OIG PERFORMANCE REVIEW:

NWEA Test Administration

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

OVERVIEW

- **Performance Review:** The OIG's Performance Analysis Unit, which uses data to identify broad issues, analyzed three years of data, focusing on Spring 2018; interviewed 20 students and 10 teachers, almost all in schools with unusual results; and spoke with testing experts.

- **Key Takeaways:** The OIG found concerning levels of unusually long test durations, high pause counts and other irregularities. Especially long tests and high pause counts had a higher occurrence of unusually high gains. Though high-stakes, the test is administered with insufficient security protocols. Improved administration is needed for CPS to have the most confidence possible in test results.

- **Key Recommendations:** Reduce long durations, preferably by setting a time limit; trim high pause counts; institute proctor and proctor-data reforms; implement tighter controls with the help of a test security expert. If the test vendor cannot provide recommended enhancements, consider another test vendor with the test security expert's help.

TEST STAKES

Students

- Helps determine promotion from 3rd, 6th and 8th grades.
- 7th-grade results affect admissions to SEHS and other competitive HS programs; results from other grades used for grade 5-8 SEES admissions.

Teachers

- 20% of Reading and Math teacher evaluations are tied to student growth based on a Value-Added Model.

Principals

- 35% of principal evaluations are tied to test results.
- Test results impact Independent School Principal status.

Schools

- 60% of an elementary school's SQRP level is based on test results.
- Test results of students from four "Priority Groups" carry extra SQRP weight.

DISTINCTIVE FEATURES

- **Length:** Math is 52 or 53 questions; Reading is 42 or 43.
- **Untimed:** Students can take as long as they want. The test vendor generally expects the test to be completed in 45 to 75 minutes.
- **Adaptive:** Correct answers result in harder questions; wrong answers result in easier questions. Students are expected to get half the questions at their achievement level right and half wrong.

PAUSES, TIME OUTS AND DURATION

- **Pauses:** The test can be paused for a break, which replaces the current question with a new question of similar difficulty.
- **Time Outs:** After 25 minutes of inactivity, generally the test will time out. When students are logged back in and continue testing, they are given a new question of similar difficulty. Note: pause counts do not distinguish between pauses and time outs.
- **Duration:** Each test's reported duration only reflects time spent on answered questions. Time spent on questions that were paused or timed out is excluded.

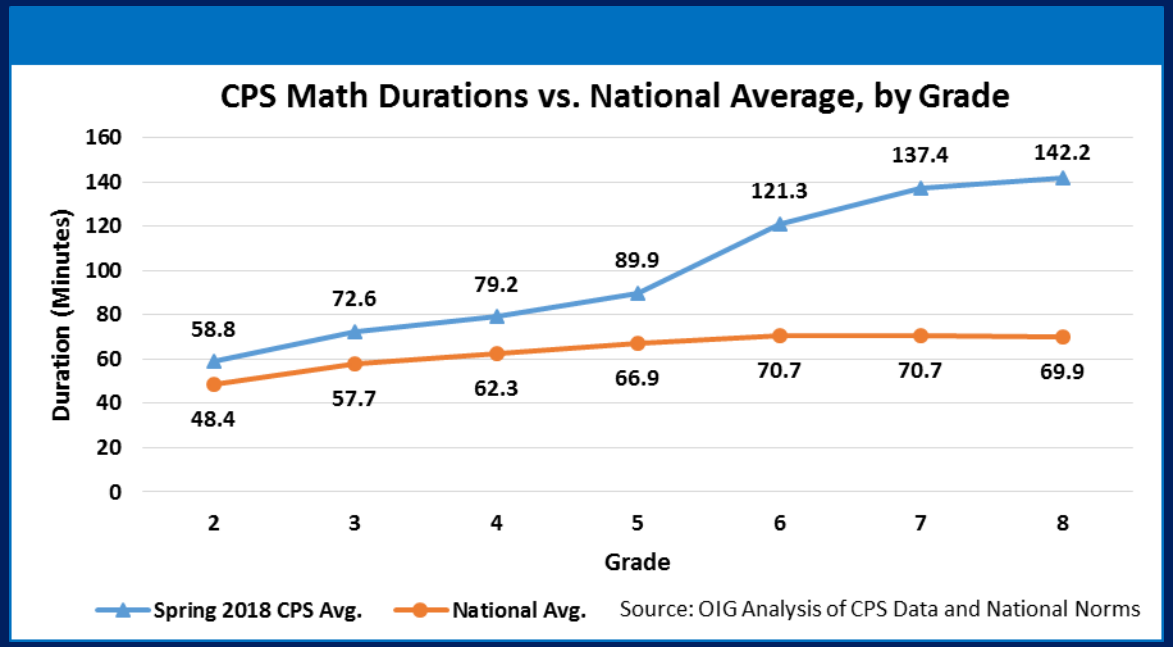
DATA FINDINGS

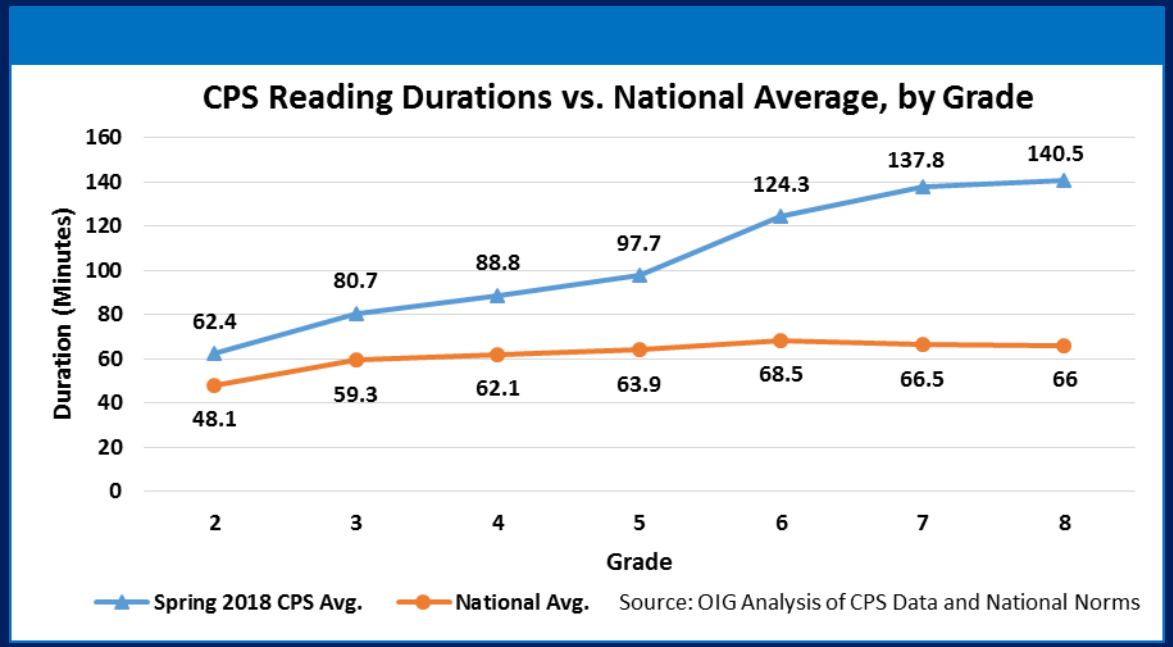
1) Durations

2) Pauses

3) Growth

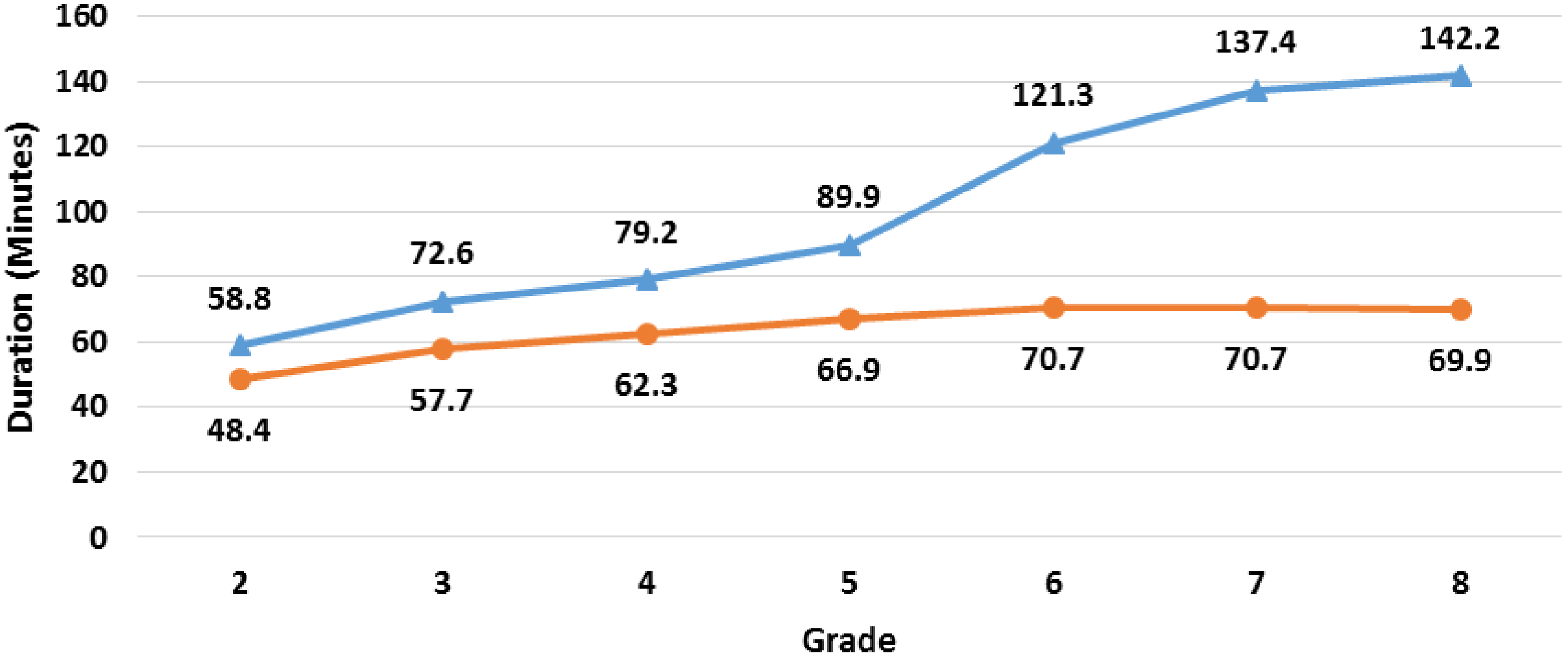
AVERAGE CPS TEST DURATIONS VS. NATIONAL AVERAGE





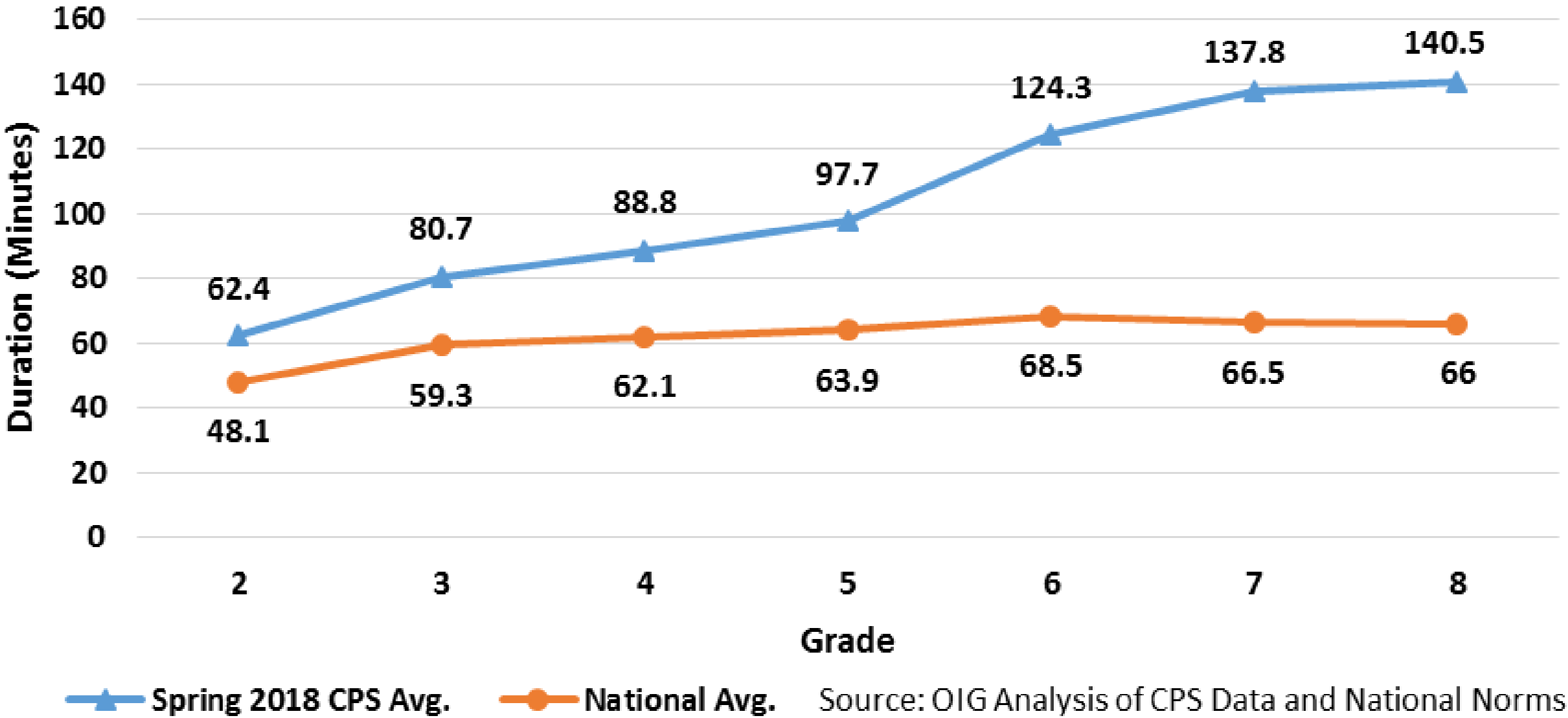
- In Spring 2018, CPS's average test durations in both Math and Reading were longer than the national average in every grade tested.
- CPS Durations in grades 6, 7 and 8 were especially far from the norm.
- CPS's high-duration averages were not driven by Diverse Learners, who were *less likely* than non-Diverse Learners to take long tests.

CPS Math Durations vs. National Average, by Grade



Source: OIG Analysis of CPS Data and National Norms

CPS Reading Durations vs. National Average, by Grade



Source: OIG Analysis of CPS Data and National Norms

CPS TESTS THAT TOOK MULTIPLE TIMES NATIONAL NORMS

Table: CPS Test Durations vs. National Norms		
2018 Test Duration vs. National Norm	Tests	% of Tests
All CPS 3rd – 8th grade tests	320,561	100%
At least 2 times the national norm	82,824	25.8%
At least 3 times the national norm	24,269	7.6%
At least 4 times the national norm	7,448	2.3%
At least 5 times the national norm	2,388	0.7%

- 1 out of 4 CPS tests took twice as long as national norms; thousands took 3, 4 or 5 times.
- Tests that took 3 or more times national norms were clustered in certain schools. About 20% of such tests were in 2.8% of schools.

NWEA RE: TEST DURATIONS

- "Tests should be administered in a reasonable time as guided by the test duration norms" first published in August 2018.
- "The validity of any test is based on certain presumptions about the conditions under which the test was administered [including] . . . . that tests are administered in Chicago in conditions that do not differ substantially from testing conditions elsewhere.''
- "When test durations exceed normal test durations by large amounts of time, CPS's ability to make accurate inferences may be compromised."
- NOTE: NWEA could not tell the OIG at what point a duration became problematic as it had not researched this issue.

DURATION PROBLEM WORSENING OVER TIME

Spring 2016 to Spring 2018

Table: Increases in Avg. Spring CPS Durations from 2016 to 2018		
CPS Grade	**Math Increase**	**Reading Increase**
3	12.4%	17.2%
4	12.5%	21.7%
5	10.7%	22.6%
6	17.4%	18.5%
7	22.2%	24.0%
8	18.0%	18.5%
Source: OIG Analysis of CPS Data.		

- Each spring, 2016 to 2018, CPS duration averages were above national norms.
- Over those three SYs, durations jumped by double-digit percentages in each test.
- By subject, the biggest duration hikes were in 7th grade, where results impact SEHS admissions.
- A recent OIG analysis showed that every test duration jumped yet again in 2019.
- **If no action is taken, durations may well continue to increase, putting CPS results at increasing risk.**

UPDATED NWEA GUIDANCE: TEST DURATIONS

Since the OIG's report, in December 2019, NWEA published updated guidance on maintaining test integrity. NWEA addressed many issues raised by the OIG's review, including excessive test durations and pauses intended to skip difficult questions. NWEA's recommendations included that district procedures state:

1) Average test durations of classrooms and grade levels "should not substantially differ" from its norms.

2) Durations should remain relatively consistent across terms.

3) Test durations should be monitored periodically to ensure consistency across schools, classrooms, and terms.

UPDATED NWEA GUIDANCE: EXCESSIVE DURATION

NWEA's updated guidance did not set a clear standard for excessive durations, but did offer examples of problematic durations. That included "Mr. Dietzen's" fifth-grade class, which averaged 150 minutes on a Spring test. Said NWEA:

- "This average test duration is beyond the 99th percentile of all NWEA tests for this grade. Mr. Dietzen's tests are unreasonably long, given that there is no reason for students to need to average 2.5 hours to complete a MAP Growth assessment.
- "Durations this long invalidate comparisons between his students' test results and NWEA norms, because the conditions vary so much from the typical test durations for NWEA students.
- "He should be coached to encourage reasonable testing durations by making sure his students are aware of the need to try their best, and that there is minimal benefit to the student to take that long to complete their assessments."

DURATION PROTOCOL

Current CPS rules state that all students, including general education students:

- can take as long as they want;
- are allowed “frequent breaks” and
- can take the test over “multiple days.”

About 30% of CPS tests take multiple days to complete; a few students and teachers described tests that took a week or more.

WHAT EXPERTS SAID ABOUT UNTIMED, HIGH-STAKES TESTS

High-stakes tests should be administered in one sitting, or in self-contained units, according to the chief of investigations at a test security firm.

- "There are too many things that can happen during those breaks that can affect the validity of the test results."

A professor of educational measurement/testing expert told the OIG that he "would definitely set a time limit" for such a high-stakes test. He noted that on 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."

DATA FINDINGS

1) Durations

2) Pauses

3) Growth

CPS TESTS BY TIMES PAUSED OR TIMED OUT

	Total	0	1-4	5-9	10-14	15-19	20+
Tests	302,993	145,424	145,388	10,524	1,149	290	218
Students*	152,128	91,221	94,309	8,904	1,053	268	180
Schools*	463	459	462	401	165	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.

- More than 12,000 tests (about 4%) were paused 5 or more times in Spring 2018.
- Tests with 5 or more pauses were clustered in certain schools — Over 20% were in just 1.7% of schools

% OF 2018 TESTS WITH 5 OR MORE PAUSES, BY GRADE

Grade	% of CPS Tests w/ 5+ Pauses
3	1.7%
4	1.8%
5	2.3%
6	4.6%
7	7.0%
8	7.1%
All	4.0%

- The highest pause rates were in 7th and 8th grade, which both can carry high stakes for students.
- The tests with the highest pause counts were more likely to be taken by *non*-Diverse Learners than Diverse Learners.

PAUSE RULES NEEDED

- The pause function, which replaces the current question with a new question of similar difficulty, is not adequately addressed by current CPS rules.
- Proctors currently could be pausing tests to help students skip questions they can't answer, as warned in an April 2018 CPS Audit. Such actions were reported in a small number of OIG interviews.
- NWEA told the OIG: "The validity of the assessment can be compromised if tests are paused for the purpose of producing a new question."

EXAMPLE OF HIGH-PAUSE 8TH-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
Duration*: 4:25		8:46	29	53
*NWEA's duration excludes time spent on paused questions and time while the test is paused.				

This is an example of one of the 200+ tests (out of more than 300,000) with at least 20 pauses. On the last day of this student's 3-day test:

- Her test was paused 14 times in under 2 hours. All pauses were too brief to be time outs.
- There were 4 consecutive pauses in just 9 minutes.
- The OIG sees no reason for the proctor to pause and resume the test so frequently in such a short time period. Note: this level of detail regarding pauses is not readily available to CPS.

TIME OUT REFORMS NEEDED

- NWEA told the OIG: “In general, there is no reason for a student or proctor to allow a question to time out . . . . [Doing so] should be considered a possible ‘gaming’ practice.”
- Also: “The assessment’s validity may be compromised if a student intentionally allows a question to time out because he/she does not know the answer and would like a new question.”
- OIG interviews with a small sample of students indicated some students are intentionally timing out questions so they can get new questions. This may be occurring more often in the high-stakes grades of 7th and 8th, where pause rates were the highest.

DATA FINDINGS

1) Durations

2) Pauses

3) Growth

SOME LEGITIMATE REASONS FOR HIGH GROWTH

- Previous test score was uncharacteristically low.
- Student made large strides in language proficiency and was therefore better able to show his/her true ability.
- Student was diagnosed with a learning disability and given needed accommodations for the first time.
- Student clicked with a particularly effective teacher, a new curriculum or a new learning strategy.

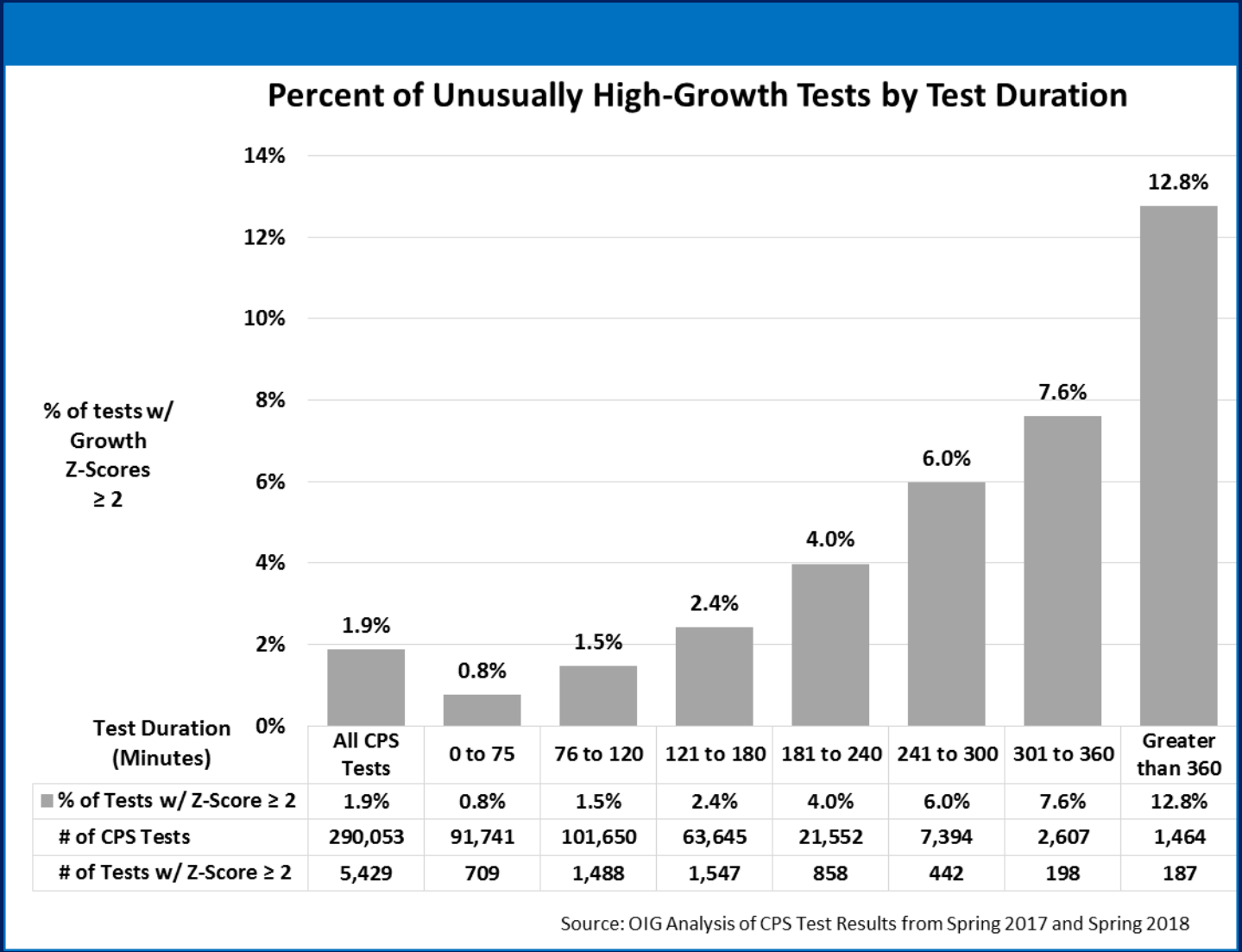
HOW THE OIG ANALYZED STUDENT GROWTH

The OIG analyzed student growth by comparing each test's score gain from Spring 2017 to Spring 2018 to the gains of other CPS tests:

- Taken by students in the same grade level and subject, and
- Taken by students with the same Spring 2017 score.

The OIG defined "unusually high growth" as two or more standard deviations above the average growth of comparable tests (in other words, having a "z-score" of at least 2). In grades 3-8, 1.9% of CPS tests met this standard.

DURATION AND UNUSUALLY HIGH GAINS

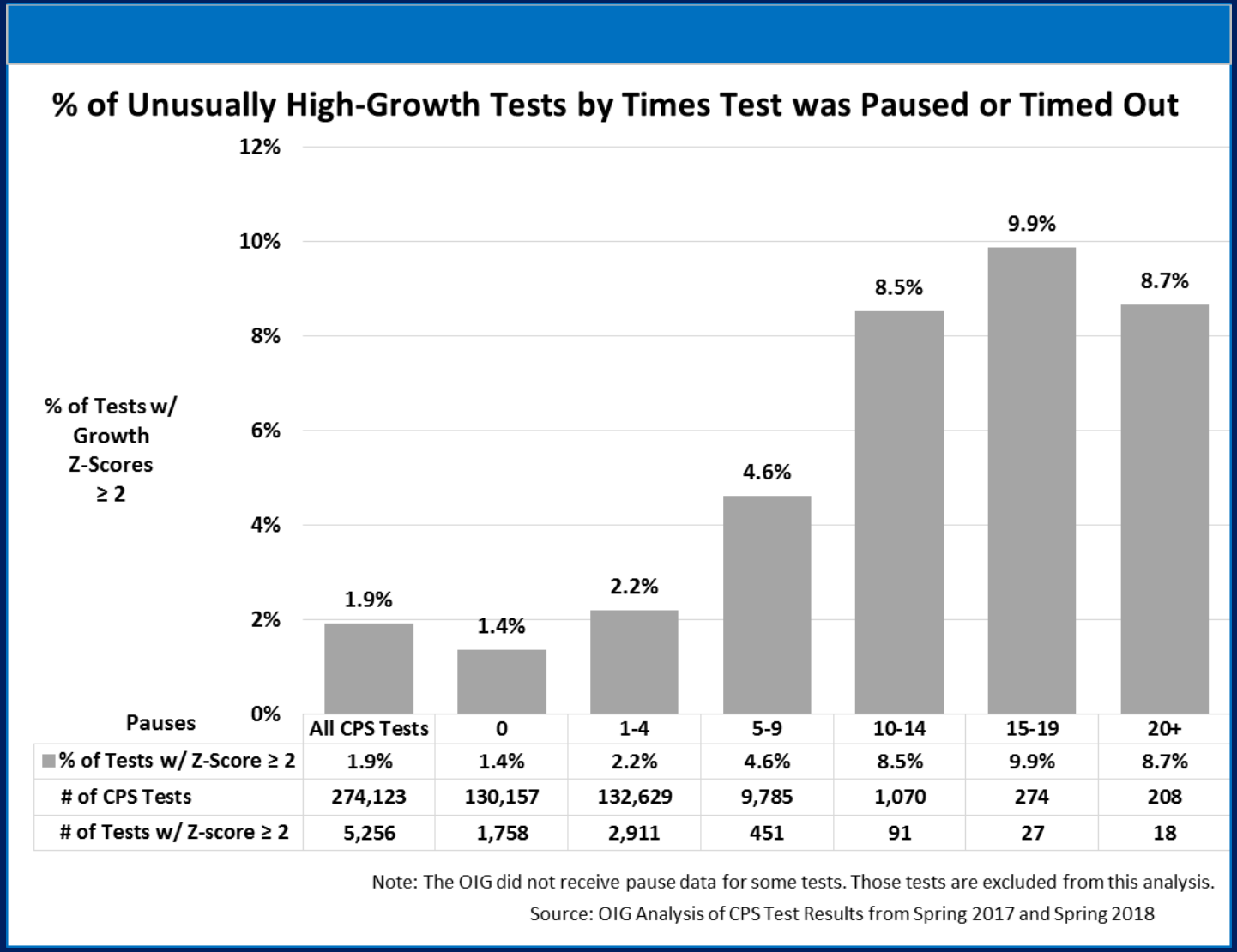


Test Duration (Minutes)	All CPS Tests	0 to 75	76 to 120	121 to 180	181 to 240	241 to 300	301 to 360	Greater than 360
% of Tests w/ Z-Score ≥ 2	1.9%	0.8%	1.5%	2.4%	4.0%	6.0%	7.6%	12.8%
# of CPS Tests	290,053	91,741	101,650	63,645	21,552	7,394	2,607	1,464
# of Tests w/ Z-Score ≥ 2	5,429	709	1,488	1,547	858	442	198	187

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

- As the length of a student's test increased, so did the occurrence of unusually high growth. The data does not establish a cause for this.
- This pattern was true for both Diverse Learners and non-Diverse Learners.

PAUSES AND UNUSUALLY HIGH GAINS



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 Test Results from Spring 2017 and Spring 2018

- There was a similar pattern with pauses and growth but, again, the data does not establish a cause.
- This pattern was true for both Diverse Learners and non-Diverse Learners.
- One expert noted that, other than for Diverse Learners who require accommodations, there's "no educational explanation for why pauses would improve scores."

INSUFFICIENT DATA

NWEA currently does not provide CPS key data that would be helpful in detecting test irregularities, such as:

- the proctor of each test;
- who paused a test and whether it was paused or timed-out.
- The number of pauses on each test (though NWEA can provide this on request). The OIG recommends that CPS analyze this data annually.
- How many days each test took, the length of those test sessions, and which dates the test was worked on. This can be requested as a custom report.

TEACHERS AS PROCTORS

- CPS Reading and Math teachers whose evaluations are tied in part to their students' test results currently can be the sole proctors of their students.
- NWEA says: In high-stakes situations, students should be proctored by their teachers as well as a second proctor who has "no direct investment" in the students' performance, according to January 2017 NWEA guidance.
- NWEA also said: A second proctor protects the integrity of testing results and teachers from false accusations of cheating.

PRIOR CPS AUDIT AND REFORMS

An April 2018 CPS Audit of testing protocols found that CPS controls for *detecting* as well as *preventing* irregularities needed to be strengthened. In response, CPS instituted several reforms, such as

- Boosting preventive training by requiring all proctors and test coordinators to undergo training and pass a 5-question "exit slip."
- Having three departments (Assessment, SQMR, and Audit) collaborate on a new method of flagging unusual test results, and
- Auditing a much larger number of test sites in 2018 and 2019 than in 2017.

MORE REFORMS NEEDED

Despite some reforms, the OIG found that some key Audit suggestions were never executed or were insufficiently executed. For example:

Not Implemented	Insufficiently Implemented
• Analyze pauses as part of the process of flagging schools with unusual results. • Try to avoid having Reading/Math teachers proctor their own students. • Include a warning about the penalties for cheating in the Test Security Agreement all proctors must sign.	• Some schools or classrooms with multiple flags were not audited. • Training and exit slip did not adequately cover the test's unique features, improper test procedures, or how to prevent test irregularities. Note: One test security guidebook recommends "clear examples of what behavior is unacceptable," but these were not provided in the training.

CONCLUSION

- The OIG conducted a Performance Review of the administration of the NWEA test. This was not an investigation of test cheating.
- The review did not establish malfeasance. Instead it recommended systemwide reforms to proactively prevent future problems.
- The OIG found a concerning pattern of unusually long durations and high pause counts in a minority of tests that could reflect some degree of gaming and/or cheating. Interviews with a small sample of students in schools with unusual results uncovered some reports of test administration irregularities, gaming and even some cheating.
- To be clear, unusual results also could be benign. However, even if benign, particularly long durations may compromise the validity of comparisons to NWEA norms. New guidance from NWEA makes this especially clear.

RECOMMENDATIONS

1) **Reduce durations**, preferably by setting a time limit.

2) **Take concrete steps to limit pauses**, including by providing clear instructions on the right and wrong way to use pauses.

3) **Find an auditable way to record the proctor** of each test taken so that analysts can look for trends among proctors. Documenting proctors also should deter cheating as it will establish clearer accountability.

4) **Focus on auditing proctors**, rather than grades and subjects, by using new proctor data. Cheating and gaming tactics described to the OIG were done with the knowledge, or at the instigation, of proctors, who ranged from teachers to SECAs to coaches.

5) **Prohibit teachers from being the sole proctors of their students** if their REACH evaluations are tied to test results. Add a second proctor who would be responsible for the test's integrity. Note: NWEA recommends a second proctor in high-stakes tests.

RECOMMENDATIONS (CONT.)

6) **Bolster training and the exit slip** that must be passed for someone to proctor. Include guidance on improper pauses and long durations. During training, cite the OIG as an office to be contacted with concerns about test irregularities.

7) **Add penalties for test cheating** to the Test Security Agreement that all proctors must sign.

8) **Hire a test security expert** for help and guidance in addressing the OIG's recommendations and other OIG concerns. If NWEA cannot provide recommended security features, this expert should help CPS write an RFP for a new vendor.