondents (69%) indicated they conducted less performance-based testing than typical for these months in 2020, which is consistent with observed differences in Q-interactive usage data. Most respondents (69%) indicated they prioritized assessment as they usually do during May–August 2020. Although nearly half (47%) indicated they limited performance-based testing to only the cases with the most serious need, over a third (40%) of respondents did not limit testing in this way. When asked if there were particular types of assessment they are delaying until after the pandemic subsided, about a fourth (24%) of respondents mentioned autism in the context of a popular measure that cannot be administered while wearing masks, but no other response was clearly predominant.

Two common approaches to address safety concerns related to the spread of COVID-19 are to use PPE or test examinees via tele-assessment while they are located either in a testing office or at their home. Practitioners were asked to estimate the percentage of all Q-interactive evaluations during May–August 2020 for which they used PPE and tele-assessment. Some hybrid face-to-face and tele-assessment approaches involve various configurations, so the total percentages do not need to approximate 100%. For example, the examinee can be located a) in a separate office in the practitioner's suite for some tasks, and then in the same room as the practitioner to use PPE for tasks that involve manipulatives, or b) at their home for some tasks, and then in the same room as the practitioner to use PPE for tasks that involve manipulatives. Table 9 summarizes the practitioners' average estimated percentages of evaluations during this time period conducted with PPE and tele-assessment.

Table 9. Practitioners' Estimated Percentages of May–August 2020 Q-interactive Evaluations Using PPE and Tele-Assessment

	Percentage			
Method	**Mean**	**Median**	**SD**	**N**
Face-to-Face with PPE	89	100	26	696
Tele-assessment with examinee in testing office	7	0	22	808
Tele-assessment with examinee located in their home	18	0	34	808

Note. *N* column provides the number of respondents that provided an estimate of percentage of evaluations conducted using each method.

As shown, the typical practitioner relied heavily on PPE when conducting performance-based testing. The typical practitioner relied on some type of tele-assessment 25% of the time.

Table 10 summarizes the percentages of practitioners using Q-interactive with various types of PPE or safety measures (total respondents *N* = 699). For the purposes of the survey, options were limited to common types of PPE or safety measures that might impact test performance.

Table 10. Percentage of May–August 2020 Q-interactive Practitioners Using Various Types of PPE and Safety Measures

PPE or safety measure	Percent
Examiner using mask	92.0
Client using mask	91.3
Plexiglass shield or other physical barrier	67.1
Examiner using face shield	49.6
Examiner using gloves	29.0
Protecting paper stimuli with lamination or other methods	26.8
Client using face shield	19.3
Examiner using eye protection	19.3
Conducting testing outside	11.0
Client using gloves	7.3
Client using eye protection	3.4
I did not use any of these particular PPE or safety measures	0.7

Note. The term "client" was used to refer to the examinee to avoid respondent confusion between the words *examiner* and *examinee*.

As shown, the vast majority (>90%) of practitioners conducted assessments during this time with both examiner and examinee wearing a face mask. Use of a physical barrier between examiner and examinee was reported by over half (67%) of respondents, and about half of examiners wore a face shield. Other safety measures were used less frequently.

Practitioners using PPE while testing with Q-interactive in May through August 2020 were queried about their experience of the impact of PPE on results, as follows: "Based on your experience/observations, how much does the use of PPE impact performance-based test results and interpretation?" Their responses are summarized in Table 11.

Table 11. Percentages of May–August 2020 Q-interactive Practitioners Indicating Various Levels of PPE Impact on Performance-Based Test Results

Impact	Percent
No impact	9.9
Minimal impact	50.9
Some impact	34.8
Significant impact	4.4
Total respondents	***N* = 699**

As shown, only about 10% of practitioners with performance-based testing experience with PPE have the impression that PPE had no impact on results and interpretation were not affected. While each respondent may have interpreted these terms differently, it is assumed that those who responded "minimal" were indicating that while PPE had an impact on results and interpretation that impact was slight. Those who responded that PPE had "some" impact or "significant" impact are assumed to be expressing greater concern about the impact of PPE on results and interpretation.

These 699 respondents were asked to discuss which aspects of performance-based testing with PPE have presented the greatest challenge. Of primary concern was examiner difficulty hearing the examinee due to a mask (35%), examinee difficulty hearing the examiner due to a mask (23%), and the physical barrier (e.g., plexiglass) causing difficulty with administration (15%). In the context of discussing masks and difficulty hearing one another clearly, respondents most

often mentioned auditory memory tasks such as WISC–V Digit Span (7%) and phonological processing tasks (5%). In the context of a plexiglass barrier or of social distancing, respondents also described problems administering tasks with manipulatives such as WISC–V Block Design (11%) or using stimulus books or response booklets (6%).

This same group of practitioners (*N* = 699) were asked to rate various examples of PPE and safety measures for potential to impact performance-based test results and interpretation. A four-point rating scale was assigned to provide an overall weighted average (e.g., ratings of "no potential impact" were assigned a rating of 1, those of "minimal impact," a rating of 2; those of "some impact," a rating of 3; and those of "significant potential impact," a rating of 4). Table 12 summarizes the percentages rating each measure at these various levels as well as the weighted average.

Table 12. May–August 2020 Q-interactive Practitioner Rating of PPE or Safety Measure Potential Impact on Performance-Based Test Results and Interpretation

PPE or safety measure	None	Minimal	Some	Significant	Weighted average
Conducting testing outside	10.6	25.1	40.9	23.4	2.8
Examiner using mask	7.7	33.5	44.5	14.3	2.7
Client using mask	6.0	36.6	42.7	14.7	2.7
Client using gloves	18.9	38.0	34.2	8.9	2.3
Plexiglass shield or other physical barrier	22.3	46.1	25.9	5.8	2.2
Client using face shield	17.8	50.9	27.1	4.3	2.2
Examiner using face shield	27.2	49.6	19.9	3.4	2.0
Client using eye protection	27.9	48.9	20.1	3.1	2.0
Examiner using gloves	36.1	42.8	18.1	3.0	1.9
Protecting paper stimuli with lamination or other methods	37.4	46.8	13.8	2.0	1.8
Examiner using eye protection	42.7	44.0	12.1	1.2	1.7

Note. The most frequent rating endorsed for each PPE or safety measure appears in bold font.

As shown in Table 12, the average impact of several PPE/safety measures was rated as greater than the value of 2 (i.e., the weighting assigned for "minimal"). Practitioners expressed the greatest concern about the impact of conducting testing outside on results and interpretation. The most frequently used aspects of PPE used in Q-interactive evaluations during May–August 2020 were examiner mask and examinee mask (> 90% of Q-interactive evaluations involving PPE as shown in Table 10). The results in Table 12 indicate close to 60% of practitioners rate the use of masks as potentially having some to significant impact on results.

Hypotheses

We anticipated that both achievement and cognitive test scores would be consistently lower in the May–August 2020 sample than in the May–August 2019 sample for various reasons. First, the aforementioned projections of lower scores on classroom-based tests due to school interruption in spring 2020 (Kuhfeld et al., 2020a) suggested achievement test scores may be lower in 2020. Second, nearly half of practitioners who engaged in performance-based testing in May–August 2020 indicated that they limited performance-based testing to only the cases with the most serious need. Thus, May–August 2020 scores were expected to reflect more severe clinical impairment than those of May–August 2019. Third, practitioners who used PPE during Q-interactive performance-based testing indicated a very high rate of mask use on the part of both examiner and examinee (i.e., > 90% indicated testing with both examiner and examinee

wearing a mask). Practitioners who participated in the survey indicated that examiner and examinee masks are the two aspects of PPE with the greatest potential to impact performance-based test results.

Data Analysis Plan

Cumulative distribution function (CDF) plots of the composite score distributions for the two time periods of interest (May–August 2019 and May–August 2020) were visually inspected to evaluate the nature of the distributions. The means, standard deviations, and standard differences of subtest and composite scores were examined to evaluate the effect sizes of any mean differences across years.

Results

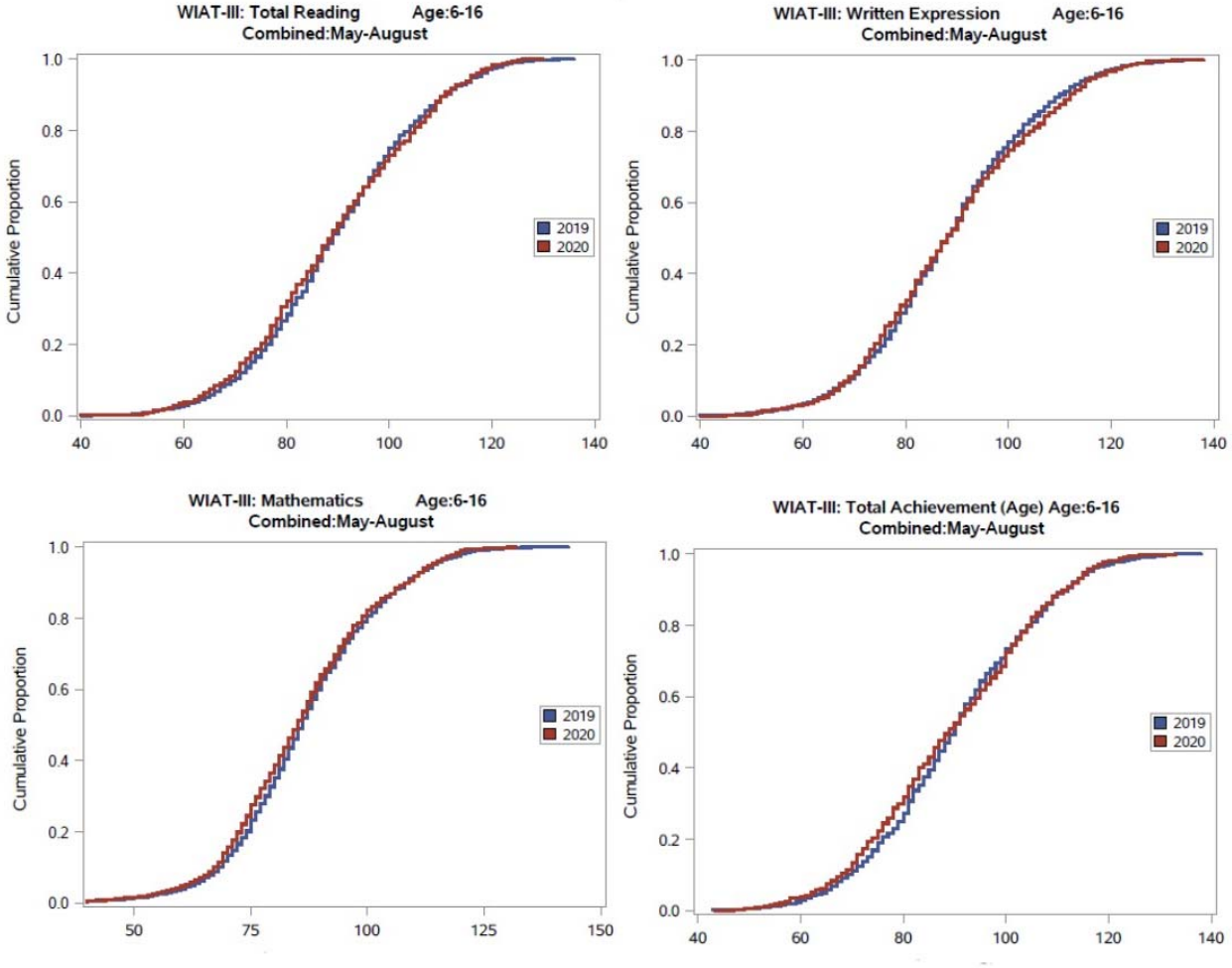
Age-based standard scores are reported for the WIAT–III and KTEA–3 for consistency in reporting results alongside the WISC–V. Results using grade-based standard scores for the WIAT–III and KTEA–3 were similar to those reported here for age-based scores.

Several conventions for reporting psychometric results are followed. The term standard difference refers to Cohen's *d*. Cohen's original suggestions for effect size descriptions indicated that .20 is characterized as small, .50 as moderate, and .80 as large (Cohen, 1988, 1992). Although these ranges do not fully describe all aspects of effect size interpretation, for the purposes of simplicity and of consistency with the technical and interpretive manuals of these tests, values for Cohen's *d* that range from .20 to .49 are reported as small effect sizes. Values that range from .50 to .79 are reported as moderate effect sizes, and values of .80 or greater are reported as large effect sizes.

WIAT–III

The May–August 2019 and May–August 2020 WIAT–III composite CDF plots appear in Figure 1, overlaid for visual inspection to determine similarities and differences that are present between the two sets of data. The blue lines represent the May–August 2019 distribution of composite scores and the red lines represent the May–August 2020 distribution for the same months. The composite scores appear on the horizontal axis. The vertical axis represents the cumulative proportion of examinees obtaining scores equal to or lower than that particular score.

Figure 1. WIAT–III Core Composite Score Density Plots



As shown, the plotted lines are highly similar for each composite score. These data suggest that the May–August 2019 and May–August 2020 scores are distributed similarly across the range of possible scores. There is not an observable difference in proportions of examinees scoring at lower levels as we predicted.

Table 13 presents the mean WIAT–III composite and subtest scores for the May–August 2019 and May–August 2020 samples.

Table 13. WIAT–III Performance in May through August 2019 and May through August 2020

Composite/ Subtest score	2019			2020			Standard difference[a]
	Mean	*SD*	*N*	Mean	*SD*	*N*	
Total Reading	89.4	15.7	4,033	88.4	16.1	1,576	-.06
Basic Reading	90.0	16.6	5,655	88.7	16.7	2,493	-.08
Written Expression	88.1	15.5	3,234	87.4	15.9	1,210	-.04
Mathematics	86.7	15.7	6,347	85.0	15.5	2,590	-.11
Reading Comprehension & Fluency	89.4	15.0	3,936	89.0	15.0	1,502	-.03
Oral Language	91.4	17.0	2,661	88.9	17.3	1,531	-.15
Total Achievement	89.6	15.9	1,424	87.9	16.2	580	-.11
Word Reading	90.4	18.0	6,713	90.0	18.4	2,973	-.02
Reading Comprehension	90.8	16.1	6,617	90.1	15.6	2,783	-.05
Pseudoword Decoding	90.4	16.8	5,808	88.2	16.5	2,584	-.13
Oral Reading Fluency	88.7	15.6	5,026	87.9	15.8	1,992	-.05
Early Reading Skills	84.2	15.7	1,805	79.1	15.8	653	**-.33**
Spelling	87.7	15.5	5,722	86.0	16.1	2,627	-.11
Sentence Composition	88.9	17.0	4,531	86.5	17.1	1,919	-.14
Essay Composition	91.4	18.0	3,093	92.4	17.2	1,175	.06
Alphabet Writing Fluency	91.0	14.7	1,543	84.7	14.3	569	**-.43**
Math Problem Solving	86.8	17.3	6,906	85.0	17.0	2,819	-.10
Numerical Operations	88.6	15.7	7,030	86.6	15.4	3,017	-.13
Math Fluency	86.5	14.8	3,383	83.5	14.3	1,440	**-.20**
Math Fluency-Addition	86.9	15.5	3,561	83.5	15.1	1,497	**-.22**
Math Fluency-Subtraction	86.7	15.3	2,661	84.4	14.5	1,142	-.15
Math Fluency-Multiplication	87.0	15.0	3,506	84.9	14.7	1,483	-.15
Listening Comprehension	94.3	16.6	3,701	93.2	17.0	2,033	-.07
Oral Expression	90.4	16.7	2,732	87.7	17.2	1,578	-.16

[a] The Standard Difference is the difference of the two test means divided by the square root of the pooled variance, computed using Cohen's (1996) Formula 10.4.

As shown, the mean differences in composite scores across the two time periods are negligible. None produced an effect size that approached .20.

Among the subtests, small negative effect sizes were present for the mean differences of Alphabet Writing Fluency, Early Reading Skills, Math Fluency, and Math Fluency-Addition. The Math Fluency-Addition mean differences were examined by age (i.e., 6–8, 9–11, 12–16) to determine if results differed by age group. The effect sizes were small among children aged 6–8 and 9–11, and negligible for the oldest age group. The sample sizes were insufficient to examine the mean score differences of Alphabet Writing Fluency and Early Reading Skills separately by age.

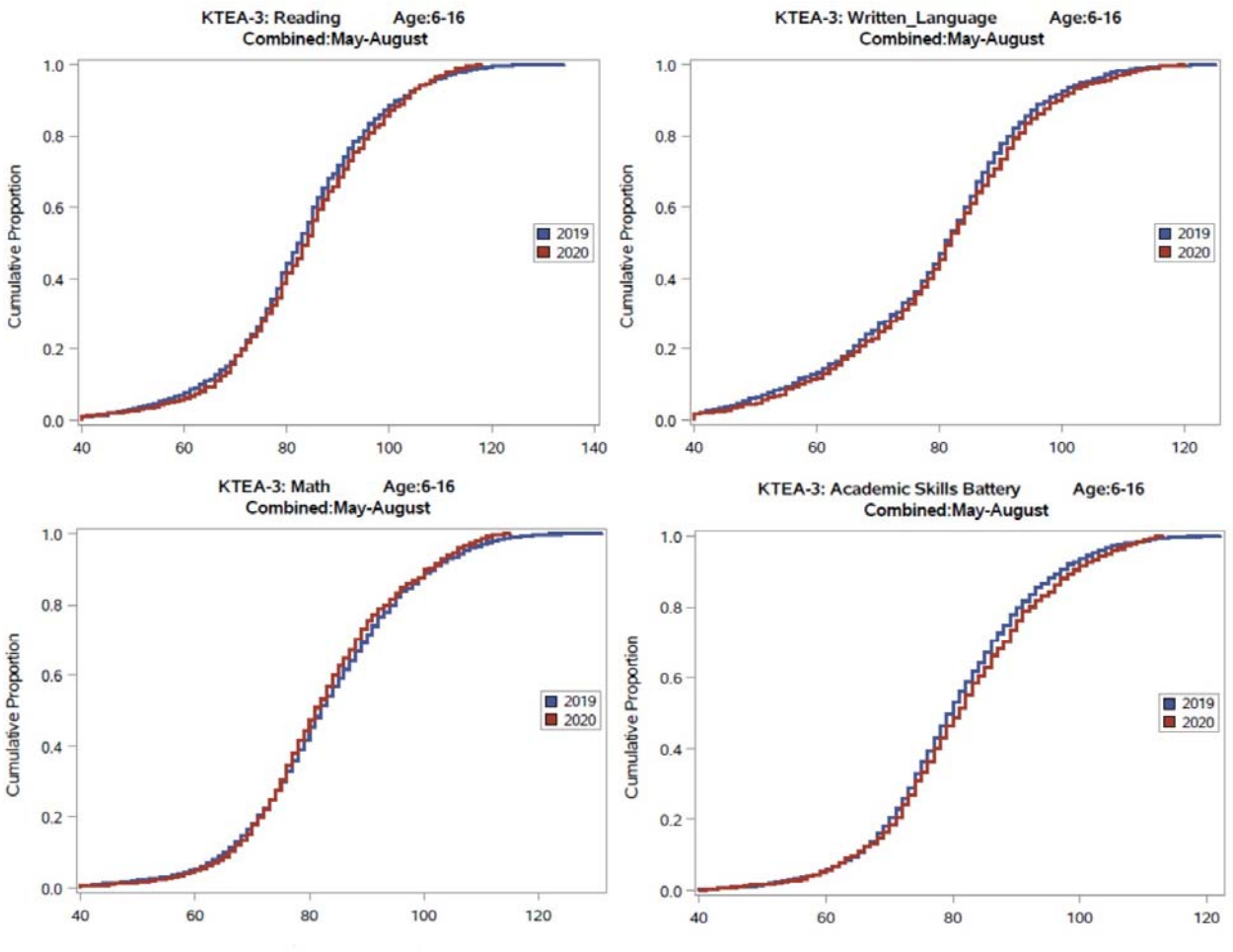
The most frequently administered WIAT–III subtests were those measuring math computation, word reading, math problem solving, reading comprehension, and spelling skills.

KTEA–3

The May–August 2019 and May–August 2020 KTEA–3 composite score CDF plots appear in Figure 2. The blue lines represent the May–August 2019 distribution of composite scores and

the red lines represent the May–August 2020 distribution. The composite scores appear on the horizontal axis, and the vertical axis represents the cumulative proportion of examinees obtaining scores equal to or lower than that particular score.

Figure 2. KTEA–3 Composite Score Density Plots



As shown, the plotted lines are highly similar for each composite score. These data suggest that the May–August 2019 and May–August 2020 scores are distributed similarly across the range of possible scores. There is not an observable difference in proportions of examinees scoring at lower levels as we predicted.

Table 14 presents the mean KTEA–3 core composite and subtest scores for the May–August 2019 and May–August 2020 samples.

Table 14. KTEA–3 Performance in May through August 2019 and May through August 2020

Composite/ Subtest score	2019			2020			Standard difference[a]
	Mean	*SD*	*N*	Mean	*SD*	*N*	
Reading	82.5	15.3	5,758	83.4	14.9	1,640	.06
Math	82.7	14.5	5,534	82.1	13.5	1,482	-.04
Written Language	79.2	16.0	3,592	80.3	16.0	1,065	.07
Academic Skills Battery	80.2	12.9	3,131	81.3	13.4	845	.15
Letter & Word Recognition	84.0	16.1	6,654	86.5	16.2	2,211	.15
Reading Comprehension	83.6	15.3	6,219	84.0	14.8	1,735	.03
Nonword Decoding	84.4	12.5	3,774	85.1	12.2	1,310	.06
Phonological Processing	86.2	15.8	2,462	86.3	14.9	799	.01
Word Recognition fluency	83.4	14.6	1,900	84.6	15.4	828	.08
Decoding Fluency	83.9	15.2	1,181	85.5	15.0	529	.10
Silent Reading Fluency	88.2	14.3	2,659	90.1	14.6	840	.13
Reading Vocabulary	86.2	14.6	1,343	85.4	14.1	332	-.06
Math Concepts & Applications	83.7	15.3	6,001	84.2	14.7	1,726	.03
Math Computation	84.7	15.8	5,967	84.3	15.5	1,735	-.02
Math Fluency	87.8	14.8	2,127	87.1	14.9	798	-.05
Written Expression	79.1	17.6	4,691	79.7	17.1	1,294	.04
Spelling	81.7	16.8	4,828	82.5	16.8	1,647	.05
Listening Comprehension	87.3	15.3	1,808	90.1	14.7	602	.19
Writing Fluency	88.2	17.8	1,271	86.6	16.9	566	-.09
Oral Expression	79.9	18.0	708	83.9	18.2	188	**.23**
Associational Fluency	94.2	20.8	871	96.4	19.2	217	.11
Object Naming Facility	87.7	14.6	1,021	88.7	13.1	352	.07
Letter Naming Facility	85.9	15.2	1,172	84.4	13.8	456	-.10

[a] The Standard Difference is the difference of the two test means divided by the square root of the pooled variance, computed using Cohen's (1996) Formula 10.4.

As with the WIAT–III, the mean differences in KTEA–3 composite scores across the two time periods are negligible. None produced an effect size that approached .20, and only the Math composite score difference was in the predicted direction (i.e., lower in May–August 2020 than in May–August 2019). The other means of composite scores were slightly higher in May–August 2020 than in May–August 2019.

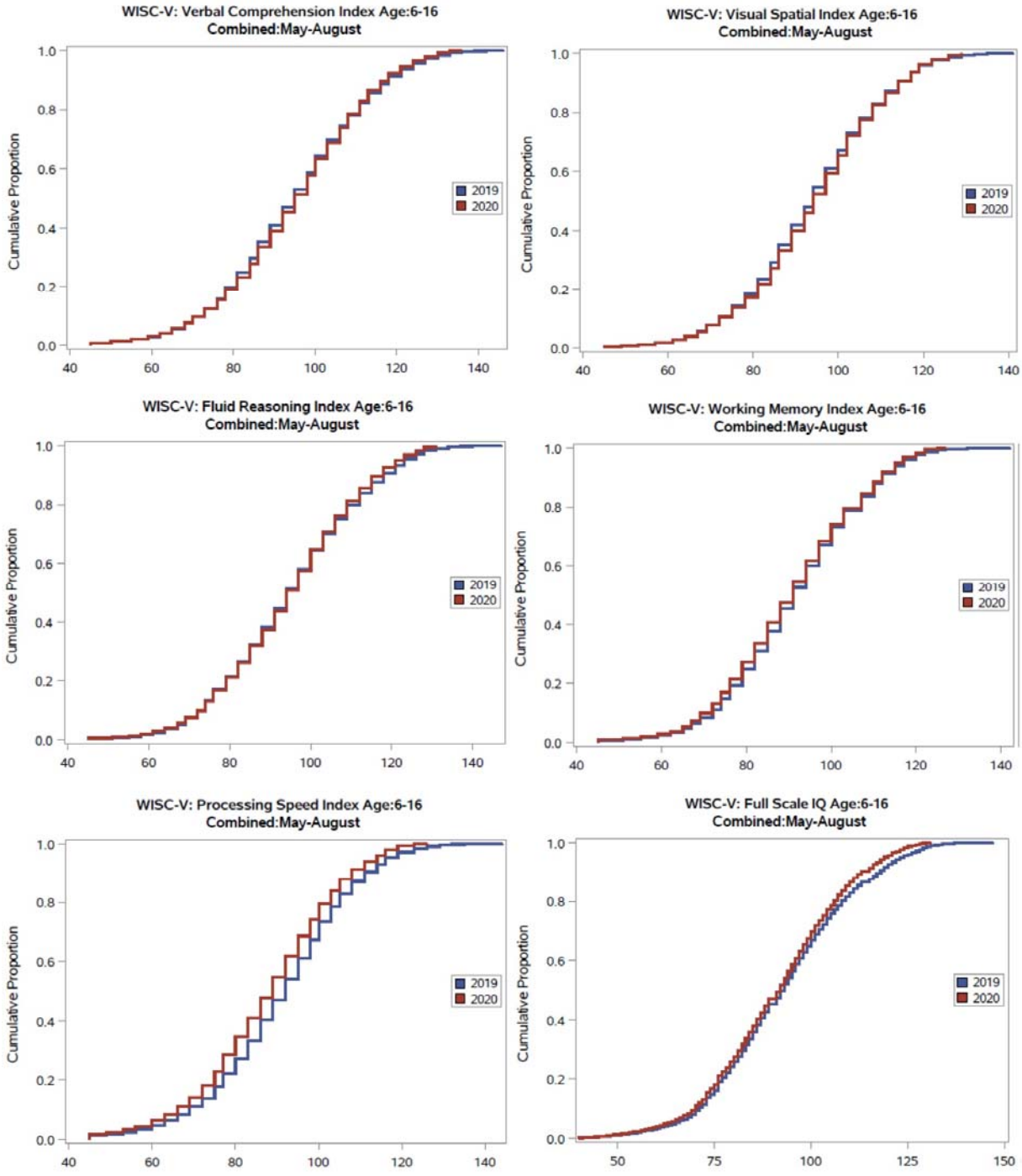
Among the subtests, all effect sizes are negligible with the exception of the KTEA–3 Oral Expression subtest, which only slightly exceeded .20. Notably, this difference across years was positive rather than negative: May–August 2020 scores were higher than May–August 2019 scores. This result can be contrasted with the WIAT–III Oral Expression mean difference, which was nearly as large as the KTEA–3 mean difference but in the opposite direction. The small effect for Math Fluency that was observed on the WIAT–III is not present for the KTEA–3 Math Fluency subtest.

The most frequently administered KTEA–3 subtests were those measuring word reading, reading comprehension, math problem solving, math computation, and spelling skills.

WISC–V

The 2019 and 2020 WISC–V primary composite score density plots appear in Figure 3. The blue lines represent the 2019 distribution of composite scores and the red lines represent the 2020 distribution for the same months. The composite scores appear on the horizontal axis, and the vertical axis represents the cumulative proportion of examinees obtaining scores equal to or lower than that particular score.

Figure 3. WISC–V Primary Composite Score Density Plots



As shown, the plotted lines are highly similar for each composite score. These data suggest that the 2019 and 2020 scores are distributed similarly across the range of possible scores. Of all composite scores, the Processing Speed Index mean score in 2020 appears to diverge most from that of 2019; a greater number of examinees appear to have obtained lower scores in 2020, although the two distributions are very similar.

Table 15 presents the mean WISC–V subtest and composite scores for the 2019 and 2020 samples.

Table 15. WISC–V Performance in May through August 2019 and May through August 2020

	2019			2020			
Subtest/ Composite score	**Mean**	***SD***	***N***	**Mean**	***SD***	***N***	**Standard difference[a]**
Verbal Comprehension Index	94.9	17.9	21,797	95.0	17.5	9,413	.01
Visual Spatial Index	93.7	16.0	18,815	94.1	15.8	8,125	.03
Fluid Reasoning Index	95.1	17.4	22,010	94.7	16.8	9,456	-.02
Working Memory Index	91.6	15.8	18,610	90.6	15.9	8,143	-.06
*Processing Speed Index	91.4	15.7	2,556	87.1	16.1	5,733	-.27
Full Scale IQ	92.9	18.0	20,937	91.1	17.1	8,191	-.10
Similarities	8.9	3.5	21,416	8.7	3.4	9,549	-.04
Vocabulary	9.2	3.5	21,299	9.3	3.5	9,393	.04
Information	8.6	3.3	2,425	8.8	3.4	1,085	.06
Comprehension	9.3	3.8	1,158	9.0	3.4	592	-.07
Block Design	9.0	3.2	21,678	8.8	3.1	8,936	-.05
Visual Puzzles	8.9	3.2	18,366	9.1	3.2	8,605	.07
Matrix Reasoning	9.0	3.5	22,032	8.9	3.4	9,515	-.03
Figure Weights	9.3	3.2	21,428	9.2	3.2	9,352	-.03
Picture Concepts	8.5	3.1	1,563	8.2	3.2	622	-.08
Arithmetic	8.1	3.1	918	7.9	3.0	504	-.07
Digit Span	8.1	3.1	21,553	7.9	3.0	9,470	-.09
Picture Span	9.1	3.2	18,218	8.8	3.2	8,217	-.07
Letter-Number Sequencing	7.9	2.9	1,465	7.7	2.8	527	-.08
*Coding	8.1	3.1	2,093	7.3	3.2	6,576	-.25
Symbol Search	8.2	3.2	18,092	8.0	3.1	7,924	-.07
Naming Speed Literacy	88.9	15.6	1,742	86.1	15.6	700	-.18
Naming Speed Quantity	91.8	14.6	1,454	89.1	14.4	707	-.18
Immediate Symbol Translation	93.6	12.2	1,126	91.5	11.9	361	-.17
Delayed Symbol Translation	94.2	12.2	539	92.7	12.2	199	-.12
Recognition Symbol Translation	96.5	13.5	487	93.8	14.1	185	-.20

[a] The Standard Difference is the difference of the two test means divided by the square root of the pooled variance, computed using Cohen's (1996) Formula 10.4.

Note. Due to issues with the WISC–V Coding subtest in digital format in 2019, an a priori decision was made to exclude the data associated with that format from the 2019 analyses for the Coding subtest scaled scores and for the Processing Speed Index. Hence, only data associated with presentation of the test in paper format (i.e., response booklet) is included.

As shown, the mean differences in composite scores across the two time periods are generally negligible. Only the Processing Speed Index standard difference produced an effect size that exceeded .20 and thus was substantial enough to be described as a small effect (Cohen, 1988, 1992). For the remaining composite scores, the average score differences were essentially indistinguishable from May–August 2019 to May–August 2020.

It should be noted that although the Coding and Processing Speed Index analyses excluded data based on the Coding digital scores, *the analyses for the Full Scale IQ did not exclude data based on the Coding digital scores because the Coding scaled score only contributes 1/7 to the Full scale sum of scaled scores*. If those data based on the digital format of the Coding subtest were excluded, however, the average May–August 2019 Full Scale IQ would be slightly lower, and the observed standard difference would be slightly smaller.

The three mean differences with small effect sizes (i.e., Processing Speed Index, Coding, and Recognition Symbol Translation) were examined by age (i.e., 6–8, 9–11, 12–16) to determine if results differed by age group. No clear pattern was present. The effect sizes of the mean differences were small in all cases except for Recognition Symbol Translation for children aged 12–16, for whom the effect size was negligible.

Although small in magnitude, the largest subtest effect sizes occurred on Coding and Recognition Symbol Translation. These subtests, which measure cognitive processes known to be clinically sensitive to many conditions assessed with the WISC–V (e.g., specific learning disorder, ADHD, autism spectrum disorder; Wechsler, 2014b; Raiford et al., 2015, 2016), trended toward lower scores in 2020.

Discussion

The aim of this study was to use available data obtained from convenience samples to provide preliminary information about how clinical practice and test scores may have shifted following the onset of the pandemic and subsequent educational disruption. The samples are not randomly selected or demographically matched, and for these reasons, the results should be considered preliminary. Data included survey responses from practitioners and observed test score data in Q-interactive. Score distributions and means obtained from selected clinical assessments administered in May–August 2020, following educational interruption during the COVID-19 outbreak, were compared with those from May–August 2019.

Fewer tests were administered on Q-interactive in May–August 2020 as compared to May–August 2019. It is important to note that the spring 2020 school closures affected almost all students in U.S. public and private schools aged 6–16. It is therefore highly unlikely that the May–August 2020 data reflects performance of only students who did not experience educational disruption, as the 2020 sample sizes were 33–47% those of May–August 2019.

Despite fewer tests being administered, the composite score distributions were highly similar. Effect sizes of the mean differences for scores obtained from individually administered, academic achievement and cognitive ability tests across these two time periods were almost universally negligible. We expected differences for a number of reasons. First, projections from large scale classroom assessment data predicted lower fall 2020 testing performance following the widespread school interruption in spring 2020. Second, about half of the practitioners who used the Q-interactive platform during May–August 2020 and responded to our survey indicated that they limited performance-based testing during this time period to clients with the most serious need. Finally, these practitioners reported very frequent usage of PPE during evaluations and indicated that they believed the use of PPE would negatively impact results.

For the two individually administered measures of academic achievement, the composite score distributions were highly similar, the composite score level differences were negligible, and only a few subtests of the WIAT–III showed small negative effects when mean scores were compared across May–August 2019 and May–August 2020. These few subtest-level differences were not replicated on the KTEA–3; in fact, the KTEA–3 mean subtest scores were generally slightly higher in May–August 2020 relative to May–August 2019.

For the WISC–V (i.e., cognitive ability measure), the composite score distributions were also highly similar and the mean score differences were negligible; only processing speed showed a small decrease relative to May–August 2019. Processing speed is generally the most clinically sensitive of all WISC–V primary index scores (Wechsler, 2014b; Raiford et al., 2015, 2016), and nearly half of Q-interactive practitioners surveyed indicated they limited performance-based testing in May–August 2020 to clients with the most serious need. Thus, more clinically sensitive index scores were expected to be slightly lower in 2020. In support of this explanation, there was a noticeable trend among the WISC–V scores toward larger differences (while still small) on two of the most clinically sensitive and diagnostic subtests (e.g., Coding and Recognition Symbol Translation). *It is important to note that an a priori decision was made to exclude the data associated with Coding in digital format from all analyses, so this difference is* ***not*** *due to technology but reflects small performance differences in the Coding subtest in paper format (i.e., response booklet) across the two timeframes (i.e., May–August 2019 and May–August 2020).* These results may reflect the trend for practitioners to prioritize performance-based testing for clients with the greatest clinical need, thus those with more clinically severe problems.

Interestingly, the mean difference between the two time periods for the WISC–V auditory working memory tasks (i.e., Digit Span and Letter-Number Sequencing) was negligible. The absence of meaningful difference was surprising because of the widespread use of masks and because many practitioners who completed the survey had the impression that mask use interfered with the examinee and examiner hearing one another and may result in performance differences on auditory memory tasks.

Limitations and Future Directions

It is important to note the limitations of the present study. First, this study provides an omnibus test to examine whether scores on individually administered clinical tests, in the presence of significant disruption to the field of clinical assessment, are largely similar to those of the prior year. School interruption and remote instruction were widespread in the spring of 2020. Since that time, furthermore, practitioners have commonly engaged in a number of accommodations during the testing process to increase safety (e.g., use of PPE and other safety measures, tele-assessment). Due to the nature of the data, it is only possible to obtain general information about these trends, and not to examine the impact of any single one of these issues. Despite these limitations, however, the observed differences in scores and composite score distributions across May–August 2019 and May–August 2020 from individually administered achievement and cognitive ability tests are negligible.

Second, it is important to note that these results are limited to a referred population. Survey results indicated that practitioners are not foregoing any specific type of clinical assessment other than autism spectrum disorder evaluations using one common test that cannot be administered while the examiner and examinee wear masks. However, it is possible that there are other differences in the clinical conditions for the population tested in May–August 2020 versus May–August 2019. Furthermore, no information is available regarding the examinees' clinical conditions or family/educational circumstances during the pandemic. As is always the case in clinical assessment, it is important to consider the impact of historical and current life situations on an examinee's performance. Because almost half of the practitioners indicated that they were engaging in performance-based testing with only the clients in most severe need, we anticipated differences in test scores for the May–August 2020 sample, but none were evident.

Third, the samples were not randomly selected or matched; rather, they consist of the entire population of referred examinees tested on these measures using Q-interactive. It is also important to note that the sample sizes vary by year. Only the examinee's age is known. It is possible that the samples differ on other demographic variables (e.g., socioeconomic status, parent education level, race/ethnicity, sex) that are known to explain some variability in academic achievement and cognitive ability scores. However, because these are referred populations taking clinical diagnostic tests, the severity of the clinical condition can have a greater impact on scores than demographic variables. For example, the special group studies in the WIAT–III, KTEA–3, and WISC–V technical manuals show that the scores often differ across clinical and demographically matched, nonclinical control samples.

Finally, it may be helpful to extend the analyses in several ways. Results by age and month are not included. In addition, results from the fall of 2020 were not yet available when this paper was written. Future research will replicate and expand upon the analyses to examine these data in greater detail and provide insight into any possible effects over time.

Questions have been raised about the impact of school closures and remote learning on students' educational achievement levels. On an individual basis, these are important questions that warrant careful consideration in the context of a clinical evaluation. Because the present study compared the academic achievement levels of two different samples, the results do not indicate whether the academic achievement levels of individual examinees changed over time. However, examining sample differences across time periods is useful for detecting systematic differences in educational achievement levels, and none were found. One possible interpretation is that the educational disruption that occurred in 2020 may have differentially impacted certain individuals more than others. For example, students with a low SES background with more limited access to assessment resources (and the resources necessary to support remote learning) may not have been well-represented in our sample.

It is interesting to note that, counter to expectations, the May–August 2020 results map onto those of May–August 2019 very closely despite any shifts in clinical assessment practices that are related to the COVID-19 pandemic. Given that no obvious or consistent differences in means or composite score distributions were found in the present study, these results provide preliminary information that suggests that the normative data provided for clinical assessments continue to provide a valid and appropriate reference point for score interpretation. Guidance is provided to help practitioners interpret clinical, norm-referenced academic achievement results during or after a significant educational disruption (see Breaux, 2020).

References

Breaux, K. (2020). Excerpt from the Wechsler individual achievement test technical and interpretive manual (4th ed.): Special Considerations for Score Interpretation Following Educational Disruption. Retrieved from https://www.pearsonassessments.com/content/dam/school/global/clinical/us/assets/telepractice/special-considerations-for-score-interpretation.pdf

Cohen, B. H. (1996). *Explaining psychological statistics.* Brooks & Cole.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.

Cohen, J. (1992). Quantitative methods in psychology: A power primer. *Psychological Bulletin, 112*(1), 155–159.

Education Week. (2020). *Map: Coronavirus and school closures*. Retrieved from Education Week https://www.edweek.org/ew/section/multimedia/map-coronavirus-and-school-closures.html

Kaufman, A. S., & Kaufman, N. L. (2014). *Kaufman test of educational achievement* (3rd ed.). NCS Pearson.

Kuhfeld, M., Soland, J., Tarasawa, B., Johnson, A., Ruzek, E., & Liu, J. (2020a). *Projecting the potential impacts of COVID-19 school closures on academic achievement*. (EdWorkingPaper: 20-226). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/cdrv-yw05

Kuhfeld, M., Soland, J., Tarasawa, B., Johnson, A., Ruzek, E., & Lewis, K. (2020b). *How is COVID-19 affecting student learning? Initial findings from fall 2020*. Retrieved from https://www.brookings.edu/blog/brown-center-chalkboard/2020/12/03/how-is-covid-19-affecting-student-learning/

National Association of School Psychologists. (2020). *The Pandemic's Impact on Special Education Evaluations and SLD Identification*. Retrieved from https://www.nasponline.org/resources-and-publications/resources-and-podcasts/covid-19-resource-center/return-to-school/the-pandemics-impact-on-special-education-evaluations-and-sld-identification

Pearson. (2009). *Wechsler individual achievement test* (3rd ed.). NCS Pearson.

Raiford, S. E., Drozdick, L. W., & Zhang, O. (2015). *Q-interactive special group studies: The WISC–V and children with autism spectrum disorder and accompanying language impairment or attention-deficit/hyperactivity disorder* (Q-interactive Technical Report 11). Pearson. Retrieved from https://www.pearsonassessments.com/content/dam/school/global/clinical/us/assets/wisc-v/q-i-tr11-wisc-v-adhd-autism.pdf

Raiford, S. E., Drozdick, L. W., & Zhang, O. (2016). *Q-interactive special group studies: The WISC–V and children with specific learning disorders in reading or mathematics* (Q-interactive Technical Report 12). Pearson. Retrieved from https://www.pearsonassessments.com/content/dam/school/global/clinical/us/assets/wisc-v/wisc-v-learning-disorders-reading-math.pdf

United States Census Bureau. (2020). *Children characteristics: 2019 American Community Survey 1-year estimates subject tables*. Retrieved from U.S. Census Bureau

https://data.census.gov/cedsci/table?q=children%20in%20the%20U.S.&tid=ACSST1Y2019.S0901&hidePreview=false

Wechsler, D. (2014a). *Wechsler intelligence scale for children* (5th ed.). NCS Pearson.

Wechsler, D. (2014b). *Wechsler intelligence scale for children technical and interpretive manual* (5th ed.). NCS Pearson.