Office of
Inspector General
Chicago Board of Education
Nicholas Schuler, Inspector General

SUMMARY REPORT

OIG 18-01670

September 26, 2019 (Revised October 18, 2019)

PERFORMANCE REVIEW: NWEA TEST ADMINISTRATION

SYNOPSIS

Every spring, CPS students take a test produced by the Northwest Evaluation Association that carries important stakes for students, teachers, principals and schools. In addition, school and district officials make curricular decisions that can involve CPS resources based on these exams. Thus, NWEA tests are so integral to so many aspects of CPS that the accuracy of their results is paramount.

A performance review by the CPS Office of Inspector General found a concerning level of unusually long test durations, high pause counts and other irregularities during CPS's Spring 2018 administration of this untimed, computer-based test. This occurred in a minority of cases, but enough to be worrisome and to warrant action.

The OIG's Performance Analysis Unit found that tens of thousands of students are taking at least twice the national average duration to complete their NWEAs; some are taking three, four and five times. At some CPS schools, a test that the average student nationally completes in roughly an hour has turned into a multi-day or even a week-long event. Excessive durations can make it difficult to accurately compare CPS results to national norms, NWEA warned the OIG. CPS's average durations, in every grade and subject, have been above national norms since at least 2016 and have increased in each of the two years since. Thus, if no action is taken, CPS durations could continue to grow, putting CPS results at increasing risk.

Longer durations could occur for benign reasons but OIG interviews of students and teachers at schools with unusual results produced reports of a variety of improper testing procedures that could have added to durations. This included everything from attempts to game the test to coaching to outright cheating.

An April 2018 CPS audit called for stronger NWEA controls. Despite some resulting reforms, current security procedures remain inadequate. Therefore, the OIG recommends that CPS, with OIG input, hire a test security expert to help CPS address the many concerns outlined in this Summary Report.

Table of Contents

Appendices:

CASE INITIATION

This investigation followed numerous complaints to the OIG over the years about alleged NWEA cheating. Many of these complaints were anonymous. However, at least one teacher contended the massive jump in NWEA scores one of her students experienced the previous year must have been the result of cheating and voiced concern that her teacher evaluation would suffer as a result.

FINDINGS

The OIG makes the following findings based on a performance review of CPS results in third- through eighth-grade NWEA Reading and Math tests given in the spring of 2017 and the spring of 2018, as well as additional NWEA data, OIG research and OIG interviews:

1. **NWEA is so integral to so many aspects of CPS that the accuracy of its results is paramount.** NWEA is an untimed, adaptive, computer-based test with high-stakes for many different CPS parties. Its results impact, to varying degrees: student promotions in third, sixth and eighth grades; for seventh graders, admission to selective-enrollment high schools and selective programs; the evaluations of Math and Reading teachers; principal evaluations; Independent School Principal status; and school SQRP levels. NWEA results also help drive curricular decisions on the school and district level that can involve CPS resources.

2. **In the spring of 2018 alone, tens of thousands of CPS students took far longer than the national norm to complete their NWEA tests**. Unusually long durations can occur for many benign reasons, including the high-stakes nature of many CPS tests, but they also can be an indicator of cheating or of attempts to game the test. In 2018, nearly 83,000 tests (more than one out of four) took double the national norm to complete and more than 24,000 took at least three times. Many unusually long tests were clustered in certain schools and certain grades within those schools. Some students in schools with unusually long durations described to the OIG improper testing procedures that would have affected durations, including proctors who allotted students a certain number of pauses (and the new questions they produced) per question number or per test day; teachers who told students to write down both the questions and the answers of difficult questions and then collected that information, sometimes for use in later lessons; and proctors who read aloud questions on the Reading test, rephrased questions or signaled to students when their answers were wrong. In addition, the OIG was told that some students were lingering over questions

until they "timed out" after 25 minutes, resulting in new questions once proctors resumed the tests. Diverse Learners were less likely to have long tests than non-Diverse Learners.

3. **Some CPS tests displayed unusually high pause counts.** ~~Close to 11~~More than 12,000 tests were paused at least five times and more than 1,~~500~~600 were paused at least 10 times, an OIG analysis showed. Tests with high pause counts often were clustered in certain schools. NWEA warned that "the validity of the assessment can be compromised if tests are paused for the purpose of producing a new question." Under such circumstances, students "have a 50 percent chance of getting a new question that might be more favorable," with those chances increasing "considerably" if the same question number is paused multiple times, NWEA told the OIG. Students in schools with high average pause counts described ways in which proctors paused tests so students could skip hard questions, or students themselves allowed questions to "time-out" after 25 minutes of inactivity. Although Diverse Learners are often allowed frequent breaks as a testing accommodation, the tests with the heaviest pause counts were predominantly taken by non-Diverse Learners.

4. **CPS NWEA tests with longer durations and higher pause counts were more likely to show unusually strong growth.** This pattern held true with both Diverse Learners and non-Diverse Learners. Strong growth can occur for positive or benign reasons but, especially when combined with unusual levels of pauses or durations, it can result from improper testing procedures.

5. **Unusually long durations and excessive pauses can reflect conditions that undermine the integrity of test results.** According to NWEA experts, the integrity of the NWEA assessment can be compromised if tests are repeatedly paused for unacceptable reasons (such as an intentional attempt to replace one question with another). Large duration changes from one spring to the next can negatively impact the accuracy of a student's growth score. Unusually long durations can compromise CPS's ability to make accurate inferences from NWEA results because longer durations "make it difficult to compare the results to NWEA's norms," NWEA warned the OIG. In addition, according to a NWEA publication on high-stakes NWEA tests: "Manipulating the testing process can significantly undermine the accuracy of student test results and negatively influence decisions based on these results."

6. **CPS's longer-than-normal durations are getting worse over time.** In every grade, ~~second~~third through eighth, CPS's average NWEA Spring test durations were longer than the national duration norm in 2016 and increased in every grade

and subject even more in both 2017 and 2018. By the end of the three-year test period, average test durations, in every grade and subject, had jumped by double-digit percentages. Thus, if no action is taken, CPS durations could well continue to move farther and farther from national averages, putting CPS results tied to national norms at increasing risk.

7. **Even excessively long test durations that do not involve misconduct can be an inefficient use of student and teacher time**. Students described classmates sitting with their heads on their desks, sometimes for multiple days, while others tested. Some teachers said they could not teach any students until all their students had completed testing.

8. **CPS's NWEA tests are being administered with insufficient security protocols for such a high-stakes exam.** For example:

   a) Many Reading and Math teachers test their own students, with no additional proctor, even though NWEA specifically recommends that high-stakes NWEAs be given by a student's teacher as well as a second proctor who has no investment in the test's outcome.

   b) Neither CPS nor NWEA keep an auditable record of who proctored each student test, making it impossible to determine if unusual results are consistently appearing in tests administered by certain proctors. Although many students are tested by their Reading and Math teachers, some described being tested by someone unconnected to their classrooms or the subjects being tested, so CPS auditors cannot assume that a student's Reading teacher gave that student his or her Reading test.

   c) Auditors currently are flying blind when they visit a classroom based on unusual results the year before because they have no idea if the proctor they are observing was in the room the previous year when the unusual results occurred. This is not an efficient use of CPS resources.

   d) Checklists used by auditors currently do not require auditors to record when small-group testing is used. Some general education students described being tested in small groups — something not allowed in CPS NWEA rules — by proctors who broke testing protocols.

   e) CPS's current NWEA data file does not include key data needed to determine if results are unusual. This includes the number of pauses per test, the kind of pause (a time-out by a student or a pause by a proctor) and the number of test days.

f) Current CPS NWEA training and "exit slip" questionnaires that must be passed to proctor the NWEAs are inadequate. Importantly, they do not sufficiently cover both unacceptable as well as acceptable behavior. The OIG is not listed as an office that should be contacted, anonymously or not, with concerns about test irregularities.

g) The Test Security Agreement that all proctors must sign does not list the penalties for cheating, which can include termination.

9. **Insufficient action followed an April 2018 audit about NWEA protocols**. The OIG credits the CPS Departments of Student Assessment and School Quality Measurement and Research for proactively requesting the audit but CPS's response to it did not go far enough in producing reforms. Audit recommended a series of changes, but the OIG found that some key suggestions were never executed, were partially executed or were poorly executed. Despite new 2018 proctor training, the OIG found evidence of a host of 2018 testing irregularities.

RECOMMENDATIONS

1. Move toward a reduction in duration times, preferably by setting a reasonable time limit for general education students. CPS should analyze its durations annually to monitor this situation.

2. Take concrete steps to limit pauses and develop clear instructions to proctors and test administrators about the right and wrong way to use pauses.

3. Find an auditable way to record the proctor of each test taken, preferably as a test data field.

4. Use new proctor data to guide the selection of audit sites. Audits should be tied to those proctors whose classrooms produced unusual test results, not on a previous year's test results in a certain grade and subject with no clue as to whether the proctor being audited is the same proctor who was in the room the previous year when unusual results occurred.

5. Prohibit Reading or Math teachers whose evaluations are tied in part to their students' NWEA results from being the sole proctors of their students. One solution is adding a second proctor who has no stake in the test. One proctor should be held responsible for the integrity of the test session — preferably the proctor with no stakes in the test.

6. Bolster NWEA training and the exit slip that currently must be passed for a CPS staff member to proctor a test. Include advice on how to guard against improper

pauses and unusually long durations, among other things. During training, cite the OIG as an office that may be called or emailed, anonymously or not, with test irregularity concerns.

7. Insert penalties for test cheating in the Test Security Agreement all proctors must sign to give them fair warning and to help deter cheating. This was recommended by Audit in April of 2018 but never implemented.

8. With the OIG's assistance, hire a test security expert for help and guidance in addressing the above concerns and others, including: improving the current methods for identifying test results warranting audits and working with NWEA on enhanced data security features such as adding proctors, pauses, timeouts and number of days tested to NWEA's CPS data. NWEA's current option to renew, worth up to $2.2 million, ends June 30, 2020. This security expert should help CPS and NWEA address the concerns in this Summary Report in any renewal of NWEA's contract — or preferably by the Spring 2020 testing season. If NWEA cannot provide needed reforms, this expert should help CPS write a Request for Proposal for a new test vendor. The OIG should be involved in the hiring of a test security expert and be kept apprised of any contracting changes or RFPs for a new test vendor resulting from this report.

ACKNOWLEDGEMENTS

The OIG's Performance Analysis Unit gratefully acknowledges the cooperation and assistance of the following people, who contributed to this report in varying degrees:

- From NWEA: Jacob Carroll, Senior Director, Privacy/Information Security; John Cronin, Vice President of Education Research; and Jennifer Potter, General Counsel.
- From Caveon LLC: John Fremer, President of Caveon Consulting Services; Dennis Maynes, Chief Scientist; and Marc Weinstein, Vice President of Caveon Investigative Services and Chief Privacy Officer.
- From the University of North Carolina at Chapel Hill School of Education: Gregory Cizek, the Guy B. Phillips Distinguished Professor of Educational Measurement and Evaluation and past president of the National Council on Measurement in Education.
- From the Harvard University Graduate School of Education: Andrew Ho, the Charles William Eliot Professor of Education, a testing and measurement expert and currently a member of the National Assessment Governing Board.

- From CPS: Peter Leonard, Director of Student Assessment; Jeffrey Broom, Director of School Quality Measurement and Research; Linda Brown, senior business effect analyst, Internal Audit and Compliance; and Zipporah Hightower, executive director, Principal Quality.
- From CPS's value-added vendor, the ECRA Group: John Gatta, CEO, and Gina Siemieniec, President of Research and Analytics.
- From Other Testing Companies: Trent Workman, Pearson Vice President of School Assessment; Steve Kromer, President of AIR Assessment, American Institutes for Research.
- More than 30 CPS students, teachers and parents who were interviewed about NWEA tests by the OIG.

BACKGROUND

A. How the NWEA MAP Works

CPS's primary assessment measure for students in grades 2 through 8 is a test produced by the Northwest Evaluation Association, or NWEA, called the Measures of Academic Progress, or MAP, Growth test. This test is often referred to within CPS as simply "NWEA."

Since SY 2012-13, CPS schools have been required to administer this test each spring in Reading and Math. Fall and Winter NWEA exams are generally optional.

The NWEA MAP is a web-based, multiple-choice, computer-adaptive test, meaning it adapts to each student's ability level based on student responses. Correct answers are followed by harder questions and incorrect answers are followed by easier questions. The test is constructed so that students generally are expected to correctly answer about half of the questions posed at their achievement level. This adaptive feature makes it virtually impossible for students to copy answers from neighboring students because they likely are seeing different test questions.

A NWEA MAP Math test contains 52 or 53 questions while a Reading test contains 42 to 43 items. Usually two items are field questions that are being tried out on students but do not count in their scores.

B. Distinctive Features

Untimed — One distinctive feature of the test is that it is untimed, even for general education students.

A sample NWEA testing schedule CPS distributes to schools assumes most students will complete the test in an hour.

NWEA expects students will finish the test in about 45 to 75 minutes, with high-performing students taking longer in some cases, according to an NWEA education blog entitled "Testing Duration: How Long is Too Long to Spend on the MAP Growth Assessment?"

To guide school districts, NWEA publishes the average duration of each test, by grade and subject. Across grades 3 to 8 in Reading and Math, average Spring durations were around an hour long. To be precise, they ranged from 57.7 minutes to 70.7 minutes, based on all students who took the Spring 2017 NWEA MAP tests.

Pauses — A proctor can pause[1] the NWEA test while a question is still on the screen if a student needs a bathroom, lunch or wiggle break, according to NWEA. Once a proctor resumes the test from a break, a new question appears that is of "similar difficulty in the same area" as the last question on the screen, NWEA told the OIG. In addition, under certain circumstances,[2] a pause can generate both an entirely new Reading passage and a new Reading question. The pause feature is designed to ensure that breaks cannot be used to obtain answers to pending questions.

However, if an existing question that stumps a student is paused, a student has "a 50 percent chance of getting a new question that might be more favorable,'' according to NWEA.

And, "If they pause multiple times on the same question, that probability can be considerably higher."

Time-outs — In addition, after 25 minutes without an answer the test will "time-out" and send students back to the Login page. The proctor then must resume the test, resulting in a new question of similar difficulty. Mere movement of a computer mouse does not reset the time-out clock.

Disengagement Alert — A disengagement alert is triggered when a student answers three successive questions with rapid guesses. With this feature, an alert displays on

[1] A proctor can pause the test from the proctor console or at the student's screen or tablet. Hypothetically, students who knew the keyboard command could pause their own tests but they would need a PIN number from the proctor's screen to resume it. The OIG found no evidence that students were pausing or resuming their own tests.

[2] According to NWEA, the Reading test is designed to provide a student with a new question after a pause or a timeout related to the existing Reading passage. However, if another test question related to the existing passage is not available, the test will generate both a new passage and a new test question.

the proctor console, warning that a student is disengaged. NWEA recommends pausing the test and speaking with the disengaged student before resuming.

NWEA STAKES

Annual Spring NWEA tests play a critical role in CPS. They carry important stakes for a variety of CPS parties. CPS's use of the test for teacher evaluations is unusual in that of the 9,500 school districts and other clients that administer NWEA, only a minority of them use NWEA for teacher evaluations, one NWEA expert told the OIG.

A. For Students

Promotion Stakes — NWEA carries stakes for students in the third, sixth and eighth "benchmark" grades. Under the 2013 revision to the CPS Promotion Policy, benchmark students scoring at or above the 24th national percentile in NWEA Reading and Math who also have passing grades in Reading and Math are promoted to the next grade. Students scoring below the 24th percentile but at or above the 11th in NWEA Reading or Math and who have at least Cs in Reading and Math are promoted with supports. Those scoring at or below the 10th percentile in NWEA Reading or Math are referred to summer school.

Elementary Admission Stakes — NWEA scores are used to determine whether CPS students can apply to selective-enrollment elementary schools or programs in grades 5 to 9. To sit for the admissions test for seventh- and eighth-grade Academic Centers, student must pre-qualify by scoring at or above the 45th percentile in NWEA Reading and Math. To sit for the admissions test for Classical, International Gifted, and Regional Gifted Centers in grades 5 to 8, students must score at or above the 60th percentile in Reading and Math. Students with an IEP applying for selective-enrollment programs in grades 5 to 8 must score at or above the 50th percentile in either Reading or Math and at or above the 40th percentile in the other subject to be eligible for admissions testing. Final selection for Academic Centers and International Gifted Programs is tied to a point system, one third of which is based on NWEA Reading and Math results.

Selective-Enrollment High School Admission Stakes — A student's seventh-grade NWEA scores are critical to admission to the system's coveted selective-enrollment high schools. CPS general-education applicants to selective-enrollment high schools must score a minimum percentile of 24 on the seventh-grade NWEA in both Reading and Math to sit for the selective-enrollment admissions exam. In addition, those who sit for the exam are accepted in part based on their seventh-grade NWEA scores, which constitute a maximum 300 of a total possible 900 points.

Other High School Admission Stakes — Many high school programs, including the district's popular International Baccalaureate programs, use seventh-grade NWEA scores as part of their admissions process.

B. For Teachers

Teacher Evaluations — Some CPS teacher evaluations are tied to NWEA scores as part of the REACH (Recognizing Educators Advancing Chicago's Students) evaluation process.

For Reading, English and Math teachers in grades 3 to 8, including teachers of Diverse Learners, 20 percent of their REACH teacher evaluation is based on their students' NWEA growth from one Spring to the next.

Each student's NWEA growth is calculated based on a value-added model that is intended to measure the impact of an educator on the academic growth of his or her students and to adjust for factors beyond the teacher's control. According to experts from CPS's value-added vendor, the model looks at 10 different factors in projecting what a student's growth should be in any one NWEA test subject: two years of NWEA Reading and Math scores; one year of Illinois test scores in the same subject; and a student's Individual Education Program status, English Language Learner status, mobility, homelessness and income. The model assumes all test scores are accurate.

C. For Principals

Principal Evaluations — NWEA scores comprise 35 percent of a K-8 principal's evaluation, broken down in the following way:

- 20 percent is based on school-wide NWEA Reading and Math gains (10 percent per subject);
- 5 percent is based on the school-wide percent of students meeting or exceeding national average growth norms in Reading and Math; and
- 10 percent is based on priority-group gains in Reading and Math (5 percent per subject). The four priority groups[3] are: African American students, Hispanic students, English Learners and Diverse Learners.

ISP Applications — To be eligible to seek Independent School Principal status, a K-8 principal must meet various threshold requirements, including heading a school

[3] If a school has fewer than 30 members of a priority group, that priority group is not assessed separately and its weight is added to the overall growth metric.

with at least a 50 percentile NWEA score in both Reading and Math for the last three school years. Once all eligibility criteria is met, principals must meet nearly 50 additional criteria, with more than a dozen based on various NWEA scores and three based on SQRP levels, which are largely tied to NWEA on the elementary level. ISPs enjoy less Network and District oversight and may serve as mentors.

D. For Schools

SQRP Levels — Most CPS K-8 schools receive a School Quality Rating Policy level of from 1+ to 3, with 1+ being the highest, based largely on their NWEA results. These levels are closely watched by parents trying to decide where their children should apply — and sometimes even where their family should live.

As of the 2017-2018 SY,[4] which was the key school year analyzed by the OIG, 60 percent of a standard elementary school's level was tied to NWEA scores, broken down the following way:

- 25 percent: a school's Reading percentile growth and Math percentile growth in Grades 3-8 were worth 12.5 percent each.
- 10 percent: Up to 5 percent was based on priority group percentile growth in Math (1.25 % each for African American students, Hispanic students, English Learners and Diverse Learners). Additionally, up to 5 percent was based on priority group percentile growth in Reading.
- 10 percent was based on students meeting or exceeding national average growth norms in Reading and Math.
- 10 percent: a school's Grade 3-8 national attainment percentile in Reading and Math counted for 5 percent per subject.
- 5 percent: The Grade 2 Math and Reading attainment percentiles were each worth 2.5 percent.

Students Counted Multiple Times — Note that under the SQRP metric, some student scores can be counted multiple times. For example, a theoretical fifth-grade African American Diverse Learner who exceeded national average growth norms in math could contribute to 1) the school's math percentile growth, 2) its priority group math growth for African Americans, 3) its priority group math growth for Diverse Learners, 4) its percent of students meeting or exceeding national average math growth norms, and 5) its national attainment percentile in math.

[4] This weighting system was amended for the 2018-19 school year.

E. Curricular Decisions

District officials make curricular decisions — some of them involving CPS resources — based on NWEA data and trends so it is critical that such decisions be based on accurate data. Teachers and principals also use the test diagnostically to determine in what areas students need help.

METHODOLOGY

The OIG analysis focused on mostly Spring 2018 Reading and Math NWEA results of third through eighth graders. To calculate gains, the OIG also used second- through seventh-grade 2017 NWEA results.

A. Data Pulled

As part of its analysis, the OIG Performance Review Unit pulled from the CPS Data Warehouse various data fields from the "comprehensive data file" that NWEA provides to CPS. That included second- through eighth-grade scores from the Spring tests of 2016, 2017 and 2018.

However, the unit ultimately decided to target the most recent scores then available — third- through eighth-grade Spring 2018 scores — and the gains in scores from Spring 2017 to Spring 2018. The OIG started its focus with third grade because third is the first grade in which NWEA carries student promotion stakes.

Data points pulled from the CPS comprehensive data file for NWEA tests included: Student Names, Student Identification Numbers, Grades and Subjects tested, RIT[5] scores, National Percentile scores, Durations, Start Dates, Test Years, Home Rooms, School Names, School IDs, Test IDs.

The OIG inserted into this data other information from the Data Warehouse, including student Limited English Proficiency status; Diverse Learner status and a Student Food Service indicator.

Over many months, the OIG Performance Review Unit also received specially-requested custom reports directly from NWEA that included: the number of Fall 2017 and Spring 2016, 2017 and 2018 pauses by Test ID; the End Dates of each Test ID in Spring 2016, 2017 and 2018; pause data and the NWEA comprehensive data file data for non-CPS students seeking fall 2019 admission to CPS selective elementary and high schools; and seven reports from NWEA's Item Record (also

[5] NWEA scores its tests on its RIT, or Rasch Unit, scale.

called "play-by-play" data) involving more detailed information about seven student tests.

To guard against inclusion of possible mechanical or testing errors, the OIG eliminated from ~~the data~~duration-average calculations outlier tests with durations of more than 1,000 minutes.

B. Calculating Gains

Different experts may prefer different ways of analyzing student gains. However, following discussions with and input from Gregory Cizek, professor of educational measurement and evaluation at the University of North Carolina at Chapel Hill and past president of the National Council on Measurement in Education, the OIG settled on the following methodology to calculate gains:

1. The OIG looked at all SY 2017-2018 CPS students in grades 3-8 who had taken a test in both Spring 2017 and Spring 2018 and had progressed normally through the grades over those two years.
2. For students with the same starting RIT score, test subject and grade level, the OIG calculated the average RIT growth from 2017 to 2018 and the standard deviation of that growth.
3. The OIG then calculated a growth "z-score" for each individual student. The z-score is equal to the number of standard deviations that the student's growth was above or below the average growth for students with the same starting score, test subject and grade level.

The advantages of this methodology are that it compares students who started with the same score in 2017 and uses a CPS norm rather than a national norm. CPS tests are high-stakes while not all NWEA tests factored into NWEA's national norms are high-stakes.

Analysts must be cautious about interpreting z-scores that were calculated from small sample sizes. ~~However, of~~To be conservative, the ~~5,666 CPS students with~~OIG eliminated from all growth ~~z-scores of 2 or more, only 7~~calculations those tests that were compared to 50 or fewer ~~students~~tests with the same starting score, test subject and grade level.

C. Association Between Durations and Gains

To determine if there was an association between students who took longer to complete the tests and students with large increases in their test scores, the OIG divided students into groups based on their Spring 2018 test duration. They were

grouped in the following duration ranges: 0 to 75 minutes;[6] 76 to 120 minutes; 121 to 180 minutes; 181 minutes to 240 minutes; 241 minutes to 300 minutes; 301 minutes to 360 minutes; and greater than 360 minutes.

For each range, the OIG calculated the percent of students whose RIT score growth was two or more standard deviations above the average growth for students in the same grade, taking the same subject test, with the same Spring 2017 RIT score.

D. Association Between Pauses and Gains

To determine if there was an association between students whose tests were paused frequently and students with large increases in their test scores, the OIG performed a similar analysis to the duration analysis above.

Test results from the Spring of 2018 were grouped in ranges by number of pauses: 0; 1 to 4; 5 to 9; 10 to 14; 15 to 19 and 20 or more pauses. As above, for each range, the OIG calculated the percent of students whose score growth was two or more standard deviations above the average growth for students in the same grade, taking the same subject test, with the same Spring 2017 RIT score.

For ~~about 35~~more than 17,000 of ~~299~~roughly 320,000 tests — mostly from charter schools — NWEA did not provide pause or end-date data, which ~~is~~are not part of the comprehensive data file that NWEA traditionally gives CPS.

E. Probability of High Gains Clustering at Certain Schools

In order to determine whether there were any clusters of students with unusually high growth scores at certain schools, the OIG used a version of a statistical procedure called Fisher's Exact Test that was suggested by Dennis Maynes, Chief Data Scientist at Caveon LLC.

This test uses four numbers to compare a sample to the population from which it was taken and to determine the probability that a random sample taken from the population would include as many or more extreme results as were actually found in the sample being analyzed. The probability, or p-value, is based on the following numbers: the total population (in this case, the number of CPS students who took the test in a given grade and subject), the number in that population with a certain property (in this case, the number of CPS students in that grade and subject with a growth z-score of two or higher), the sample size (in this case, the number of students in a given school who took the test in a given grade and subject), and the

[6] The NWEA blog "Testing Duration: How Long is Too Long to Spend on the Map Growth Assessment?" says that, in general, NWEA expects students to complete a MAP Growth test in about 45 to 75 minutes.

number in the sample with that property (in this case, the number of students in that school, grade, and subject with a growth z-score of two or higher).

These numbers were used to calculate the probability that a random sample of CPS students in a given grade and subject would have as many or more students with a 2 or higher growth z-score as were actually found in each school. For example, in third-grade math, across the district, ~~542~~533 of 24,~~627~~025 tests (2.2%), had growth z-scores of 2 or more. If a particular school had 50 third-grade math tests, and 5 had z-scores of 2 or higher (10%), then Fisher's Exact Test would result in a p-value of about .~~0048~~0049. This means that there is just under a 0.5% chance that a random sample of 50 CPS third graders would have 5 or more students with a z-score of at least 2.

The OIG calculated this probability for every school, at every grade level, in every subject. In all, there were 5,461 different combinations of schools, grade levels, and test subjects in the OIG's data.

The schools' growth p-values provide a useful approximation of how unlikely a school's growth was to have occurred by chance. By using a count of the number of students with high gains rather than the average gain, this method also minimizes the impact of any one outlier gain.

However, although Fisher's Exact Test assumes that a random sample has been taken, the students in a particular school are certainly not a random sample of CPS students. So, its results must be interpreted cautiously. Maynes told the OIG: "I wouldn't take the results and say 'You're convicted.' I would say there is potentially smoke. We should follow up on it to try to find explanations."

Maynes also noted that statistics alone cannot prove someone cheated on a test because "statistical data by themselves do not measure intent." So, additional information is needed to determine whether cheating was involved.

F. Interviews of Students, Teachers and Parents

To conduct further inquiry into unusual results, the OIG called students who attended the grades in schools that produced among the largest number of long durations, the highest number of pauses or the largest clusters of high gains. Some students were contacted based on other information indicating concerns about their NWEA test scores. In total, with parental permission, the OIG talked to 20 students by phone about their NWEA tests.

Early on, one fifth-grader whose third-grade Reading score jumped from the 4th percentile to the 99th told the OIG he could not recall his third-grade Reading

teacher or NWEA test from up to two years earlier. After that, the OIG's student interviews often focused on older students, particularly seventh and eighth graders, and the most recent results then available (the 2018 tests). A handful of students were interviewed about their fourth- through sixth-grade NWEAs.

The OIG talked to students about a year after their 2018 tests so it's possible their recollections were not perfect. Some also may have chosen not to acknowledge improper testing procedures because they feared getting their teachers, or themselves, in trouble.

If students related information indicating a teacher had committed a test administration infraction, the OIG whenever possible attempted to contact those teachers by phone.

The OIG also called a few teachers randomly, without any indication that any impropriety was occurring in their classrooms or their schools, in order to get a sense of test administration procedures at a cross-section of schools. In total, the OIG talked to ten CPS teachers by phone about NWEA tests.

Not one teacher admitted improperly administering tests in ways described by students. But several said some students appeared to be, or could have been, intentionally timing out questions. This came from both teachers in schools with unusual results and teachers chosen randomly.

The OIG also interviewed a handful of current or former CPS parents who contacted the OIG with concerns that their child's test scores were the result of cheating.

G. Interviews with Experts

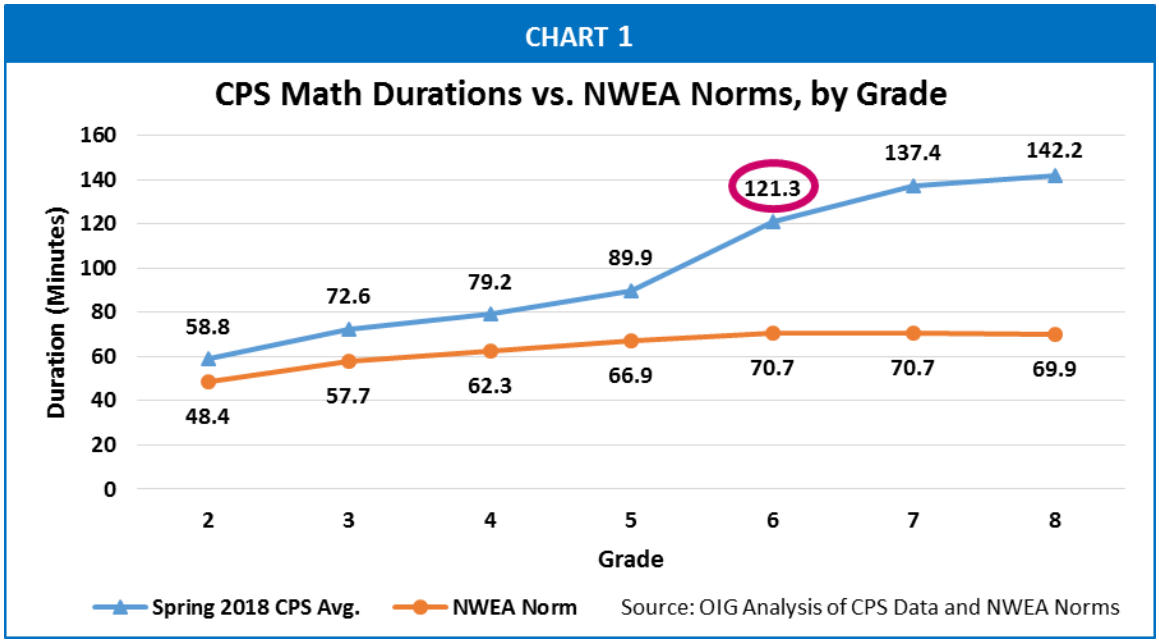
To inform its research, the OIG held discussions or email conversations with a variety of experts from NWEA and Caveon LLC, as well as two academic experts in educational testing. These experts are cited throughout this report.

DATA FINDINGS

A. Duration Findings

The OIG found that in every tested grade, in both Reading and Math, the average CPS student in 2018 took longer than the national norm[7] to complete a NWEA test, as shown in **Chart 1**.

[7] NWEA's duration norms for its Reading and Math MAP Growth tests are based on all students nationally who took those tests in the Spring of 2017. This includes Diverse Learners.



In particular, the gap between CPS average Math test durations and NWEA norms significantly increased in grades six, seven, and eight, all of which carry high stakes for students. (See **Appendix A** for a similar Reading duration comparison.)

Note that NWEA excludes from the duration count the amount of time a test was paused. The duration reflects the cumulative number of minutes that items appeared on a student's computer screen.

It is also notable that CPS's high durations are not driven by Diverse Learners who need extended time. As **Appendix B** shows, Diverse Learners were less likely to have long tests than non-Diverse Learners.

There could be many benign reasons for long durations. For example, the high stakes many students face concerning promotion and high school admissions could prompt some to take longer than they would if no such stakes existed. Non-native English speakers may take more time reading questions, especially when stakes are involved.

At a minimum, the long CPS durations indicate that at some schools, testing conditions were different than those of NWEA's national sample. According to NWEA:

> The data from CPS's MAP tests is compared to the norms for growth and performance from NWEA's nationally representative sample. For the inferences from that comparison to be accurate, CPS testing conditions

> should be reasonably reflective of testing conditions of other schools that administer MAP throughout the United States. . . .
>
> Test durations that vary excessively from the norms, or test durations that differ significantly between terms, may pose a risk to the accuracy of inferences made from the results of NWEA assessments. If test durations exceed the norms by an excessive degree, NWEA believes it may be reasonable to take steps to bring testing durations closer to norms.

This concept is also expressed in a NWEA blog called "Testing Duration: How Long is Too Long to Spend on the MAP Growth Assessment?" Its author, NWEA Vice President of Education Research John Cronin, noted that a test score is supposed to be a legitimate estimate of a student's ability. But, Cronin wrote, certain conditions could misrepresent that ability, including "allowing students to take multiple hours to complete the assessment in a single sitting" and "frequent interruptions or pauses initiated by the proctor."

Gregory Cizek, a professor of educational measurement and evaluation at the University of North Carolina at Chapel Hill and past president of the National Council on Measurement in Education, expressed concern about the length of some CPS durations. Said Cizek: "Assuming that these are not students who require accommodations, some of your times are so different that it calls into question the validity of these scores or at least making a comparison to what the NWEA is supposed to measure."

In fact, as **Appendix B** shows, non-Diverse Learners were more likely to take long tests than Diverse Learners, who can receive accommodations allowing extra time. While 4~~.1~~ percent of non-Diverse Learner tests had durations of more than four hours, ~~only~~ 2.~~6~~5 percent of Diverse Learner tests took that long.

In addition, even fourth- and fifth-grade tests — which carry stakes for teachers, principals and schools but not students — showed longer durations than their national norms. (Again, see **Appendix A**.)

According to NWEA, to guide its clients as to how long students normally take to complete a NWEA test, NWEA releases Average MAP Growth Test Durations for each subject-matter test and grade.

However, since at least the Spring of 2016, each of CPS's average durations, in every grade and subject, have exceeded national norms. Each of those average durations proceeded to increase in each of the following two years.

In fact, as shown in **Table 1**, from Spring 2016 to Spring 2018, each of CPS's average durations jumped double-digit percentages.

Table 1: Increases in Avg. Spring NWEA Durations from 2016 to 2018

CPS Grade	Math Increase	Reading Increase
3	12.4%	17.32%
4	12.5%	21.7%
5	10.7%	22.6%
6	17.64%	18.65%
7	22.2%	24.0%
8	18.0%	18.65%

Source: OIG Analysis of CPS Data

Seventh-grade Reading experienced the largest hike, of 24 percent. In Math, the largest duration increase also occurred in seventh grade, which grew by 22.2 percent.

Thus, CPS's duration problem[8] is getting worse over time. If no action is taken, CPS durations may continue to move farther and farther from national norms. This would increase the risk of CPS results that are tied to national norms being distorted, based on what the OIG has been told.

Several experts recommended examining durations as one point of inquiry into possible CPS test anomalies. Long durations can be an indicator of cheating, although this clearly is not a certainty. But the extra time can be used in ways that violate normal test administration procedures.

Analyzing CPS duration data, the OIG found that tens of thousands of CPS students far exceeded the national duration norms in 2018, as shown in **Table 2**.

In 2018, nearly 83,000 tests in grades 3 to 8 Reading and Math — or more than one out of every four tests — were taken by students who needed at least twice the national average duration to complete those tests.

Table 2: CPS Test Durations vs. National Norm

2018 Test Duration vs. National Norm	# of Tests	% of Tests	Students*	Schools**
All CPS 3rd – 8th grade tests	320,561	100%	160,906	498
At least 2 times the national norm	82,824	25.8%	55,630	495
At least 3 times the national norm	24,269	7.6%	17,853	482
At least 4 times the national norm	7,448	2.3%	5,832	401
At least 5 times the national norm	2,388	0.7%	1,966	258

*Reflects number of students with the indicated duration ratio on at least one test.

**Reflects number of schools with at least one test with the indicated duration.

Source: OIG Analysis of CPS Data

[8] The data CPS receives from NWEA includes the duration of each test taken, so CPS already has the ability to monitor duration data without any extra information from NWEA.

More than 24,000 tests were taken by students who needed at least three times the national average duration. Nearly 7,500 tests were completed in four or more times the national duration norms.

Notably, tests with long durations often were concentrated in certain schools. Of the more than 24,000 tests with durations that were at least three times the national norm, ~~nearly 5,000~~more than 4,700 (or almost 20 percent) were clustered in just 14 schools out of roughly 500 (or almost 3 percent of all schools). At each of these 14 schools, more than a third of all tests took three times the national norm to complete.

When Reading and Math durations were averaged by grade at each school, several high-stakes seventh- and eighth-grade tests at certain schools popped up among the top 25 tests systemwide with the longest durations over the CPS average for that subject and grade. This is detailed in **Appendix C**.

However, at a handful of schools, even some fourth- and fifth-grade tests — which have no stakes for kids but do have stakes for teachers, principals and schools — appeared in the top 25 tests for average durations over the CPS average.

Interestingly, one OIG analysis indicates the longer seventh-grade durations appear to be a CPS phenomenon.

Non-CPS eighth graders who took the Fall 2018 NWEAs as part of the admission process for outsiders seeking fall 2019 admission to CPS's coveted selective-enrollment high schools completed the Math test in an average 82.5 minutes and the Reading test in an average 85.2 minutes, NWEA data showed. But a few months earlier, CPS students who took the Spring 2018 seventh-grade NWEAs to determine their fall 2019 admission to SEHSs, as well as all other CPS seventh graders, averaged far longer durations, of just over 137 minutes in each subject.

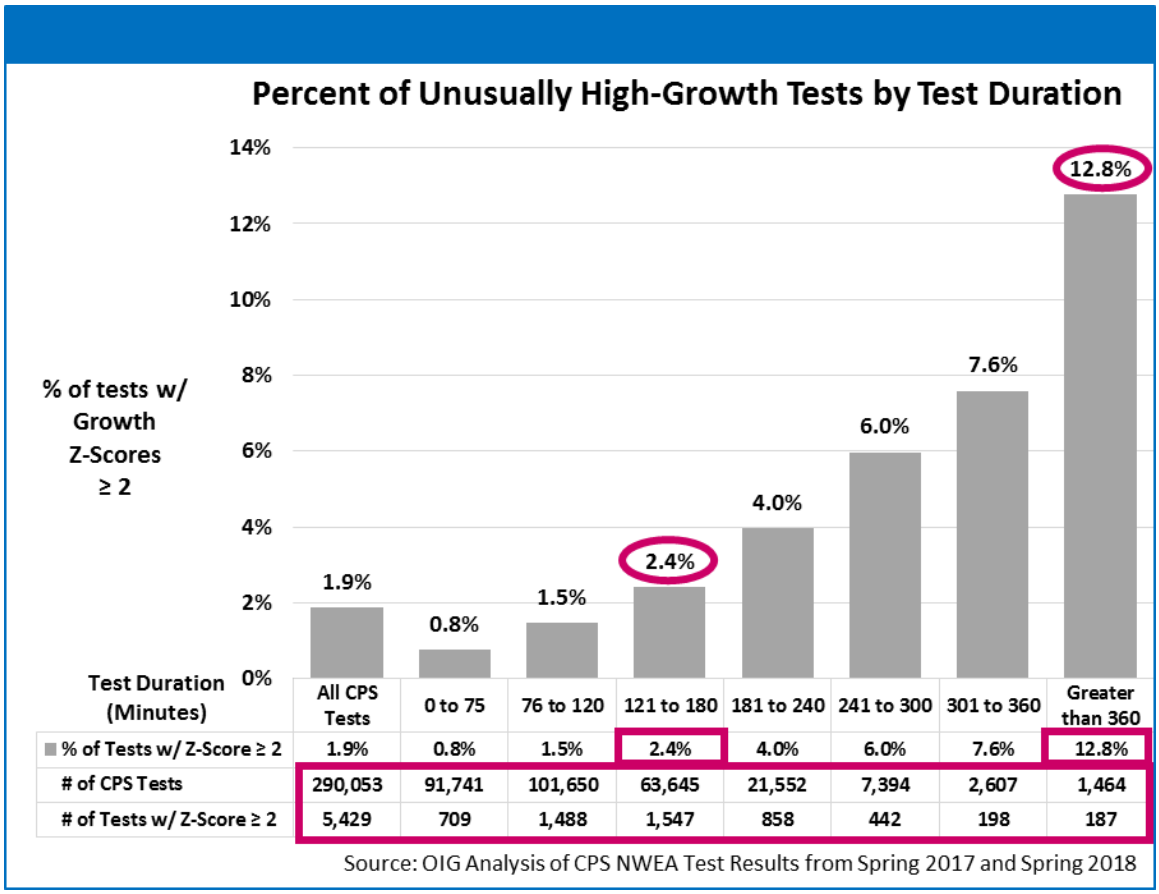
Thus, CPS seventh graders took 62 to 66 percent longer than their closest non-CPS counterparts to complete their NWEAs.[9] Note that both groups of students take these NWEAs in CPS schools.

[9] Non-CPS eighth graders taking the NWEAs as part of the admission process to Fall 2019 selective-enrollment CPS high schools also took far fewer pauses than all CPS seventh graders, including those seeking Fall 2019 admissions to selective-enrollment high schools. Of 4,~~950~~112 outsider eighth-grade tests in the fall of 2018, only 20 had 5 or more pauses, or ~~.4~~close to 0.5 percent. However, 7 percent of all CPS seventh-grade tests taken just a few months earlier, in spring of 2018, had 5 or more pauses, an OIG analysis showed.

The OIG also analyzed the test durations of SY 2017-18 CPS students who took NWEA tests in both Fall 2017[10] and Spring 2018. Here, the average durations, by subject and grade, of the Spring 2018 tests were longer than the Fall 2017 tests — by anywhere from 28 percent to more than 50 percent.

In addition, another OIG analysis indicates a connection between CPS durations and large test gains. As illustrated in **Chart 2**, the longer students took on their 2018 NWEA tests, the greater likelihood they had of posting a growth z-score of at least 2, the threshold set by the OIG for high growth. (As described in the Methodology section of this report, a growth z-score of 2 or higher means that a student's growth on a particular test was 2 or more standard deviations above the average growth for students with the same starting RIT score, at the same grade level, taking the same subject test.) This "high growth" standard was met on 1.9 percent of CPS third- through eighth-grade tests in Spring 2018.

By the time students were taking 5 to 6 hours (or 301 to 360 minutes) to complete a NWEA test — most likely spilling over into at least a second day of testing — 7.6



[10] Fall 2017 data provided by NWEA did not include charter data.

percent were posting a growth z-score of 2 or more. That's four times the systemwide average.

Students who took more than 6 hours were 6.~~6~~7 times as likely as the average CPS student to post unusually large gains.

Although Diverse Learners were more likely overall to achieve an unusually high growth score (2.~~67~~6% of Diverse Learners' tests had z-scores of at least 2 versus non-Diverse Learner's 1.~~77~~76% of tests), both groups of students were more likely to achieve high growth if they spent longer on their tests.

(See **Appendix D** for versions of the OIG's duration analysis that separate Diverse Learners and non-Diverse Learner students.)

NWEA told the OIG that large differences in durations between terms can affect the accuracy of growth calculations, saying:

> For example, if a group's average test duration was 50 minutes in spring 2017 and was 180 minutes in spring 2018, there would be reason to believe that a growth comparison could be inaccurate (*i.e.*, the academic growth of the students was not accurately measured because one test duration was much longer than the other). Additionally, under this scenario, comparisons to the NWEA norms would not be meaningful. The average difference in test durations from one year to the next under NWEA norms is within 10 minutes, making it difficult to draw meaningful inferences if the subject group has much larger differences in test durations.

So what could have been occurring during CPS's unusually long test durations that might have impacted student gains? The OIG turned to students and teachers in the grade levels of schools with long durations to try to find out.

One eighth grader who was listed in CPS records as an English Learner said that in her testing room, the teacher read the Reading questions to maybe half the class — something that would consume extra time and is not permitted. This student's Reading test duration, without pauses, was about 3 ½ hours; her math duration was close to 4 ½ hours.

Other students, none of them listed in 2017-18 records as Diverse Learners or English Learners, described to the OIG various improper uses of pauses that would have impacted test durations and possibly test scores.

One eighth grader said she could see classmates not even trying to answer some math questions. Three or four kids were intentionally letting questions time-out so they could get different questions, this student said. However, a friend timed-out questions a lot and still got a bad score, the student said.

The proctors would get upset about the time-outs because graduation was nearing, this eighth grader said. The proctors would tell students, "If you're just going to sit there and not answer the problem, it's not going to help you."

This student said she did not intentionally time-out questions. However, she said she was stressing and crying about the test and it took her six or seven days to finish.

An OIG summary of the custom "play-by-play" report of this student's eighth-grade Math test is presented in **Table 3** below.

Table 3: Example of High-Duration CPS Math Test Over 7 Test Days

Date	Start Time	Time from First to Last Question (Hr:Min)	Questions Paused or Timed Out	Questions Answered
5/23/2018	11:35 AM	1:37	3	4
5/24/2018	11:14 AM	1:53	4	3
5/25/2018	10:11 AM	2:58	6	5
5/29/2018	8:58 AM	6:26	7	9
5/30/2018	11:55 AM	3:09	5	7
5/31/2018	10:51 AM	4:14	6	7
6/01/2018	8:51 AM	1:53	0	18
NWEA Duration*: 8:03		**22:13**	**31**	**53**

*NWEA's Duration data excludes time when the test was paused. However, NWEA does not keep records of the exact time when each pause occurred or how long the test was paused. Pauses are reflected in the play-by-play data as questions that were put on the student's screen at a certain time but never answered. In contrast, the Time from First to Last Question column includes daily down-time during pauses.

Source: OIG Analysis of NWEA Spring 2018 Play-by-Play Data

This student took just over eight hours in total duration, excluding breaks, to complete her Math test. However, including daily (but not overnight) breaks, her play-by-play showed her test stretched over 7 school days and more than 22 hours.

Her play-by-play also shows that on some days, she paused or timed-out more questions than she answered. Although she only answered three to five questions on each of the first three days, on the seventh day she answered 18 questions in under two hours — with zero pauses — to complete the test. After 31 pauses or time-outs over seven days, her final Math score was around the 60th percentile.

In addition, her RIT score increased by several points on the final day of the test, indicating that she had successfully handled some tougher questions. So, apparently,

when she wanted or needed to buckle down and answer even challenging questions without pausing, she could.

At the same school, a seventh grader said the test "timed out" several times on her, but she conceded that she would not tell the IG if she was intentionally timing out in order to get another question. She said she answered about 17 of 53 math questions over close to 4 ½ hours on the first day of the test. She finally finished the Math test on the second day but needed three days to complete her Reading test, she said.

A seventh grader at a different school said she suspected some students were intentionally timing out because they had long Reading passages and hoped they would be replaced with shorter ones. This student took nearly six hours, without pauses, to complete her Math test.

An eighth grader at the same school said about five questions on the Math test timed out on her. "A lot of people waited for [the test] to time out. They think it's gonna give them an easier question but it doesn't work like that. It remains the same,'' the student said. This student took nearly six hours, excluding pauses, to complete her Math test, duration data showed.

A seventh grader at the same school said students would have rather seen a question time out than answer it incorrectly because they were worried about the impact the test results would have on their ability to get into certain high schools.

"We were so worried about high school. We didn't want our score to drop. . . . I know for a fact that guessing — we would never do that," the student explained.

Testing in both the morning and afternoon, the student said she only made it to question five on the first day of the 53-question Math test. She tried to pace herself to make it to the twenties the second day. Data showed she took more than 600 minutes in total duration, excluding pauses, to complete her Math test. The student estimated it took her four days to finish that test and 4 ½ to finish her Reading test.

This student also reported another improper testing procedure that could add to duration. She said her Math teacher advised students to write down on scratch paper both the question and all the answers of any challenging questions. The information on the scratch paper was collected each day and used later for class review. (Copying questions during a test for future reference is not only time-consuming, it's improper, according to NWEA.)

At yet another school, a seventh-grade Diverse Learner told the OIG that her teacher gave hints during her Math test. If she was doing a problem wrong, the teacher would say she needed to re-read a problem, the student said. Such hints, and the re-

working of problems they can cause, can lead to longer durations. This student's Math test took nearly 6 ½ hours, without pauses.

In addition, unusually long tests eat up instructional time.

Students described sitting with nothing to do, their heads on their desks, after finishing their tests ahead of colleagues. Some teachers expressed frustration that they were unable to teach any students when some of their students had not finished their tests.

Most Spring 2018 NWEAs were finished in a day but just shy of 30 percent of them took multiple days to complete, an OIG analysis indicated.

As **Table 4**[11] shows, at least half the Spring 2018 tests in 113 of 463 schools were multiple-day events, indicating that such tests were concentrated in certain schools.

Table 4: Schools by % of Tests that Took Multiple Days

# of Schools	% Multiday Tests
58	75% or more
55	50% to 74.9%
91	25% to 49.9%
259	Less than 25%
463	All Schools

Source: OIG Analysis of Spring 2018 CPS NWEA Data from Grade 3-8 Tests

In some cases, only one class period might have been devoted to the test each day. In other cases, students may have tested several hours over multiple days, as was the case with many interviewed by the OIG.

However, the sample testing schedule CPS distributes to schools assumes that most students will take an hour to complete their NWEA tests.

In a blog entitled "Testing Duration, How Long is Too Long to Spend on the MAP Growth Assessment,'' even NWEA warns that "An efficient measure of student learning shouldn't have the student away from the classroom for several hours at a time.''

However, when given a summary of CPS duration data, NWEA declined to say which durations by grade and subject were likely to be problematic as it had not published research on this issue.

[11] Test end dates had to be specially requested from NWEA for the OIG to do this calculation. Unfortunately, the OIG could not calculate an average number of days needed to complete a test because it could not assume that every week day between the test start date and end date was a testing day. This is information the OIG recommends NWEA include in future CPS comprehensive data files as it would help CPS identify schools that are spending large numbers of days on a single NWEA test.

B. Pause Findings

Currently, CPS's NWEA Accommodations Matrix[12] specifically states that general education students, as well as Diverse Learners and English Learners, may take "frequent breaks" during their NWEA tests.

Additionally, one CPS teacher noted that there is nothing in CPS training that says students can't ask for a break, or a pause in the test, in order to get a new question.

In written answers to the OIG, NWEA called intentionally pausing a test to produce a different question improper. NWEA considers pausing a test to be appropriate if a student needs a bathroom, water or wiggle break.

In April 2019, NWEA finally added language to its Guidelines that specifically labeled a "high number of pauses" as "outside the bounds of what is expected." This language was issued after the Spring 2018 NWEA tests in question and after the OIG had contacted NWEA with questions about its test.

The pause function is intended as a security feature so that students cannot research answers to pending questions while on breaks. Instead, as mentioned previously, students face new questions of similar difficulty level in the same area when they return from breaks. If a test is paused on a hard question, a student has a 50 percent chance of getting a more favorable question, and an even greater chance if the same question number is paused multiple times.

Plus, even general education students can take the test over "several days," according to CPS's NWEA Accommodations Matrix. If a test continues into a second day, it would be suspended at the end of the first day — an action that is counted in NWEA custom reports as a pause.

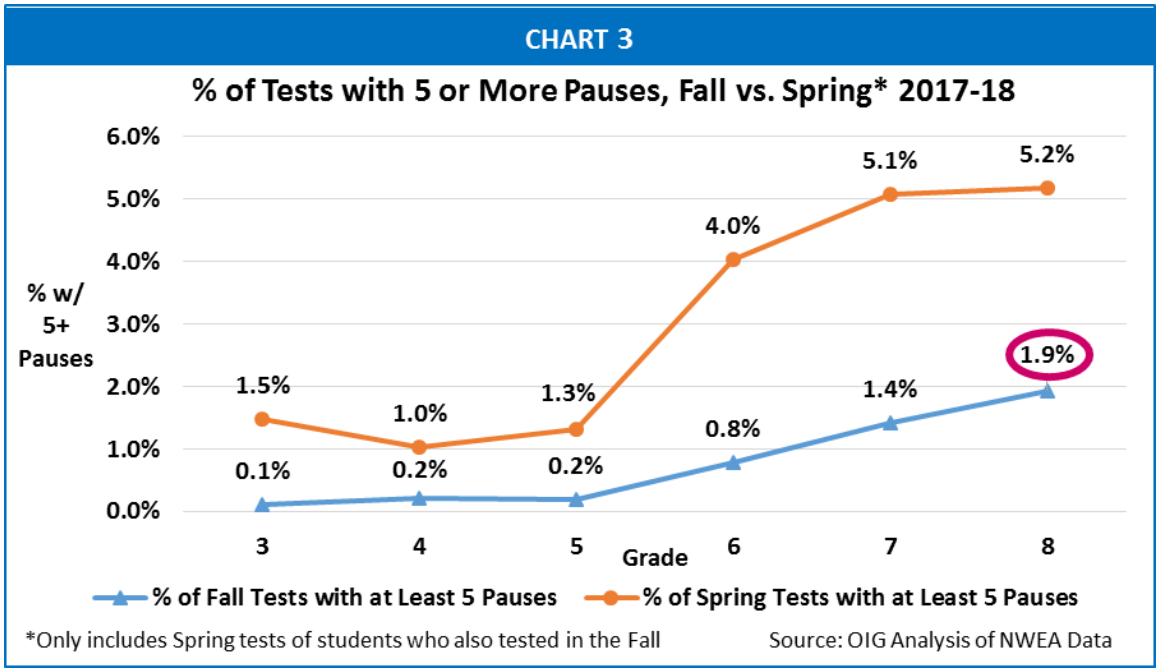
In addition, if a student walks away from his or her test with the test still on the screen, after 25 minutes without an answer, the test will automatically time-out.

Thus, pauses may occur for perfectly innocent reasons.

However, when the OIG analyzed the pauses of students who had tested in both Fall 2017 and Spring 2018, it found that these students were much more likely to have had at least five pauses on their Spring tests than on their Fall ones, as indicated in **Chart 3**.

[12] To see the Spring 2018 Accommodations Matrix, go to pages 21 to 26 of the SY18 NWEA EOY Test Administration Manual.

This suggests that, without high stakes attached to the results of the tests, students and proctors used the pause function much differently. This was especially true in the sixth, seventh and eighth grades.



Many CPS Spring 2018 tests were never paused at all. An analysis of CPS NWEA data indicated that close to 48 percent of all tests involved no pauses at all.

But some tests clearly reflected what appeared to be an unusual number of pauses, as indicated in **Table 5** below.

Table 5: CPS Tests by Number of Times Paused or Timed Out

	Total	0	1-4	5-9	10-14	15-19	20+
Tests	~~262,398~~302,993	~~25,629~~145,424	~~25,972~~145,388	~~9,273~~10,524	1,~~033~~149	~~276~~290	~~15~~218
Students*	~~132,222~~152,128	~~78,697~~91,221	~~1,508~~94,309	~~7,835~~8,904	~~945~~1,053	~~254~~268	~~78~~180
Schools*	~~429~~463	~~423~~459	~~429~~462	~~371~~401	~~146~~165	~~54~~60	24

*Reflects students and schools with at least one Reading or Math test in the indicated pause range.

Note: The OIG did not receive pause data for some tests. Those tests are excluded from this analysis.

Source: OIG Analysis of Spring 2018 CPS NWEA Data from Grade 3-8 Tests

~~Close to 11~~More than 12,000 tests had at least five pauses, more than 1,~~500~~600 had at least 10 pauses and more than 200 tests had 20 or more pauses, **Table 5** shows.

Some might question whether pauses were mostly taken by Diverse Learners whose IEPs allowed breaks (which are recorded as pauses). However, this was not the case. A much larger percentage of non-Diverse Learners' tests had at least 10 pauses than those of Diverse Learners. This is indicated in **Appendix E**.

Asked what would constitute an unusual number of pauses, one NWEA expert told the OIG: "When you see it, you probably know it."

This expert said he would not be concerned about one, two or three pauses. But if he saw a classroom averaging 10 or 12 pauses, that would concern him.

Unfortunately, CPS NWEA data does not reflect which students were tested as a class. CPS's comprehensive data file includes a student's homeroom number, but the OIG interviewed few students who were tested with their homerooms. For example, many students took the Reading test with their Reading teacher. But some general education students told the OIG they were pulled out of their subject-matter classes or homerooms and tested in small groups. One Diverse Learner said he was tested one-on-one, with a proctor sitting right next to him.

Due to all these different ways of testing students, even Battelle for Kids roster information indicating which students were assigned which Math and Reading teachers is not necessarily a reliable record of who proctored which students.

Table 6: % of Tests w/ 5 or More Pauses by Grade

Grade	% of Tests w/ 5+ Pauses
3	1.7%
4	1.~~9~~8%
5	2.~~4~~3%
6	4.6%
7	7.~~2~~0%
8	7.~~3~~1%
All	4.~~1~~0%

Note: The OIG did not receive pause data for some tests, which are excluded from this analysis.

Source: OIG Analysis of CPS Spring 2018 NWEA Data

In addition, the OIG had to custom order pause counts per test as such data is not currently provided with NWEA's comprehensive data file.

If an analyst tried to aggregate this custom pause count by test subject and grade within a school, at schools with multiple classrooms per grade the data from high-pause classrooms could be diluted by data from other normal-pause classrooms in the same grade. And, at some schools even general education students are tested in small groups, according to student reports to the OIG. That means one proctor might allow one small group from the same grade a lot of pauses but another proctor might not make such allowances for other small groups from the same grade.

One OIG analysis made clear that the percent of tests with at least five pauses increases as grade levels increase. As **Table 6** shows, a significant jump in the percent of grade-level tests with five or more pauses occurred in seventh and eighth grades. Because these tests can impact high school admissions and eighth-grade graduation, they have especially high stakes for students.

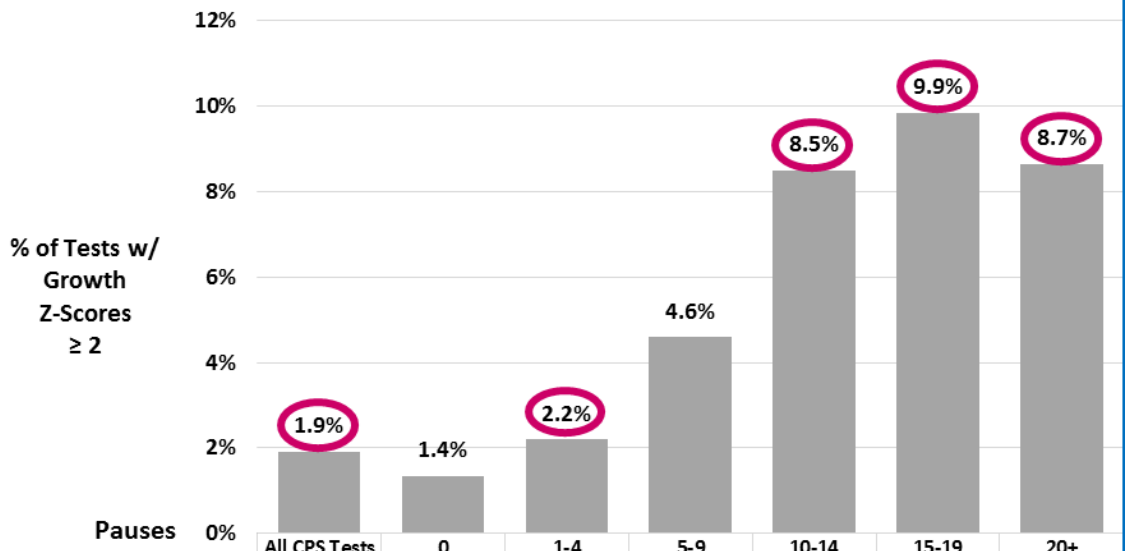
The OIG also created an average pause count for each grade and each subject-matter test by school. **Appendix F** shows the top 25 average pause counts in the district. Unlike the top 25 duration-average chart **(Appendix C)**, this top 25 pause-average chart was clearly dominated by seventh- and eighth-grade tests.

One possible explanation is that in these high-stakes grades, older students had figured out and shared how to time-out the test in an attempt to game it. After all, at least one older student admitted discussing the test during lunch breaks and several older students were aware that they could get new questions by timing out the test. One even discussed the results of this tactic with her friends.

As with durations, the OIG found a connection between the number of pauses and the percent of unusually large test gains. See **Chart 4** below.

CHART 4

% of Unusually High-Growth Tests by Times Test was Paused or Timed Out



Pauses	All CPS Tests	0	1-4	5-9	10-14	15-19	20+
% of Tests w/ Z-Score ≥ 2	1.9%	1.4%	2.2%	4.6%	8.5%	9.9%	8.7%
# of CPS Tests	274,123	130,157	132,629	9,785	1,070	274	208
# of Tests w/ Z-score ≥ 2	5,256	1,758	2,911	451	91	27	18

Note: The OIG did not receive pause data for some tests. Those tests are excluded from this analysis.

Source: OIG Analysis of CPS NWEA Test Results from Spring 2017 and Spring 2018

As shown in **Chart 4**, CPS tests with at least 10 pauses had a ~~four~~roughly 4.5 to ~~five~~5 times higher chance of achieving a growth z-score of 2 or more than CPS tests overall.

~~Similar patterns appeared among the tests of~~ Both Diverse Learners and non-Diverse Learners were more likely to achieve unusually high growth if they paused more often. (See **Appendix G** for versions of the OIG's pause analysis that separate Diverse Learners and non-Diverse Learners.)

This is curious because, as UNC's Cizek put it, other than for students with special needs who require accommodations involving breaks, there's "no educational explanation for why pauses would improve scores."

So why were some schools experiencing such high Spring 2018 pause levels? The OIG turned to students to find out.

One eighth grader reported that, in her Math testing room, if a student raised a hand and said he or she didn't know the answer to a question, the teacher would pause and then resume the test so a new question would appear. This teacher allowed each student two such free pauses per testing day, the student said.

At her school, the NWEA was "always a two-day process, never rushed," the student said. During the Math test, her teacher provided not only lunch breaks but two or three class-wide water or bathroom breaks during which the test was paused and then resumed. This student said she took four additional, individual breaks on her own, leaving her question on the screen during those breaks without pausing the test (something the proctor should not have allowed).

However, an OIG summary of this student's custom play-by-play report, shown in **Table 7**, indicates this student was understating her pause activity.

Table 7: Example of High-Pause Eighth-Grade Math Test

Date	Start Time	Time from First to Last Question (Hr: Min)	Questions Paused or Timed Out	Questions Answered
5/25/2018	9:35 AM	5:40	10	21
5/29/2018	2:15 PM	1:26	5	9
5/30/2018	10:48 AM	1:40	14	23
NWEA Duration*: 4:25		**8:46**	**29**	**53**

*NWEA's duration data excludes time when the test was paused. However, NWEA does not keep records of the exact time when each pause occurred or how long the test was paused. Pauses are reflected in the play-by play data as questions that were put on the student's screen at a certain time but never answered.

Source: OIG Analysis of NWEA Play-by-Play Data

This student's Math test had 29 pauses over the nearly 4 ½-hour duration, without breaks, needed to complete her Math test. That included, on the third and final day of the test, 14 pauses in less than two hours.

Not only was this test paused many times (a total of 29) but it was sometimes paused on consecutive question numbers or several times in a short period of time, further detail in the play-by-play showed. For example, over one 35-minute period on the last day of the test, it was paused nine times, including four times in a row over nine minutes. Because these pauses occurred within such a short period of time, they could not have been the result of time-outs, which occur after 25 minutes of inactivity. Instead, the test was likely paused on the same question number by the proctor four times in a row and resumed each time by the proctor.

At another school, a Diverse Learner said his proctor would pause his sixth-grade Reading test up to three times in a row to net him new questions before finally requiring him to answer the question.

At this same school, a former CPS general education student said two of the proctors in his fourth-grade Math testing room told students who were stumped on a question to raise their hands so the proctors could pause the test and resume it with a new question. Students merely had to say, "Can I get a new question?'' when they didn't know an answer. If a student still didn't know the answer the adults would give clues, such as nodding their heads "yes" or "no," the student said. However, on the Reading test, which featured different proctors, students were not allowed to ask for new questions, the student said.

This technique of intentionally asking for pauses to replace one question with another presumably would add to a duration count because it would mean the student was faced with additional questions.

Pauses also would add to the duration count if they occurred after an intentional, or even accidental, "time-out" — another technique reported by students.

One student whose eighth-grade math test had 36 pauses and a duration of close to six hours, excluding breaks, said the test timed out on her many times. However, she admitted that she probably would not tell the IG if these time-outs were intentional on her part.

At the same school, another student said she believed some students in her classroom were intentionally timing out questions in both Reading and Math because they were frustrated, although she did not do so. Such students just sat at

their desks, not even trying to answer questions. This student's Math test had 31 pauses and her Reading test had 22 pauses.

"In general, there is no reason for a student or proctor to allow a question to 'time out,' " NWEA told the OIG. Doing so "should be considered a possible 'gaming' practice."

Nationally, the average student takes the NWEA in about an hour, which probably involves fewer breaks than many CPS students enjoyed. Some CPS teachers seemed to be using breaks in a way that could advantage students. For example, one student said her teacher told kids shortly before breaks that they would have five minutes to finish their questions before the break or, if they were on a "question they didn't know well," they could just wait for the test to pause.

C. Growth Findings

An OIG analysis indicated that high-gaining students (whose RIT score growth was 2 or more standard deviations above the mean growth for students testing in the same grade and subject with the same starting RIT score) tended to be clustered in certain grades and subjects within certain schools. Some schools might have only one high-gaining test, by subject and grade. But a few schools seemed to have multiple high-gaining tests, by subject and grade.

Experts cautioned the OIG that there could be many benign reasons scores could increase by unusually large amounts.

For starters, the student's previous year's score could have been uncharacteristically low, for a multitude of reasons — heavy absences, personal trauma, or test anxiety to name a few. This could make the following year's scores look unusually high.

In addition, if students greatly improved their language proficiency or were diagnosed with a learning disability and provided accommodations, they might be better able to show their true ability and experience a large score jump.

One teacher who teaches half the eighth-grade Math classes in his school said roughly half of his students are not native-English speakers. Some may not even have gone to school full-time before they arrived in America, he noted. He attributed his school's large eighth-grade Math gains to his students' increasing English proficiency as well as their hard work.

Students also may experience strong gains following a switch to a new curriculum or learning strategy or lessons with a particularly effective teacher. One teacher said she believes her students did better when she looped with them from fourth to fifth

grade. A seventh-grade teacher attributed high Math gains to the introduction of Algebra in her grade.

To identify clusters of high-growth students, the OIG conducted an analysis that is explained in the Methodology section of this report. The OIG's analysis resulted in a p-value, or probability-value, that a random sample of CPS students in a given grade and subject (for example, third-grade Math) would have as many or more high-growth students as a particular school actually had in that grade and subject.

These probabilities should be interpreted cautiously because CPS students in a particular school are not actually a random sample of CPS students and because there can be legitimate reasons for a cluster of high-growth students. However, the p-values are a useful approximation of how extreme a particular cluster of high-growth students is.

Dennis Maynes, Chief Data Scientist at Caveon, who suggested the test used by the OIG to calculate p-values, also recommended that the OIG set a very conservative threshold to flag schools as having high growth. He suggested choosing as a threshold a probability level that would be very unlikely to flag any schools if each group of students analyzed was a representative sample of CPS students in the same grade taking the same subject test. Based on the number of combinations of schools, grades, and subjects being analyzed, Maynes recommended looking at p-values of less than one in a million, or, to be very conservative, less than one in a billion.

If the clusters of students that were analyzed were all representative samples of CPS students in their grade and subject, the OIG would expect essentially zero[13] of its 5,461 combinations of schools, grades, and subjects to have probability-values of less than one in a million.

So, according to Maynes, "Any classrooms that you flag at this level, something is different in this classroom than other classrooms. . . . We don't know what that is, but something is different."

Maynes also recommended that, because all p-values below these thresholds are very small, the OIG should not "worry about the exact number. It just means pay attention."

[13] To be exact, the OIG would expect .005461 out of 5,461 combinations to have p-values of less than one in a million and .000005461 combinations to have p-values of less than one in a billion. Either is so close to zero that the OIG would not expect to flag any school/grade/subject clusters if they were all representative samples of CPS students in their grade and subject.

So, high-gaining schools/grades/subjects flagged by the OIG are worth further inquiry, to determine both whether their tests were administered properly and what the schools might be doing well to help their students achieve such unusual growth.

The OIG calculated that ~~57~~55 grades/subjects at ~~43~~40 schools, including ~~six~~five at Bouchet Math and Science Academy, had growth probability-values of less than one in a *million*. ~~Nineteen~~Twenty of those ~~57~~55 grades/subjects, including two each at Dixon ~~Elementary~~and Spencer, had probability-values of less than one in a *billion*.

Table 8 shows the ~~19~~20 schools/grades/subjects with clusters of high-growth students that would have had less than a one-in-a-billion chance of occurring in a random sample of CPS students testing in that grade and subject, along with their average test durations and average number of times each student's test was paused or timed out.

Rather than show the exact p-value calculated, the table shows the number and percent of tests in the school/grade/subject that had growth z-scores of 2 or more compared to the district-wide percent of students in that grade/subject with growth z-scores of 2 or more. (For the ~~57~~55 schools/grades/subjects with p-values of less than one in a million, see **Appendix H**).

The various school/grade/subject combinations are sorted from the least probable down.

Table 8: Clusters of High-Growth Students with Less than a One in a Billion Chance of Occurring in a Random Sample of CPS Students in that Grade and Subject

School	Grade	Subject	Avg. Duration (Hr: Min)	Avg. Pauses	# of Tests	Tests w/ Growth Z-Score 2+	% w/ Growth Z-Score 2+	CPS % w/ Growth Z-Score 2+*
Chavez	3	Math	1:54	0.6	~~104~~103	29	~~27.9~~28.2%	2.2%
Faraday	4	Math	1:~~26~~27	0.1	~~27~~26	16	~~59.3~~61.5%	2.2%
Ruggles	7	Reading	2:26	1.3	27	14	51.9%	1.6%
Clinton	8	Math	2:~~35~~38	1.1	~~93~~75	~~22~~20	~~23~~26.7%	2.3%
Greeley	6	Math	2:04	0.2	57	16	28.1%	2.0%
Farnsworth	5	Math	1:~~52~~53	0.~~4~~5	~~54~~52	~~17~~16	~~31.5~~30.8%	2.3%
~~Greeley~~Pickard	~~6~~8	Math	~~2:03~~3:42	~~0.2~~1.8	~~58~~47	~~16~~15	~~27.6~~31.9%	2.~~0~~3%
Budlong	8	Math	4:34	5.8	76	18	23.7%	2.3%

School	Grade	Subject	Avg. Duration (Hr: Min)	Avg. Pauses	# of Tests	Tests w/ Growth Z-Score 2+	% w/ Growth Z-Score 2+	CPS % w/ Growth Z-Score 2+*
~~Pickard~~	~~8~~	~~Math~~	~~3:38~~	~~1.8~~	~~48~~	~~15~~	~~31.3%~~	~~2.3%~~
Bouchet	6	Math	4:31	3.8	56	15	26.8%	2.0%
~~Peck~~Spencer	~~3~~5	Math	1:~~34~~40	~~1~~0.4	~~163~~48	~~23~~14	~~14.1~~29.2%	2.~~2~~3%
Dixon	6	Math	5:52	8.4	~~57~~56	14	~~24.6~~25.0%	2.0%
~~Spencer~~KIPP - One	5	Math	~~1:37~~2:10	~~0.3~~N/A**	~~51~~82	~~14~~17	~~27.5~~20.7%	2.3%
~~KIPP - One~~Peck	~~5~~3	Math	~~2:11~~1:34	~~N/A**~~ 1.4	~~85~~162	~~17~~22	~~20.0~~13.6%	2.~~3~~2%
~~Von Linne~~Mason	3	~~Math~~Reading	~~3:43~~2:18	1.3~~.7~~	~~59~~35	~~14~~11	~~23.7~~31.4%	~~2.2~~1.8%
~~Mason~~Von Linne	3	~~Reading~~ Math	~~2:16~~3:43	~~1.2~~3.7	~~38~~57	~~11~~14	~~28.9~~24.6%	~~1.8~~2.2%
Ravenswood	3	Math	1:36	1.5	61	14	23.0%	2.2%
~~Fiske~~Dixon	~~4~~8	Math	~~1:49~~5:12	11.0~~.8~~	~~48~~89	~~12~~16	~~25~~18.0%	2.~~2~~3%
~~Dixon~~Fiske	~~8~~4	Math	~~5:13~~1:49	~~10.9~~0.8	~~96~~45	~~16~~12	~~16~~26.7%	2.~~3~~2%
Whistler	5	Reading	2:~~51~~55	1.2	~~33~~31	9	~~27.3~~29.0%	1.~~5~~4%
Spencer	6	Math	2:00	0.2	45	11	24.4%	2.0%

*For tests taken in the same subject by students of the same grade level.

**This school's pause data was not provided to the OIG.

Note: These clusters of high-gaining students are sorted by their probability of occurring in a random sample of CPS students testing in the same grade and subject, although the OIG is not including the exact probability in this chart. The school/grade/subject combinations are sorted from the least probable down.

See: Methodology for an explanation of the OIG's analysis of student growth from Spring 2017 to Spring 2018.

Andrew Ho, a professor of educational measurement at the Harvard Graduate School of Education, cautioned that a probability-value for high gains could be considered a flag for schools that are doing really well. A very low p-value, Ho noted, means that the model being used cannot account for what has occurred. Ho would label this as surprising rather than suspicious unless there is "strong convergent evidence that cheating has occurred." He would supplement growth analysis with "qualitative or additional quantitative checks in the spirit of understanding how that data came to be and not in an accusatory or stigmatizing way."

Other experts also advised using multiple measures to flag unusual results. In keeping with this advice, the OIG looked not only at surprising gains but also at unusually large durations and pauses. The OIG also interviewed students and teachers in schools and grades with unusual results.

Some of the schools/grades/subjects with growth p-values of less than one in a billion also had high average durations or pauses. For example, as shown in **Table 8**, Dixon's sixth- and eighth-grade math tests each averaged more than eight pauses and five hours of duration per test. These results are clearly worth further exploration.

As an example of what some individual high-growth student results look like on paper, ~~**Table 9**~~the table below shows the Spring 2017 versus the Spring 2018 percentile gains and duration changes of the five highest-gaining students from Dixon Elementary's Spring 2018 sixth-grade Math NWEAs. Note that their 2018 tests took at least seven hours and they paused anywhere from 9 to 19 times.

Examples of Dixon 6th Grade Math Students with Highest Growth*

Student	Spring 2017 Percentile	Spring 2018 Percentile	Spring 2017 Duration (Hr: Min)	Spring 2018 Duration (Hr: Min)	Spring 2018 Pauses
A	17	93	1:52	7:05	12
B	54	98	3:05	7:14	19
C	28	90	0:57	8:11	18
D	68	96	4:46	8:22	9
E	80	98	4:53	10:16	16

* See: Methodology for an explanation of the OIG's analysis of student growth from Spring 2017 to Spring 2018.

Source: OIG Analysis of CPS and NWEA Data

But some schools/grade/subjects with p-values of less than one in a billion did not display extraordinarily long durations or pauses. For example, Faraday Elementary's fourth-grade Math tests averaged about 1.5 hours and had almost no pauses. These results, too, are worth further exploration.

As an example of what high-growth students looked like at Faraday, data for its five highest-gaining students from its Spring 2018 Math test are listed below:

Examples of Faraday 4th Grade Math Students with Highest Growth*

Student	Spring 2017 Percentile	Spring 2018 Percentile	Spring 2017 Duration (Hr: Min)	Spring 2018 Duration (Hr: Min)	Spring 2018 Pauses
F	13	93	0:41	1:44	0
G	20	93	0:37	1:32	0
H	10	89	0:48	1:32	0
I	57	98	0:45	1:21	1
J	37	93	0:59	1:32	0

* See: Methodology for an explanation of the OIG's analysis of student growth from Spring 2017 to Spring 2018.

Source: OIG Analysis of CPS and NWEA Data

Notice that even Faraday's highest gaining students had, at most, one pause and were able to complete their tests in less than double the national duration norm, which was 62.3 minutes in fourth-grade Math.

CPS AUDIT OF NWEA TESTING PROTOCOLS AND CPS RESPONSE

Sometime before the Spring 2018 NWEAs, the CPS Departments of Student Assessment and School Quality Measurement and Research asked the CPS Office of Internal Audit and Compliance to review "the governance and internal controls" surrounding NWEA tests.

The completed audit — dated April 2018, shortly before the Spring 2018 NWEAs began in mid-May — noted that previously there was not "an established and documented accountability process for schools and teachers with a history of potential test irregularities." Even after various methods were used to flag schools as high risk the previous year, 40 percent of flagged schools (12 of 30) were not audited on site in 2017, as intended.

The audit listed as "high risk" the finding that CPS's controls for *detecting* irregularities needed strengthening. It listed as "moderate risk" the finding that CPS's controls for *preventing* irregularities needed a boost.

In addition, this audit warned that schools with a significant number of test irregularities can suffer "potentially unreliable test scores and inaccurate measures of student progress."

In response, the Department of Student Assessment created a training PowerPoint for test administrators and proctors. All such personnel had to watch the

PowerPoint and pass a five-question "exit-slip" test afterwards to ensure they understood proper administration procedures before they could give the 2018 NWEAs.

According to a NWEA Plan Implementation memo, updated in January of 2019, Student Assessment was "confident" that this new training, combined with other compliance steps, "helped ensure a smooth and secure administration of the NWEA assessment" in the Spring of 2018.

However, an OIG review of the PowerPoint used in proctor training found that it lacked sufficient guidance on how to prevent test irregularities — even though the training was intended to avert such irregularities. The PowerPoint also contained at least one error — it falsely stated that a disengagement alert[14] would notify proctors when a student was lingering too long over a question. To be correct, this alert is only triggered when a student answers three successive questions with rapid guesses.

Three accommodation rules for general education students underwent some changes by the Spring 2018 test administration, at CPS's request. For example, by 2018, proctors were no longer allowed to read Math questions or answers to general education students. However, the proctor PowerPoint contained no indication that these changes were ever explained. No teacher questioned about this by the OIG was aware of these changes.

Concerning the exit slip, its questions were not focused on detecting or addressing test irregularities. Instead, its questions included: "When is the last day of testing for benchmark grades (third, sixth, and eighth)?" and "How many units are in the NWEA test?"

And, to pass the exit slip, proctors needed to get only four of five questions correct.

Asked why the training PowerPoint did not explain some of NWEA's unusual features, such as time-outs, one Student Assessment official said some training changes were being considered but CPS wanted to make sure to share the right guidance without giving proctors ideas on how to spread bad practices. This is an understandable concern.

However, some of the students interviewed by the OIG seemed to know more about how the NWEA test functioned in regards to time-outs and pauses than the teachers the OIG interviewed.

[14] In response to an OIG inquiry, NWEA is looking into whether it is possible to provide an alert when a student lingers too long on a question.

It appears that, at some schools and among some older students, the cat is already out of the bag on how to game the test. Meanwhile, some teachers expressed astonishment that tests could be intentionally timed out. Several said they thought such time-outs were due to mechanical problems with the test. Some didn't know if pauses produced harder, easier or same-difficulty-level questions. One teacher said she was never trained on the difference between pausing and suspending a test, although she used both options.

Some teachers said they would appreciate more in-depth training and suggestions on how to guard against intentional time-outs, an unusual number of student requests for breaks and long durations.

Every teacher the OIG questioned about the exit slip said they passed it on the first try, yet few knew many details about how the NWEA worked. At least one teacher said the exit slip should have more questions.

The audit recommended that a Test Security Agreement, which for years proctors were required to sign before administering any NWEAs, should include a warning of the potential penalties for test cheating, which can include termination. To date, these penalties have yet to be added to the form.

The 2018 audit also recommended that CPS establish a flagging process to monitor, identify and respond to test irregularities. It suggested that at least seven NWEA data points, along with other criteria, be used to flag unusual test results, including the number and length of pauses. However, CPS never analyzed its 2018 pauses in response to this April 2018 recommendation, although it inquired about having NWEA do some kind of analysis more than a year later, in May 2019, in response to some tips, according to NWEA and one CPS official.

In addition, neither NWEA nor any outside experts were consulted in developing this flagging process, according to a Student Assessment official. One Caveon expert questioned why the eventual flagging analysis relied on Spring-to-Spring changes in both RIT scores and percentiles, when the two measures are very closely aligned. This expert suggested using just the RIT scores. The only other flag is based on the change in Spring-to-Spring durations. This flag would not catch classrooms with chronically long tests. The OIG recommends also looking at duration lengths (not their change) to identify schools, or grades and subjects within them, with excessively long tests.

The audit suggested that schools receive on-site audits based in part on the results of this flagging process, which is tied to the previous year's test results. However, some schools with multiple flagged tests based on 2018 results were never audited

in 2019, according to documents provided by the Department of Student Assessment. In other cases, 2019 auditors visited flagged schools, but went to different Reading and Math classrooms, by grade, than those that had been flagged.

The audit also recommended that CPS explore the rotation of proctors so that teachers would not be proctoring their own students — at a minimum in schools identified as high-risk through the Assessment flagging system and other criteria. One CPS Student Assessment official said this recommendation was viewed as not feasible and noted that students benefit from a familiar testing environment. However, several teachers told the OIG that years ago they were required to use two proctors during state tests. This is also what NWEA recommends in a document called "Guidance for Administering NWEA MAP/MPG Assessments When Results Are Used for High-Stakes Purposes." At least one of the proctors should have no stakes in the test, according to this NWEA document.

In the wake of the 2018 audit, Student Assessment developed what purported to be a 2019 "Anonymous Fraud Reporting Form." However the second paragraph of the form indicated it was not anonymous by stating "the name and photo associated with your Google account will be recorded when you upload files and submit this form." A Student Assessment official said the form actually did not record a submitter's name but the fact that it gave the impression that it would may have discouraged anonymous complaints.

The OIG credits the Departments of Student Assessment and School Quality Measurement and Research for seeking the 2018 audit, but finds that CPS's response to it did not go far enough in addressing NWEA test administration concerns.

DISCUSSION

A. Long Durations and Excessive Pauses

The current practice of allowing students to take excessively long times to complete their tests and to pause repeatedly is concerning for several reasons.

Compromised Test Results — At a minimum, even if innocent, such practices can create testing conditions that vary significantly from the conditions under which NWEA's national sample took the test. This can risk distorting results tied to national norms, according to what the OIG has been told.

UNC's Cizek voiced this concern, saying that, "Assuming that these are not students who require accommodations, some of your times are so different that it calls into question the validity of these scores or at least making a comparison to what NWEA

is supposed to measure." As indicated in **Appendix B**, non-Diverse Learners were more likely to take longer tests than Diverse Learners.

The CPS Audit Department also recognized this danger in its April 2018 audit by stating:

> Schools with an elevated number of test irregularities can result in potentially unreliable test scores and inaccurate measures of student progress towards career and college readiness. Such inaccurate data results in CPS not being able to properly plan for instruction, academic supports and resource allocation.

Waste of Instructional Time — Even if innocent, excessively long durations that allow students to take days and days to complete a 53-question test soak up valuable instructional time. Even one seventh grader recognized this, saying that kids were "taking our slow time" to complete their NWEAs, which she conceded could throw off a teacher's lesson plans.

Added Stress for Students — NWEA's untimed nature, combined with its adaptive design, could be adding stress for students who feel compelled to work as long as possible to try to get every question right — even though the test is constructed so that they are expected to get half the questions at their ability level wrong. OIG interviews indicated some students facing high stakes "stressed out" during the test and were afraid to guess. One teacher described one high-scoring student who took more than a week to complete her eighth-grade math test as going through "mental torture."

An Indicator of Cheating or Gaming — As detailed in this Summary Report, excessive durations and pauses can be a result of a variety of cheating and gaming techniques, including the following practices reported by CPS students:

- allowing students to intentionally time-out so they can receive another question. This can result in a 50 percent chance of getting a more favorable question. Pausing multiple times on the same question can increase these odds "considerably," according to NWEA.
- allotting students a certain number of free pauses.
- requiring students to write down the questions and answers of challenging questions on their scratch papers, then collecting this information at the end of each test day and using it in future lessons;
- reading questions or answers on the Reading test aloud to students;

- coaching students on how to answer questions by nodding or shaking the head, pointing to information on the screen, telling a student to read a question again, rephrasing questions or showing students math formulas.

B. Other Concerns

Hurting Kids — Cheating ultimately can shortchange students and set them up for failure.

One mother told the OIG that taking the NWEA was almost a painful experience for her Diverse Learner son, who said he was instructed to repeatedly ask the proctor to pause the test and give him a new question if he did not know an answer. The test "would get harder and harder because they help[ed] him with it," the mother said. Eventually her son couldn't answer the questions and would start shutting down and getting upset, the mom said. But her son emerged with a high test score and now this mother worries that his next school may think he's ready for high school even though his high NWEA scores are the result of cheating. Said this mom: "You are affecting kids in the long run. He is not ready for high school but you are making it look like he is."

In another case, a student had been scoring between the 87th and the 99th percentile when he was a CPS third- and fourth-grader, data showed. But after his mother switched him to a private school that also used NWEA, his scores suddenly plummeted, the mom said. Questioning her son about his CPS scores, the mom finally realized they were the result of cheating by CPS proctors. At his new school, her son is getting evaluated for special education services — something she wonders if CPS should have done.

The OIG is concerned that inflated or distorted test scores may be preventing students from getting the help they need — and their parents from knowing they need help. NWEA acknowledges that an inflated score is "less useful as a diagnostic score."

In addition, one seventh-grade teacher voiced concern that kids who repeatedly time out questions to get new questions could be using that advantage to win selective-enrollment seats. "That's extremely unfair to kids who are legitimately trying to get into selective-enrollment high schools," the teacher said. In addition, students who gamed or cheated their way into selective enrollment schools ultimately could find themselves over their heads academically in their new surroundings.

REACH Impact — Cheating and gaming practices can improperly help boost the REACH scores of the Reading and Math teachers that allow or use such practices, according to CPS's value-added vendor.

The likelihood of the following year's Reading or Math teacher being adversely impacted if he or she honestly administers NWEA is more remote, the vendor said. This is because many factors are used to project a student's growth and a single outlier score would not have a major impact.

"Cheating can pay for a teacher. They could cheat their way to a high growth score, but it will have very little to no impact on the following year's teachers,'' the vendor told the OIG.

However, the OIG tends to doubt that many teachers understand this. That's because several teachers interviewed by the OIG about the NWEA asked the OIG to explain REACH calculations to them. One teacher even complained to this office about a student who entered her classroom with suspiciously high NWEA scores and worried that her REACH evaluation would suffer as a result. Teachers need to understand how a previous year's NWEAs affect them so that, even if they suspect cheating in a prior year, they are not tempted to repeat such practices to prevent being adversely affected by a drop in their students' test scores.

Abuse of Small-Group Testing — Current CPS guidelines do not allow the NWEA testing of general education students in small groups. However, both Diverse Learners and non-Diverse Learners reported being testing in small groups by proctors who broke test administration rules. The OIG is especially concerned that this small-group setting could be abused in the hands of the wrong proctors, particularly when the tests of priority-group students whose results can carry extra weight are involved. Therefore, the OIG recommends that the size of non-Diverse Learner testing groups be among the items monitored by auditors. Auditors also should try whenever possible to monitor Diverse Learner testing rooms to ensure appropriate testing procedures and accommodations are being used.

RECOMMENDATIONS

A. Two Proctors

As mentioned, NWEA's advice to customers who use NWEA for high-stakes purposes is that both a teacher and an additional proctor should monitor student testing. The second proctor should be someone with no direct investment in the performance of those students being tested, NWEA advises. A second proctor protects the integrity of testing results and protects teachers from false accusations of cheating, according to NWEA.

"The notion of teachers proctoring their own students is asking for trouble,'' said Caveon's Maynes. That's because the teachers' stakes are bound up in the students' stakes.

Several CPS teachers endorsed the idea of having two adults proctor the NWEAs during interviews with the OIG. They said they could use the extra help during mechanical difficulties or when students need to go to the bathroom.

In addition, several teachers said two proctors were required during CPS's administration of a now-defunct state test.

However, the OIG notes that two proctors is not a guarantee that cheating or gaming will not occur. At one school, testing protocols were broken by multiple adults who were proctoring a small group of students in one room, according to students. The OIG believes its recommendation to record proctors and to hold one responsible for administering the test properly, as discussed below, will discourage misconduct in rooms with multiple proctors.

B. Recording Proctors

Virtually every gaming or cheating scenario described to the OIG was promoted or allowed by a proctor. Therefore, maintaining an auditable record of who proctors each test is critically important in both deterring and detecting cheating.

This is not a new idea. The TILSA (Technical Issues in Large-Scale Assessment) Test Security Guidebook of May 2013 recommends that each testing entity keep detailed records of test administrators and proctors.

Neither CPS nor NWEA maintain records of who proctors each NWEA test. This puts CPS auditors in the position of flying blind when they audit test administration procedures based on the grade and subject in a school that racked up unusual test results the year before. The OIG believes auditors should be checking test administration procedures of the proctors who administered the tests that produced unusual results. Based on conversations with CPS students and teachers, proctors sometimes have no connection to the tested subject or grade. To audit what could be the wrong classrooms is a misuse of CPS resources.

Test publishers do what they are required to do by their contracts, according to Marc Weinstein, vice president of Caveon Investigative Services. They will do more for districts and states who push them to do more and who get that in their contract. What data should be collected should be raised in the Request for Proposal because "it's all about what the customer establishes as the specification,'' Weinstein said.

The multiple stakes attached to CPS's NWEA test "create a fertile ground for test misconduct. I would think that in an environment of that nature, the entity purchasing the test and using the platform would want to make sure [it] had the capability to collect important data that would help identify testing irregularities," Weinstein said.

Vendors always like to say that they cannot do anything but what they are doing now, said John Fremer, president of Caveon Consulting Services.

However, Fremer asked, if an influential school district like Chicago wanted proctor information, why would the vendor not do it? Fremer said that when he worked previously for one testing company, if Chicago wanted something, company officials would fall over themselves to see if they could do it.

Several large K-12 test vendors can capture a field of information reflecting who launched their tests, if necessary,[15] Weinstein said.

For example, Pearson Assessments can include the logon or name of each test's proctor as a field of information for every test taken, one Pearson executive told the OIG. This feature is optional, based on a client's request.

The American Institutes for Research also can collect data on each test's proctor, and some AIR state tests include the identity of the last proctor of a student's test in their data reported to a state, according to an AIR executive. AIR can include more proctors upon request, the executive said. AIR has even examined proctor information when doing data analytics to check for cheating, the OIG was told.

The proctor fields both Pearson and AIR can provide would allow CPS to analyze its test results by proctor, something that would be especially useful in identifying audit targets.

The person who launches the test, also known as the primary proctor, often is held responsible for the integrity of the test's administration, according to UNC's Cizek. "It would be most typical for the buck to stop with them so if there are any test irregularities they are held accountable," Cizek said.

NWEA suggested that CPS establish a policy that all persons present during a testing session share equal responsibility for ensuring that proper testing practices are followed. **Appendix I** contains a list of other NWEA suggestions.

[15] NWEA said it collects some proctor information but auto-deletes it in nine days. This information is voluminous and "not in a digestible format," NWEA said. Currently NWEA does not review or use it.

So accountability does not turn into a finger-pointing game, the OIG recommends that one proctor be held responsible for the integrity of test results — preferably the proctor without any stakes in the test.

As NWEA is currently constructed, a proctor must resume a test, so tests that exhibit an excessive number of pauses do so with the knowledge of — and possibly even at the instigation of — the proctor. In addition, students cannot intentionally and repeatedly time out a test without a proctor's acquiescence as proctors have to resume the test. Proctors should be given guidance during training sessions on how to avoid excessive pauses and then be held accountable for explaining them if they occur.

According to Caveon's Fremer, "There is incontrovertible evidence over the years that when you add stakes you get more efforts to cheat." So, Fremer said, when districts add stakes, they also should add steps to minimize efforts to cheat.

The OIG believes that a clear record of who proctored each test would go a long way toward minimizing efforts to cheat as well as in helping CPS identify proctors who administer tests in unusual ways.

Asked if it was mechanically possible for NWEA to provide such information, NWEA said it currently "does not have the capability to record the adults present during a test. Although such a feature may be technically feasible, it would need to be developed through NWEA's Product Management process to be evaluated fully and priced."

If NWEA cannot add this field of information to its data, the OIG recommends that CPS explore alternative ways to capture this information electronically that could be folded into its NWEA comprehensive data file. This could include some kind of electronic template that each lead proctor would be required to fill out for every test listing all proctors and students in the room with at least one field of information that would allow this data to be merged with NWEA's comprehensive data file. However, this solution would not be ideal because it would rely on manual data entry and therefore would be more vulnerable to error.

The proctor record should be used to ensure auditors are regularly assigned to observe those proctors who were in the classrooms that produced unusual results. Although the Office of Internal Audit audited many NWEA testing classrooms in 2018 and 2019, who will do such audits in 2020 is uncertain, according to one Department of Student Assessment official. The OIG recommends that these audits continue, but based on an improved flagging system and using revised audit checklists.

C. Timed Tests

Currently, the NWEA is untimed, resulting in some CPS students spending days — in some cases more than a week — on a maximum 53-question test that the average student nationally is completing in about an hour.

The head of investigations for Caveon strongly recommended against using an untimed test in a high-stakes environment.

"In my opinion, high-stakes need a controlled environment and a fixed period of time,'' said Caveon's Weinstein. With an untimed test, where breaks can last hours or days, "There are too many things that can happen during those breaks that would affect the validity of the test results."

Having an untimed, high-stakes test "doesn't make sense because it means that testing session could drag out for literally days and you'd have no idea what's happening when students temporarily stop testing," Weinstein said.

Weinstein was concerned that during the test, teachers could walk around the testing room, observe which concepts students were struggling with and then conduct mini-lessons on those concepts during breaks. Teachers can try to "harvest" questions from students during breaks in the test and then provide instruction in those areas before the test resumes, another Caveon expert said. Students also can discuss test content or communicate electronically about it during breaks, Weinstein said.

Weinstein would prefer to see a test broken into self-contained units or subparts that can be completed in one sitting than to see a test be untimed and paused repeatedly over hours or even days.

In a high-stakes environment, Weinstein said, "Test sessions should have a clear start and end time. If breaks are allowed, they should be logged. Every second of the test should be logged." All keystrokes should be logged with time-stamps, he said.

Based on accounts to the OIG, clearly some students are taking advantage of NWEA's untimed nature.

With stakes so high for students, some students said their colleagues would rather sit on questions for 25 minutes until they time out and are replaced with new questions than risk guessing incorrectly.

One teacher reported hearing concerns from other teachers that their students "looked like they were falling asleep" during their tests. The teacher was told that some kids were "just staring at the screen and spacing out until [a question] timed

out." Another reported seeing students with their heads resting on their desks *during* their tests. She was not sure if these students were napping or waiting for the test to time out.

One eighth-grade teacher estimated that maybe 40 percent of her students take at least five hours to complete their tests; perhaps 5 percent take 10 hours or more. Remember, this is for a maximum 53-question test. For one high-scoring student who took over a week to complete one NWEA, the test was like "mental torture,'' this teacher said. The student was "stressed out" about questions and on at least one day only answered one or two. Sometimes, afraid to make a mistake, she just "sat on" questions rather than guess, the teacher said.

According to this teacher, "It's crazy to me that [the NWEA] is unlimited and then these kids move on thinking all their tests will be unlimited. Number two, for some of these kids, it just stresses them out. They're just sitting there. I think if there was a time limit they wouldn't second guess themselves so much."

Cizek also recommended setting time limits for general education students on high-stakes test. With high-stakes, untimed tests, "Educators have absolutely no incentive to tell kids to finish. If I know my rating depends on this, I'm gonna tell them to keep checking their work and taking their time," Cizek said.

He also noted that students who take a long time on their tests may continue testing while their classmates receive instruction. These students, often among the lowest-performing, wind up losing more instructional time than their peers even though they could be the students who need it most.

Cizek suggested setting the time limit at two standard deviations above the test's national norm for general education students. Some students are currently taking longer than they really need to finish their NWEAs, Cizek said. If the cutoff was set at a time at which 98 percent of the population is finishing now, Cizek was confident that almost all students would be able to complete the test.

NWEA told the OIG its tests should be administered in "a reasonable time as guided by the test duration norms" produced by NWEA but it does not recommend imposing time limits on individual students because the test is meant to be untimed.

"When test durations are unreasonable or excessive, however, NWEA may support establishing benchmark expectations for the average time it takes the students to complete an assessment,'' NWEA said. (See **Appendix I** for NWEA suggestions on setting duration benchmark expectations.)

Among NWEA's suggestions were that CPS establish a policy that the district has a right to retest students with a third-party proctor "when there is a deviation that would call the validity of a student's test results into question. Such deviations could include: excessive test durations without an IEP accommodation; excessive pauses or time outs during a test; or inconsistencies in testing practices between terms without sufficient rationale."

However, because Spring NWEAs are given so close to the end of the school year, it may well be difficult to identify outlier durations or pauses and then organize a retest before the end of that school year.

The OIG believes that the multiple high stakes attached to CPS's NWEA results makes their untimed nature susceptible to abuse. The OIG recommends that CPS consult NWEA on how to maintain test validity while setting a time limit on existing NWEA tests for general education students. If NWEA rejects specific limits, CPS should consider setting rules for the retest of students whose durations exceed a certain amount of time — or even weigh switching to a different, timed test.

D. Limit Pauses

It's clear from systemwide pause data that some proctors are allowing an excessive number of pauses during NWEA tests.

While some pauses could be benign, the way they are clustered in certain schools and the comments of students at those schools as to how they occurred indicates some pauses are being used in an attempt to game the test. In addition, the OIG's data analysis shows both Diverse Learner and non-Diverse Learner students with more pauses were more likely to achieve high gains even though, as one expert told the OIG, there's no educational reason why pauses should improve the scores of non-Diverse Learner students.

CPS needs to take concrete steps to limit pauses so they cannot be used to try to provide students with more favorable questions.

The OIG recommends that CPS's NWEA training materials explain exactly how the test's pause function works and offer guidance on how to handle students with excessive pauses so that proctors and schools can be held accountable if tests show excessive pauses. Pauses for lunch or bathroom breaks shouldn't be used strategically to skip hard questions.

NWEA also advises that CPS retest students with excessive pauses (See **Appendix I**). If this policy is instituted, proctors should warn students in advance of their tests that they could be retested if their tests exhibit too many pauses.

Cases of excessive pauses should be investigated and proctors of tests that showed high numbers of pauses should be audited the following year.

In response to OIG questions, NWEA recommended that CPS encourage practices that diminish the need to pause the test, such as providing rest room breaks prior to starting the test.

NWEA also suggested that proctors maintain a "proctor log" to document each time a test is paused or resumed during a session. "If this is too onerous, require that instances of excessive pausing be documented,'' NWEA said. The OIG is not clear how a proctor would know at what point pauses became excessive unless he or she kept a cumulative count during the test, which seems very time-consuming. (Again, see **Appendix I** for further NWEA suggestions.)

The OIG would prefer to see the pause count of each test included with the comprehensive data file provided to CPS. In addition, ideally, NWEA data should distinguish between pauses initiated by proctors and time-outs, which are not initiated by proctors.

E. List Penalties

In its April 2018 Audit, Audit Department officials specifically recommended that CPS update the NWEA Test Security Agreement which every test administrator and proctor must sign so that it includes language on "the consequences of not adhering to the rules." One Student Assessment official said such language was never folded into the agreement because statements in the agreement "are affirmative action statements that confirm appropriate actions."

The OIG does not find this a valid reason for not including the penalties for improper test administration in the Test Security Agreement. Proctors and test administrators need to know the consequences of breaking the rules. Including this in a document proctors and others must sign creates a record that they were given this warning.

F. Improve Training and Exit Slip

Following the April 2018 Audit, Student Assessment created a training PowerPoint for School Test Coordinators to use in training proctors.

The OIG found this training PowerPoint inadequate in that it did not explain how NWEA pauses and time-outs work. It offered no guidance on national duration norms. It did not explain new changes in the Accommodations Matrix. Its explanation of NWEA's disengagement feature was partially incorrect. And the OIG was not listed as an office that could be called with complaints, anonymous or not, about test irregularities.

Clearly, more guidance is needed. For example, one teacher didn't know a student could intentionally sit on a question until it timed out. She also didn't think her seventh graders were "savvy enough" to figure this out, but then said, "Now that you're saying it, I guess it could be a possibility."

This teacher added: "If pausing tests and timing out is getting kids different questions or passages then, yeah, we should definitely be informed of that. . . . Not only should we be informed to watch for that but this is extremely unfair to kids who legitimately are trying to get into a selective enrollment high school and their spots are being taken by kids who are, it sounds like, kinda cheating."

Although NWEA says it shares duration norms with clients to guide them, apparently this information is not being shared with CPS teachers. Not one questioned by the OIG had any idea what the national duration norms were. Said one: "If there's a certain amount of time you are expecting, that should be explicit."

The five-question exit slip proctors had to pass by getting at least four of five questions correct did not cover enough information about proper and improper test administration procedures, especially in light of the fact that this exit slip was created after an audit had criticized CPS's NWEA "preventive controls."

Some teachers told the OIG that they would appreciate more guidance on how the test works and proper versus improper procedures. One even said exit slips should have more questions.

The TILSA Test Security Guidebook recommends that test training material "provide clear examples of what behavior is unacceptable." It goes on to add:

> One source of cheating by staff is lack of understanding about what are acceptable and unacceptable behaviors and the important reasons behind the need for accurate test results.

The OIG recommends that CPS upgrade its training materials to cover both acceptable and unacceptable behaviors. Otherwise, teachers can continue to make the argument, as one did to the OIG, that CPS training does not specifically say certain tactics are improper.

Proctors should be informed that allowing students to pause a test merely to get another question is improper. CPS may want to adjust its retest policy and advise proctors to alert students that they could face consequences, such as retests, should they attempt to improperly use the pause function. Proctors should be warned to be on the lookout for students who appear to be intentionally timing-out the test. They should be offered guidance on how to handle what appears to be intentional time-outs or excessive requests for pauses. Strict rules for handling student scratch paper

should be shared, perhaps by requiring only a second proctor who has no stakes in the test to dispose of it.[16]

CPS also may want to provide more than five exit-slip questions and to expect more than an 80 percent accuracy rate on an exit slip tied to such a high-stakes test.

G. Consult a Test Security Expert

Investigations into cheating allegations can be time-consuming, difficult and traumatic. To get to the truth, investigators may want to question students, possibly putting kids into the uncomfortable position of having to make negative comments about their teachers. Or, teachers may have to make negative comments about their students.

A more prudent approach is to be as proactive as possible on the front end by setting clear policies, providing thorough training, explaining to all parties the consequences of breaking test administration rules, and maintaining records (such as proctor and pause information) that can help detect irregularities and be swiftly used if an investigation is deemed warranted.

As Caveon's Fremer put it: "You don't want to catch them misbehaving. You want them not to misbehave."

Given all the stakes attached to NWEA, the integrity of its results is paramount. Although CPS attempted some reforms in the wake of an audit calling for tougher detective and preventive controls, those reforms did not go far enough. Therefore, the OIG recommends that CPS hire a test security expert for guidance in addressing the concerns outlined in this report. This is a highly technical area; the input of an independent, highly-qualified expert is warranted.

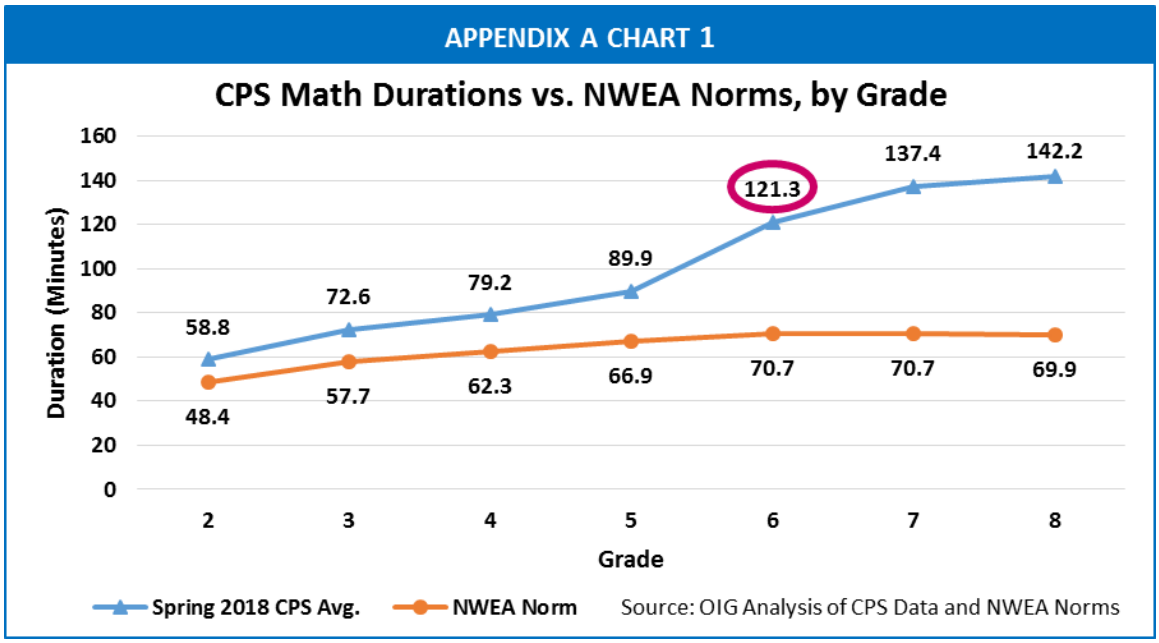
The OIG believes that NWEA does not currently provide CPS with enough information about CPS tests and testing conditions to discourage test irregularities or to adequately and readily detect and investigate unusual test results. A test security expert could work with CPS on such concerns as: setting a time-limit on CPS NWEA tests for general-education students or, at a minimum, finding a way to reduce excessive durations; reducing excessive pauses; improving the current assessment grid for identifying test results warranting an audit; improving auditor questions; bolstering training procedures, exit slips and the Test Security Agreement; ensuring that Math and Reading teachers are not the sole proctors of

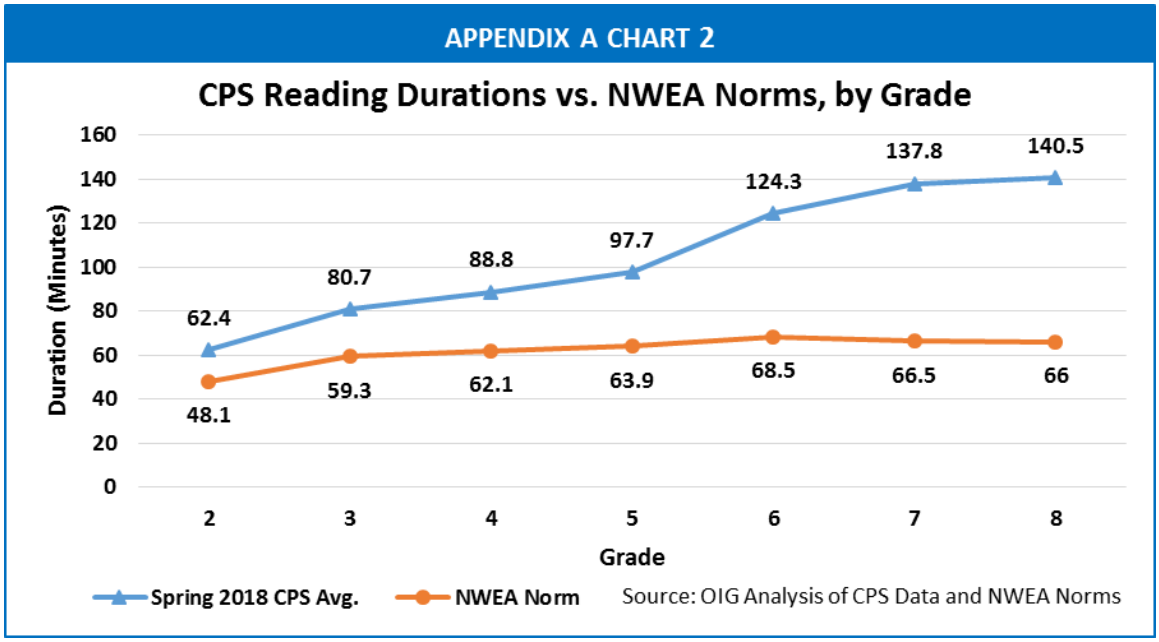
[16] One Caveon test security expert recommended that all scratch paper be a distinctive color, such as bright pink, so that it would be noticeable if a proctor kept it.

their students; obtaining additional data from NWEA such as information on pauses versus timeouts, the number of days tested and auditable information on proctors.

The current NWEA first option to renew, worth up to $2.2 million, ends June 30, 2020. With assistance from the OIG, CPS should hire this test security expert in time to help CPS incorporate NWEA security improvements into any second renewal with NWEA, or preferably even sooner — by the Spring 2020 tests. If NWEA cannot provide needed reforms, CPS should use this test security expert to help the district write ~~an~~a Request for Proposal for a new test contractor. The OIG should be kept apprised of any NWEA contracting changes or RFPs for a new test contractor resulting from this report.

Appendix A: CPS Durations vs. NWEA National Norms, by Grade





Appendix B: Spring 2018 Tests by Duration and Diverse Learner Status

Duration (Minutes)	Total	0 to 75*	76 to 120	121 to 180	181 to 240	241 to 300	301 to 360	Greater than 360
Total Tests	9,223320,561	96,238105,682	04,297111,308	54,96368,386	,99223,010	7,53983 9	68276 4	1,51257 2
Tests by DLs	,87445,243	19,14221,043	2,77613,663	5,7987,214	2,079189	682723	50257	47154
Tests by non-DLs	7,349275,318	77,09684,639	91,52197,645	58,16561,172	,91320,821	,8577,116	43250 7	1,36518
% of DL Tests		45.7146.51%	30.5120%	16.2315.95%	4.9684%	..6360%	0.6057%	0.3534%
% of non-DL Tests		29.9630.74%	35.5647%	22.6022%	7.7456%	.6658%	0.9591%	0.5352%

*The NWEA blog "Testing Duration: How Long is Too Long to Spend on the Map Growth Assessment?" says that, in general, NWEA expects students to complete a MAP Growth test in about 45 to 75 minutes.

Note: The OIG identified Diverse Learners using 2017-18 data from CPS's "SPED" Special Education indicator, which is based on whether the student has an Individualized Education Program. The OIG used 2017-18 SPED indicator data.

Source: OIG Analysis of CPS and NWEA Data from Grade 3-8 Tests

Appendix C: Top 25 Spring 2018 Average Test Durations Most Over National Norm

	School	Gd.	Subject	Tests	Avg. Pauses	Avg. Duration (Hr: Min)	NWEA Norm (Hr: Min)	% of Norm
1	Dixon	7	Mathematics	91	9.49	7:06	1:10	603%
2	Jensen	4	Reading	42	9.12	5:49	1:02	563%
3	Dixon	8	Reading	98	11.64	6:02	1:06	550%
4	Lyon	7	Reading	163	13.59	5:35	1:06	504%
5	Dixon	6	Mathematics	60	8.27	5:49	1:10	494%
6	Casals	5	Reading	45	4.11	5:13	1:03	491%
7	Fiske	8	Mathematics	40	5.15	5:37	1:09	482%
8	Burnside	8	Reading	45	9.91	5:08	1:06	468%
9	Cullen	8	Reading	20	7.10	5:04	1:06	461%
10	Casals	8	Reading	43	3.47	4:57	1:06	451%
11	Dixon	8	Mathematics	99	10.88	5:14	1:09	449%
12	Lyon	7	Mathematics	165	13.24	5:12	1:10	441%
13	CICS - Avalon	7	Reading	58	~~/A*~~5.29	4:53	1:06	441%
14	Fiske	3	Mathematics	41	2.73	4:10	0:57	434%
15	CICS – Wash. Park	6	Reading	54	~~/A*~~4.28	4:54	1:08	430%
16	CICS – Wash. Park	5	Reading	30	~~/A*~~3.37	4:34	1:03	430%
17	Casals	4	Reading	42	4.10	4:26	1:02	429%
18	Casals	7	Reading	42	3.55	4:41	1:06	424%
19	Aldridge	7	Mathematics	14	4.64	4:56	1:10	419%
20	Woodson	8	Reading	24	2.92	4:36	1:06	419%
21	Dixon	7	Reading	91	6.~~15~~14	4:37	1:06	418%
22	CICS – Wash. Park	7	Reading	47	~~/A*~~3.34	4:33	1:06	411%
23	Cullen	7	Reading	16	5.31	4:33	1:06	411%
24	Morton	7	Reading	29	6.66	4:33	1:06	411%
25	Lavizzo	8	Mathematics	40	6.58	4:44	1:09	407%

~~*This school's pause data was not provided to the OIG.~~

Note: Percents of National Norm are rounded. There are no ties.

Source: OIG Analysis of CPS and NWEA data.

Appendix D: OIG Analyses of the Relationship Between Test Duration and Likelihood of Unusually High Growth, by Diverse Learner Status

Percent of non-Diverse Learner Students' Tests with Unusually High Growth, by Test Duration			
Duration (Minutes)	**Total Tests**	**Tests w/ z-score ≥ 2**	**% w/ z Score ≥ 2**
All Tests	7,349 25 2,740	4,546 460	1.77 76%
0 to 75	7,096 75, 328	541 524	0.70%
76 to 120	1,521 89, 993	1,242 219	1.36 35%
121 to 180	8,165 57, 351	1,300 266	2.24 21%
181 to 240	9,913 633	733 727	3.68 70%
241 to 300	5,857 744	384 381	5.60 65%
301 to 360	2,432 369	177 175	7.28 39%
Over 360	1,365 322	169 168	2.38 71%

Percent of Diverse Learner Students' Tests with Unusually High Growth, by Test Duration			
Duration (Minutes)	**Total Tests**	**Tests w/ z-score ≥ 2**	**% w/ z Score ≥ 2**
All Tests	41,874 37,313	1,120 969	2.67 60%
0 to 75	19,142 16,413	222 185	1.16 13%
76 to 120	12,776 11,657	317 269	2.48 31%
121 to 180	1,798 294	313 281	4.60 46%
181 to 240	2,079 1,919	154 131	416.83%
241 to 300	682 650	68 61	9.97 38%
301 to 360	250 238	26 23	0.40 9.66 %
Over 360	147 142	20 19	3.61 38%

Note: The OIG identified Diverse Learners using 2017-18 data from CPS's "SPED" Special Education indicator, which is based on whether the student has an Individualized Education Program. The OIG used 2017-18 SPED indicator data.

Source: OIG Analysis of CPS Data from Grade 3-8 NWEA Tests. See: Methodology for an explanation of the OIG's analysis of student growth from Spring 2017 to Spring 2018.

Appendix E: Spring 2018 Tests by Pauses* and Diverse Learner Status

Pauses	Total	0	1-4	5-9	10-14	15-19	20+
Total Tests	~~162,398~~302,993	~~125,629~~145,424	~~125,972~~145,388	~~,273~~10,524	1,~~033~~149	~~276~~290	~~215~~218
Tests by DLs	~~36,797~~42,760	~~16,403~~19,014	~~18,896~~21,944	1,~~369~~649	~~92~~112	~~22~~26	15
Tests by non-DLs	~~125,601~~260,233	~~109,226~~126,410	~~107,076~~123,444	~~,904~~8,875	~~941~~1,037	~~254~~264	~~200~~203
% of DL Tests		44.~~58~~47%	51.~~35~~32%	3.~~72~~86%	0.~~25~~26%	0.06%	0.04%
% of non-DL Tests		48.~~42~~58%	47.~~46~~44%	3.~~50~~41%	0.~~42~~40%	0.~~11~~10%	0.~~09~~08%

*NWEA's data does not distinguish between pauses and time-outs.

The OIG did not receive pause data for some tests, which are excluded from this analysis.

Note: The OIG identified Diverse Learners using 2017-18 data from CPS's ~~"SPED"~~ Special Education indicator, which is based on whether the student has an Individualized Education Program~~. The OIG used 2017-18 SPED indicator data~~.

Source: OIG Analysis of CPS Data from Grade 3-8 Tests

Appendix F: Top 25 Spring 2018 Pauses* per Test

	School	Grade	Subject	Tests	Avg. Duration (Hr:Min)	Most Pauses* of One Test	Avg. # of Pauses*
1	Lyon	7	Reading	163	5:35	46	13.6
2	Lyon	7	Math	165	5:12	64	13.2
3	Dixon	8	Reading	98	6:02	24	11.6
4	Wacker	7	Math	20	2:26	23	11.4
5	Dixon	8	Math	99	5:14	27	10.9
6	Burnside	8	Reading	45	5:08	33	9.9
7	Lyon	8	Math	168	4:38	36	9.9
8	Oglesby	8	Math	40	2:30	65	9.8
9	Wacker	6	Reading	33	2:36	55	9.5
10	Dixon	7	Math	91	7:06	21	9.5
11	Pullman	7	Math	37	2:15	48	9.4
12	Jensen	4	Reading	42	5:49	35	9.1
13	Burnside	8	Math	45	4:32	29	8.7
14	Fernwood	8	Math	25	3:06	14	8.7
15	Fernwood	7	Math	20	3:06	13	8.6
~~16~~15	Wacker	7	Reading	20	2:11	23	8.6
17	Oglesby	7	Math	43	2:07	61	8.4
18	Oglesby	8	Reading	40	2:36	21	8.3
19	Dixon	6	Math	60	5:49	19	8.3
20	Wacker	8	Math	18	3:05	15	8.2
21	Fernwood	8	Reading	25	2:28	14	8.2
22	Lavizzo	8	Reading	40	4:00	20	7.8
23	Fernwood	6	Math	29	2:45	14	7.8
24	Lyon	8	Reading	167	4:23	44	7.7
25	Pullman	8	Reading	38	2:44	30	7.7

*The NWEA MAP test can be paused using a proctor's administrative console, paused using a command on the student's computer or tablet, or timed out due to 25 minutes of inactivity, all of which result in a new question when the test is resumed. NWEA's records currently cannot distinguish between types of pauses.

Source: OIG Analysis of CPS and NWEA Data

Appendix G: OIG Analyses of the Relationship Between Pauses/Time-Outs and Likelihood of Unusually High Growth, by Diverse Learner Status

Percent of non-Diverse Learner Students' Tests with Unusually High Growth, by Number of Pauses*

Pauses*	Total Tests	Tests w/ z-score ≥ 2	% w/ Z Score ≥ 2
All Tests	,601238,869	4,161319	1.8481%
0	,226114,957	1,471517	1.3532%
1-4	,076114,148	2,231318	2.0803%
5-9	048,359	332358	4.2028%
10+	,395405	127126	9.108.97%

Percent of Diverse Learner Students' Tests with Unusually High Growth, by Number of Pauses*

Pauses*	Total Tests	Tests w/ z-score ≥ 2	% w/ Z Score ≥ 2
All Tests	,79735,254	1,013937	2.7566%
0	,40315,200	277241	1.6959%
1-4	3,896481	630593	3.3321%
5-9	1,369426	9693	7.016.52%
10+	129147	10	7.756.80%

*NWEA's data does not distinguish between pauses and time-outs.

The OIG did not receive pause data for some tests, which are excluded from this analysis.

Note: The OIG identified Diverse Learners using 2017-18 data from CPS's "SPED" Special Education indicator, which is based on whether the student has an Individualized Education Program. The OIG used 2017-18 SPED indicator data.

Source: OIG Analysis of CPS Data from Grade 3-8 NWEA Tests. See: Methodology for an explanation of the OIG's analysis of student growth from Spring 2017 to Spring 2018.

Appendix H: Clusters of High-Growth Students with Less than a One in a Million Chance of Occurring in a Random Sample of CPS Students in that Grade and Subject

School	Grade	Subject	Avg. Duration (Hr: Min)	Avg. Pauses	# of Tests	Tests w/ Growth Z-Score 2+	% w/ Growth Z-Score 2+	CPS % w/ Growth Z-Score 2+*
Chavez	3	Math	1:54	0.6	~~104~~10 3	29	~~27.9~~28.2%	2.2%
Faraday	4	Math	1:~~26~~27	0.1	~~27~~26	16	~~59.3~~61.5%	2.2%
Ruggles	7	Reading	2:26	1.3	27	14	51.9%	1.6%
Clinton	8	Math	2:~~35~~38	1.1	~~93~~75	~~22~~20	~~23~~26.7%	2.3%
Greeley	6	Math	2:04	0.2	57	16	28.1%	2.0%
Farnsworth	5	Math	1:~~52~~53	0.~~4~~5	~~54~~52	~~17~~16	~~31.5~~30.8%	2.3%
~~Greeley~~Pickard	~~6~~8	Math	~~2:03~~3:42	~~0.2~~1.8	~~58~~47	~~16~~15	~~27.6~~31.9%	2.~~0~~3%
Budlong	8	Math	4:34	5.8	76	18	23.7%	2.3%
~~Pickard~~	~~8~~	~~Math~~	~~3:38~~	~~1.8~~	~~48~~	~~15~~	~~31.3%~~	~~2.3%~~
Bouchet	6	Math	4:31	3.8	56	15	26.8%	2.0%
~~Peck~~Spencer	~~3~~5	Math	1:~~34~~40	~~1~~0.4	~~163~~48	~~23~~14	~~14.1~~29.2%	2.~~2~~3%
Dixon	6	Math	5:52	8.4	~~57~~56	14	~~24.6~~25.0%	2.0%
~~Spencer~~KIPP-One	5	Math	~~1:37~~2:10	~~0.3~~N/A**	~~51~~82	~~14~~17	~~27.5~~20.7%	2.3%
~~KIPP-One~~Peck	~~5~~3	Math	~~2:11~~1:34	~~N/A**~~ 1.4	~~85~~162	~~17~~22	~~20.0~~13.6%	2.~~3~~2%
~~Von Linne~~Mason	3	~~Math~~Reading	~~3:43~~2:18	1.3~~.7~~	~~59~~35	~~14~~11	~~23.7~~31.4%	~~2.2~~1.8%
~~Mason~~Von Linne	3	~~Reading~~ Math	~~2:16~~3:43	~~1.2~~3.7	~~38~~57	~~11~~14	~~28.9~~24.6%	~~1.8~~2.2%
Ravenswood	3	Math	1:36	1.5	61	14	23.0%	2.2%
~~Fiske~~Dixon	~~4~~8	Math	~~1:49~~5:12	11.0~~.8~~	~~48~~89	~~12~~16	~~25~~18.0%	2.~~2~~3%
~~Dixon~~Fiske	~~8~~4	Math	~~5:13~~1:49	~~10.9~~0.8	~~96~~45	~~16~~12	~~16~~26.7%	2.~~3~~2%
Whistler	5	Reading	2:~~51~~55	1.2	~~33~~31	9	~~27.3~~29.0%	1.~~5~~4%
Spencer	6	Math	~~1:57~~2:00	0.2	~~47~~45	11	~~23~~24.4%	2.0%

School	Grade	Subject	Avg. Duration (Hr: Min)	Avg. Pauses	# of Tests	Tests w/ Growth Z-Score 2+	% w/ Growth Z-Score 2+	CPS % w/ Growth Z-Score 2+*
~~West Ridge~~Hayt	~~6~~4	Math	~~2:57~~1:04	~~4.3~~0.2	~~72~~105	~~13~~16	~~18.1~~15.2%	2.~~0~~2%
Esmond	4	Reading	1:59	1.0	24	8	33.3%	1.~~6~~5%
~~Hayt~~	~~4~~	~~Math~~	~~1:04~~	~~0.2~~	~~109~~	~~16~~	~~14.7%~~	~~2.2%~~
~~Ariel~~	~~3~~	~~Math~~	~~1:15~~	~~0.2~~	~~56~~	~~12~~	~~21.4%~~	~~2.2%~~
Sutherland	8	Math	2:50	0.9	54	12	22.2%	2.3%
Ariel	3	Math	1:15	0.2	56	12	21.4%	2.2%
Sherwood	3	Math	1:18	0.3	28	9	32.1%	2.2%
Bouchet	3	Reading	3:45	3.0	45	10	22.2%	1.8%
Bouchet	3	Math	3:09	3.2	48	11	22.9%	2.2%
Caldwell	5	Reading	2:21	1.3	20	7	35.0%	1.4%
Ruggles	3	Reading	1:13	0.5	35	9	25.7%	1.8%
~~Caldwell~~	~~5~~	~~Reading~~	~~2:21~~	~~1.3~~	~~20~~	~~7~~	~~35.0%~~	~~1.5%~~
~~Bouchet~~West Ridge	~~3~~6	Math	~~3:13~~2:54	~~3.2~~4.0	~~50~~69	~~11~~12	~~22.0~~17.4%	2.~~2~~0%
Young ES	3	Reading	1:27	1.2	~~92~~91	13	14.~~1~~3%	1.8%
~~Sherwood~~	~~3~~	~~Math~~	~~1:25~~	~~0.5~~	~~30~~	~~9~~	~~30.0%~~	~~2.2%~~
~~Bouchet~~	~~8~~	~~Math~~	~~3:39~~	~~2.0~~	~~39~~	~~10~~	~~25.6%~~	~~2.3%~~
~~Bouchet~~	~~3~~	~~Reading~~	~~3:42~~	~~3.1~~	~~50~~	~~10~~	~~20.0%~~	~~1.8%~~
Bouchet	8	Reading	4:11	4.0	39	9	23.1%	1.7%
Cather	7	Reading	2:~~19~~16	0.~~5~~4	~~34~~32	8	~~23.5~~25.0%	1.6%
Bell	7	Math	2:10	0.7	88	13	14.8%	2.1%
Castellanos	8	Math	2:01	0.8	~~98~~97	14	14.~~3~~4%	2.3%
~~Bell~~Penn	~~7~~4	~~Math~~Reading	~~2:10~~1:38	0.~~7~~1	~~92~~24	~~13~~7	~~14.1~~29.2%	~~2.~~1.5%
~~Penn~~Sutherland	~~4~~6	Reading	~~1:37~~2:39	~~0.1~~2.4	~~25~~60	~~7~~9	~~28~~15.0%	1.~~6~~3%
Dixon	7	Math	7:04	9.4	82	12	14.6%	2.1%
Hefferan	6	Math	3:50	2.6	22	7	31.8%	2.0%

School	Grade	Subject	Avg. Duration (Hr: Min)	Avg. Pauses	# of Tests	Tests w/ Growth Z-Score 2+	% w/ Growth Z-Score 2+	CPS % w/ Growth Z-Score 2+*
Hefferan	7	Reading	3:59	1.9	27	7	25.9%	1.6%
~~Sutherland~~	~~6~~	~~Reading~~	~~2:37~~	~~2.3~~	~~62~~	~~9~~	~~14.5%~~	~~1.4%~~
Lavizzo	3	Reading	1:59	1.5	~~36~~35	8	22.~~29~~%	1.8%
Spencer	7	Reading	2:54	1.5	40	8	20.0%	1.6%
Pullman	3	Reading	1:13	0.1	~~50~~48	9	18.~~08~~%	1.8%
Earle	6	Math	3:44	4.4	33	8	24.2%	2.0%
Norwood Pk.	4	Math	1:15	0.0	~~54~~52	10	~~18.5~~19.2%	2.2%
~~Dixon~~	~~7~~	~~Math~~	~~7:08~~	~~9.5~~	~~87~~	~~12~~	~~13.8%~~	~~2.1%~~
KIPP-Ascend	7	Math	3:~~55~~45	~~N/A**~~ 1.6	87	12	13.8%	2.1%
~~Earle~~	~~6~~	~~Math~~	~~3:40~~	~~4.4~~	~~34~~	~~8~~	~~23.5%~~	~~2.0%~~
~~Kilmer~~	~~4~~	~~Math~~	~~1:19~~	~~0.3~~	~~69~~	~~11~~	~~15.9%~~	~~2.2%~~
Mcnair	6	Math	2:56	2.7	35	8	22.9%	2.0%
~~Spencer~~	~~7~~	~~Reading~~	~~2:46~~	~~1.4~~	~~44~~	~~8~~	~~18.2%~~	~~1.6%~~
Jamieson	5	Math	1:42	0.2	83	12	14.5%	2.3%
~~New Sullivan~~Bouchet	~~4~~5	Math	1:~~32~~56	~~0.8~~1.3	~~45~~55	~~9~~10	~~20.0~~18.2%	2.~~2~~3%
~~Bouchet~~Earle	~~5~~7	Math	~~1:56~~4:28	~~1.3~~6.9	~~56~~25	~~10~~7	~~17.9~~28.0%	2.~~3~~1%
~~Shoop~~Stagg	8	~~Math~~Reading	3:~~34~~57	~~3.6~~4.2	~~34~~30	~~8~~7	23.~~5~~3%	~~2.3~~1.7%
~~Earle~~Esmond	~~7~~5	Math	~~4:26~~3:18	~~6.9~~2.0	~~26~~23	7	~~26.9~~30.4%	2.~~1~~3%
~~Skinner~~Esmond	~~6~~5	~~Math~~Reading	2:~~59~~43	~~3.6~~1.5	~~118~~23	~~13~~6	~~11.0~~26.1%	~~2.0~~1.4%
~~Stagg~~Waters	~~8~~4	~~Reading~~ Math	~~3:55~~1:19	~~4.2~~0.9	~~31~~75	~~7~~11	~~22.6~~14.7%	~~1.7~~2.2%
Ruggles	3	Math	1:06	0.4	36	8	22.2%	2.2%

*For tests taken in the same subject by students of the same grade level.

**This school's pause data was not provided to the OIG.

Note: These clusters of high-gaining students are sorted by their probability of occurring in a random sample of CPS students testing in the same grade and subject, although the OIG is not including the exact probability in this chart. The school/grade/subject combinations are sorted from the least probable down.

School	Grade	Subject	Avg. Duration (Hr: Min)	Avg. Pauses	# of Tests	Tests w/ Growth Z-Score 2+	% w/ Growth Z-Score 2+	CPS % w/ Growth Z-Score 2+*

See: Methodology for an explanation of the OIG's analysis of student ~~NWEA~~ growth from Spring 2017 to Spring 2018.

Appendix I: NWEA Suggestions on Reducing Pauses and Time-Outs and Setting Duration Benchmarks

Q: Does NWEA have any suggestions on how CPS or NWEA could reduce the instances of a proctor pausing a test to produce a new question because a student did not know the answer to the question on the screen?

A: NWEA recommends the following:

1. Establish a CPS policy that the district has the right to retest students, with a third party proctor, when there is a deviation that would call the validity of a student's test results into question. Such deviations could include: excessive test durations without an IEP accommodation; excessive pauses or time outs during a test; or inconsistencies in testing practices between terms without sufficient rationale.
2. Establish a CPS policy that students are not to be given commands or passwords that would give them the ability to pause a test on their own. CPS policy should make it clear that only the test proctor can pause a test.
3. Establish a CPS policy that all persons present during a testing session share equal responsibility for ensuring that proper testing practices are followed and that "cheating or gaming" practices do not occur.
4. Encourage CPS test administration practices that would diminish the need to pause assessments. For example, encourage teachers to provide for a rest room break prior to the start of testing.
5. Require that proctors maintain a "proctor log" to document each time a test is paused or restarted during a session. If this is too onerous, require that instances of excessive pausing be documented.

Q: Does NWEA have any suggestions on how CPS or NWEA might reduce the instances of students "timing out" questions so they can get new questions?

A: In general, there is no reason for a student or proctor to allow a question to "time out." The district could establish a policy or procedure requiring proctors to monitor the progress of students' tests to identify possible instances of "cheating or gaming" practices. Allowing items to time out should be considered a possible "gaming" practice. The district could require such instances to be documented by the proctor.

Q How would a school district determine what to set as a duration "benchmark expectation?" How would a district impose this "benchmark expectation?"

A: In establishing a test duration for an applicable grade or subject, a school district could consider NWEA's test duration norms as a benchmark. A school district should further evaluate a reasonable test duration based on the following factors, among others: the amount of classroom time that is consumed by current test durations; the disruption to school schedules; and the ability to compare test results to NWEA norms. Once established, a school's testing durations should generally be expected to fall within their benchmarks. A school district's policy should also make it clear that test conditions should be consistent from term to term. Notably, a school district should not impose time limits on individual students because some students need more time than others to complete a test.