Guidance for Administering MAP Growth Assessments to Improve Measurement Accuracy and Maintain Test Integrity

December 2019

NWEA
121 NW Everett Street
Portland, OR 97209
866-654-3246
https://www.nwea.org

Table of Contents

Section 1: Introduction

Schools use results from the MAP® Growth™ interim assessments from NWEA® in several ways. Some use the results solely for instructional purposes and school improvement. Some include results as a component of their teacher evaluation systems, to determine whether a student advances to the next grade, or as an indicator for student readiness for certain programs or interventions (such as special education or gifted and talented programs). Whether the purpose is "low-stakes" or "high-stakes," students may be impacted by decisions made from their assessment data. Thus, educators have a professional l obligation to ensure that student results on assessments are an accurate reflection of student academic achievement.

The primary purpose of MAP Growth is to accurately measure student achievement, and, when done over time, student academic growth. The validity and defensibility of decisions made from this assessment are a direct product of the integrity of the measure. While most educators strive to administer MAP Growth assessments in a manner that provides an accurate measure of their students' achievement, if the accuracy of the assessment is compromised, then decisions made from MAP Growth assessment results may be invalid. More importantly, students can be negatively impacted. For example, in Atlanta, students who thought they were performing at or above grade level discovered, as a widespread cheating scandal was revealed, that their achievement had been misrepresented and they were actually in need of academic support and assistance. The scandal eroded the affected students' and families' trust in the system and subsequent efforts to mitigate the damage on these students has been largely ineffectual (Aviv, 2014; Cottman v. The State, 2017; McCray, 2019).

In recent years, school system leaders have been under increasing pressure to improve test scores. In some settings, that pressure passes to educators who may respond by engaging in testing practices that compromise the integrity of the assessment. When those practices occur, it can compromise the primary purpose of the MAP Growth assessments, which is to accurately measure students' academic knowledge and their learning so that sound educational decisions can be made from those results. In circumstances where irregular test administration practices are used to artificially inflate achievement or growth metrics, those practices may violate the norms of the profession and the *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 2014).[1,2] In other cases, educators may unintentionally administer assessments in ways that compromise the integrity of the results, and that may lead to inaccurate instructional

[1] Standard 6.6 requires that "Reasonable efforts should be made to ensure the integrity of test scores by eliminating opportunities for test takers to attain scores by fraudulent or deceptive means. (p. 118). The standard puts the responsibility for maintaining conditions that would protect the integrity of test scores with the educator(s) administering the test. Standard 6.7 goes on to charge test administrators with the responsibility for assuring the security of test materials at all times.

[2] Throughout this guidance "the Standards" refers to the *Standards for Educational and Psychological Testing*.

decisions to the detriment of student learning. This guidance document[3] is intended to help educators, school leaders, and district administrators to improve the accuracy of their test results and protect the integrity of the testing process.[4] Consistent testing practices and accurate results lead to better educational decisions, which ultimately benefits students.

Section 2: The Importance of Written Policies

When we refer to a test[5] as having "integrity," we mean that the test was administered in a manner that is likely to produce an accurate result. A test event with diminished integrity may provide inaccurate or even misleading information that may lead to unproductive educational decisions and, for families, a flawed understanding of their child's actual achievement.

The core principle underlying test integrity is that test administration practices should be aligned with producing an accurate result. The question, "will this practice help ensure the accuracy of the student's result?" should guide test administration decisions.

The guidelines we offer in this document are grounded in the *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 2014). In the preamble to Chapter 6, the authors reference the importance of standardization in testing practices, so that "all test takers have the same opportunity to demonstrate their competencies (p. 111)." Written policies and procedures for test administration are important because they help ensure consistency of test administration practices across settings and clarify the expectations and rules for administering assessments for educators. Standards 6.6 and 6.7 place obligations on educators to maintain test administration conditions that minimize the likelihood of fraudulent results,[6][7] and to maintain the security of test content. For the purpose of this guidance document we generally rely on these three references when offering guidance as to whether a test administration practice is appropriate.

It is important for school systems to establish board policies that define expectations for professional conduct in relation to testing, to embed those expectations in testing procedures, and to provide documentation and appropriately detailed guidance with regard to the proper

[3] NWEA conducts ongoing research in this area, and the guidance throughout this document reflects guidance based on this research and questions we have received.

[4] This paper identifies many potential risks to the integrity of the testing process, but it is not an exhaustive list of the ways in which the accuracy of tests may be compromised.

[5] For purposes of this guidance document, the term "test" generally refers to a test event; that is the taking of a test and the resultant score, rather than the design or construction of the test itself.

[6] While the reference in the Standards is to fraudulent results produced by students, an educator who facilitates practices that produce fraudulent results would also violate the standards.

[7] We do not make specific references to any practice as "cheating" or "test fraud" in this guidance. Characterizing a practice as "cheating" or "fraudulent" can imply violation of a legal statute and includes the requirement to prove contractual or legal elements that must be resolved by the school system or legal authorities based on the facts of a particular incident.

test administration practices. A policy for professional conduct is important because it establishes that principle that test administration practices are integrally related to professional conduct. Detailing appropriate administration practices is essential to provide guidelines for how tests are to be administered and offers clarification in areas where professional judgment might be required. We provide some excerpts from sample policies and procedures in Appendix A and include some key areas for consideration in the following sections. This document provides guidance around what NWEA considers to be appropriate testing practice.

Section 3: Suspension, Early Ending of Test Events, and Retesting

In the interest of getting an accurate test result, it is sometimes appropriate to suspend or end a student's test session early, or to suggest retesting a student after a testing session is complete. In the Standards, Section 6.1 says that "retesting is intended to decrease the probability that a person will be incorrectly classified as not meeting some standard" (p. 114-5). In other words, the primary purpose of retesting is to protect the student's interest and ensure that decisions made from the measure are accurate.

Thus, in cases in which a student is sick, rushing through a test, or showing a lack of engagement, it may be appropriate to pause, suspend, or terminate a test with the intention of retesting the student at another time. In these situations, the student would continue the test at a later point, or potentially restart the test altogether, when he/she/they is better able to engage with the test. There are also situations when retesting may be warranted after a test has been completed.

MAP Growth tests can be **paused**, **suspended,** or **terminated** prior to their completion. Once a test is completed and a score is awarded, a **retest** can be taken if necessary.

- **Paused or Suspended Test.** A test is paused or suspended when the proctor interrupts a student's test with the intention of allowing its completion after a break (e.g., lunch or to use the restroom) or on a later date. The portion of the test completed prior to the pause or suspension is included as part of determining the final score. Pausing or suspending a test is appropriate when it is clear the student is not engaged on the test for a few items and the proctor believes the student will be able to re-engage with the test once it is resumed. However, if a student has been disengaged throughout the majority of a test, pausing or suspending the test is not appropriate and is likely to produce an inaccurate score.
- **Terminated test.** A test is terminated when the proctor or teacher ends the test without allowing it to be resumed. In this case, no score is awarded, and the student would need to start a new test from the beginning. Terminating a test is appropriate if the student is disengaged and a large portion of the test has already been completed since the score is unlikely to be accurate measure of student achievement in this circumstance.
- **Retesting a student with a completed test.** Once a test is completed, a score is awarded and becomes part of the student's longitudinal record. If retesting is justified (see

Section 3.1), the student can complete a retest. The "official" test result for a term, however, is based on the score with the lower standard error of measure (i.e. the more reliable score), and not the most recent or highest score. That is one reason why pausing, suspending or terminating a test when a student is disengaged can be preferable to allowing a test to be completed.

In general, tests should be paused or suspended prior to completion rather than allowing a student to complete a test that is likely to produce an inaccurate result that requires retesting. NWEA's proctor console provides information and notifications to proctors during testing so they can determine if intervention with a student is needed. By monitoring these notifications and intervening, engaged testing is more likely to occur, mitigating the need to terminate the test or retest a student.

Because teachers, principals, and students may be under significant pressure for their students to perform well on tests, particularly if they are used for high-stakes purposes, there may be situations where educators may seek to retest students in which the educator might benefit from a higher score. Here are three examples of situations that educators might encounter. Two of the examples NWEA does not consider appropriate testing practices while the third is considered appropriate:

- Principal Barnes' school adopts a practice in which students whose spring test scores fall short of their growth target are automatically retested.
- Principal Harrington discourages retesting students during the fall term, even when evidence justifies retesting (for example a student fell sick during the test), but retests many students during the spring term.
- Student Heather's test score fell 1 point short of the cut score for gifted programming. Because her test score is used for teacher evaluation, her teacher Green would benefit if she receives a higher score on a retest. Teacher Green writes a memo to the district testing coordinator requesting a retest for Heather because she is near the cut line for an important educational decision. The testing coordinator approves the decision.

NWEA encourages school systems to adopt policies that ensure retesting is done to improve the accuracy of results and benefit the student, and that decisions about exceptions to those guidelines are made by an impartial third party.

3.1. Retesting Guidelines

School system leadership should establish an expectation that educators emphasize the importance of every assessment and encourage all students to do their best every time. To ensure accurate student data, prior to the first round of testing, schools or districts should have written procedures that establish guidelines governing when a student should be retested.

The broad principle is that retesting a student who has a completed test is justified when there is evidence that a completed test is an inaccurate measure of the student's achievement. In

general, it should be assumed that if the proctor or educator allowed a test to be completed, then no evidence was visible during the test that would have compromised its accuracy. Thus, NWEA recommends retesting should only occur when objective evidence is produced that would indicate the validity of the test is at risk. For example, when monitoring assessments, educators and proctors might intervene when they see evidence that:

- A student may be ill or distraught during the test.
- A student refuses to take or complete the test.
- A student is rushing to complete the test items.
- A student is observed responding without reading the items.

In these, and other similar circumstances, educators should first intervene with the student to identify the reason he/she/they is not engaged with the assessment and, assuming the student is able to give his/her/their best effort, then encourage the student to try his/her/their best. If the student is still not trying on the assessment after the educator's efforts at encouragement, the test may be either suspended or terminated, based on the proportion of the test completed. The student should resume or restart the test at a time when he/she/they is better able to demonstrate his/her/their learning.

MAP Growth reports provide multiple metrics that can be used to inform decisions about retesting. For example,

1. Retesting can be justified if the student shows rapid guessing on a large portion of the test, as reflected in the Percentage of Rapid Guesses (%RG) metric on a MAP Growth Assessment (see Section 3.2).
2. Retesting can be justified if the student shows a "substantial" decline in score, as defined by the school or district, between the current and prior testing (see Section 3.3).
3. Retesting can be justified if the student's current test duration is unusually short (see section 3.4).
4. Retesting can be justified post-hoc if metrics from the test provide evidence that the score is likely to be inaccurate. These may include a low percentage of items answered correctly, an abnormally high standard error of measure, rapid guessing that had a significant impact on the students score, or other problems.

If possible, a school or district should consider implementing a system in which a principal, building administrator, or, ideally, an impartial designee reviews all retesting decisions prior to the student retaking the test. Retesting should be monitored to ensure that retesting policies are applied consistently at every testing term.

3.2. Percentage of Rapid Guesses (%RG) on a MAP Growth Assessment

The percentage of Rapid Guesses (%RG) is a MAP Growth metric that provides a nuanced view of student engagement during the testing process. A student response is flagged as a rapid

guess when the student answers an item in less than 10% of the average time it takes other students to respond to the item. Starting in Fall 2019, on the majority of tests, the MAP Growth test automatically pauses when a student rapidly guesses on six items (those guesses do not need to be consecutive), and then again after an additional number of rapidly guessed items. The number of additional pauses, and when those pauses occur, varies slightly depending on the test/subject. After automatically pausing, test proctor action is required for a student to resume testing. This auto-pause feature allows the proctor or educator to talk with the student, understand the issues, and encourage engaged test behavior, or decide to suspend the test and resume it at a later point. Research shows that notification and the action of proctors can decrease rapid guessing and improve the validity of test scores (Wise, Kuhfeld, & Soland, 2019). We encourage proctors to suspend or terminate tests at these pause points if they believe the student is unlikely to engage with the remainder of the assessment.

The student test engagement capability in the MAP Growth assessments was established to help ensure each student's assessment results reflect their content knowledge and ability. A lack of engagement typically deflates a student's score. If a RIT score is generated for a test with 30% or more of items answered lacking sufficient engagement, the score will not accurately represent a student's achievement level (Wise, S., 2019). In general, a student is considered engaged if the %RG is less than 10% or listed as N/A on student reports.

Students who rapidly guessed on 30% or more of the total test items are permitted to finish the test and will receive a RIT score. However, NWEA recommends that these students be retested if district policy allows, as this is the point at which the validity of a student's score has been substantively affected. If retesting is available, the student should test again when he or she is ready to engage in the testing process. By default, the MAP Growth system will allow retesting of students who reach the 30% rapid guessing threshold. However, if your district's policies do not permit retesting of such students, your MAP Growth System Administrator or District Assessment Coordinator will need to contact NWEA to request that the default setting be turned off.

When the %RG is between 10% and 30%, educators should consider the estimated impact of the student's lack of effort on his or her RIT score, which is also reported in MAP Growth. If you judge the Estimated Impact of Disengagement on RIT for a student to be a serious and deflating impact on a student score, consider retesting the student in compliance with your district's written policy.

3.3. "Substantial" Decline in RIT Score Between Two Test Events

A large decline in test scores between two administrations can be an indicator of test results that do not reflect a student's actual achievement level. There are circumstances in which schools may consider retesting individual students if they show a substantial drop in test score in relation to the prior term.

NWEA does not formally define what would be considered a "substantial" decline in RIT score between consecutive test events. For example, in lower grades, a decline of greater than 10 RIT

points from the most recent prior test may be indicative of low student effort on the current test, or some other factor that caused the student to score lower than expected. Whatever definition the school system chooses to use should be included in the district's written policy on retesting and should be applied at every term. For example, if the threshold for retesting a student was set as a RIT score dropping by 10 points, this should be enforced from the prior spring to the fall (i.e., summer loss) just as a student whose RIT score dropped by 10 points from fall to winter in the same school year would be retested.

3.4. Written Rationale for Retests

When deciding to retest a student, the rationale should be documented at the time it occurs. The district's testing procedures should include information on who approves retesting decisions, how these decisions are to be documented, and who is responsible for creating the documentation. This provides school leaders with the ability to track which students were retested and for what reasons.

Documenting instances of retesting can be useful for two reasons. First, it protects teachers from accusations of test manipulation if a student's test performance is questioned. Second, it ensures transparency and accountability surrounding all retesting decisions

Section 4: Monitoring Test Durations and Maintaining Consistency in Testing Conditions

MAP Growth provides a metric on student engagement called Total Test Duration, which indicates the amount of time the student took to complete the assessment. Test duration can be an indicator of whether a student gave appropriate effort during the testing process. NWEA's current research indicates MAP Growth tests completed in less than 15 to 20 minutes will likely provide inaccurate estimates of student achievement, although this may not be the case for every student who completes a test quickly.

MAP Growth is not a timed assessment. It was designed so that students were not held under unreasonable time constraints so that students could take the time needed to answer each question. Furthermore, educators should recognize that the test can take longer for high achieving students and students in the upper grades, in particular, to complete. Students with Individualized Education Plans (IEPs) or 504 Plans that indicate they need longer time to complete assessments should also be given all the time necessary to get a valid estimate of their achievement.

With that noted, the pressure on teachers to improve scores may have been one of the reasons average MAP Growth test times have increase significantly in the past several years. Between spring 2014 and spring 2018, the average test duration increased by over 21 minutes in reading and 17 minutes in mathematics. In some schools, it has become common for students to average over 2.5 to 3 hours to complete a single MAP Growth assessment. We are increasingly concerned that assessments may be taking longer to complete than is necessary to produce an

accurate score and, in some rare circumstances, that test conditions and durations in the spring are managed to be longer in an effort to inflate measured student growth (in districts that measure growth from fall to spring). Managing test conditions to artificially improve measured student growth is an unsupported use of MAP Growth.[8]

As an example, modifying test conditions to minimize students' fall scores or maximize their spring scores, which can inflate the magnitude of gains a student shows from fall to spring, is not a supported practice (the same is true across any two term pairs). For example:

- Students complete fall tests in under 45 minutes on average, while spring tests average over 2 hours.
- Students are expected to complete fall tests within one class period, but spring tests are administered over multiple class periods.
- Students who complete tests in under 20 minutes during the fall are not retested, while students who test in under 20 minutes during the spring are retested because their tests are too short.

To put student test durations into context, NWEA published a document titled "Average MAP Growth Test Durations." This document includes tables that show the average MAP Growth testing durations by content area and grade based on aggregated test durations from student tests during the 2017–2018 school year. The tables are intended to provide educators with general ranges that show how long students typically take to complete a MAP Growth assessment, as well as the typical differences in test duration across terms. Typical test durations vary based on the grade and season with elementary students taking less time than secondary students, higher achieving students taking longer than average achieving students, and tests taken later in the school year taking a few minutes longer than tests taken earlier in the school year.

In general, two expectations should be included in a district's test administration procedures. First, the procedures should state that average student test durations should not substantially differ for classrooms and grade levels of students from the published test duration ranges. Second, the procedures should state that test durations should remain relatively consistent across terms. Here are four examples:

- Mr. Smith's 5th grade math students take 75 minutes, on average to complete their spring MAP Growth assessment. The average test duration for NWEA students in 5th

[8] Referencing Section 6 of the Standards "the importance of adherence to appropriate standardization of administration procedures increases with the stakes of the test (p. 111)." By extension, adherence to appropriate standardization procedures should be expected across terms when both scores are used for a high-stakes decision.

grade is 67 minutes. Mr. Smith's duration averages at about the 60th percentile of all tests.[9] This is reasonable.

- Ms. Cossette's 5th grade math students take 22 minutes, on average, to complete their fall MAP Growth assessment. This is 34 minutes below the average time to complete fall assessments in this grade. The average duration of the class is at the 7th percentile of all tests. In this case, the assessments are too short which could result in the achievement of her students being underestimated. She should be coached to make sure her students give their best effort in both fall and spring assessments.
- Mr. Dietzen's 5th grade students take 150 minutes, on average, to complete their spring assessment in each subject. 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.
- Mr. Berend's 5th grade took 40 minutes on average to complete their fall assessment and 150 minutes on average to complete their spring assessment. A difference this large demonstrates an inconsistency in testing conditions. He should be coached to ensure testing conditions are consistent and to encourage students to provide their best effort in in both fall and spring assessments.

Test durations should also be periodically monitored by the school system to ensure that test administration practices are consistent across schools and classrooms, and that test durations remain relatively consistent across terms. If classrooms of students take notably longer than the range of regular times, steps should be taken to work with educators to reduce test durations to more reasonable lengths.

A school district's written test administration procedures should include a statement about when students should be retested based on the total amount of time they spend on their test, the difference in duration between terms, and that those expectations be enforced every term.

NWEA researchers collected testing policies from partners and used them to create examples that are included as Appendix A to this document for your consideration and use as you develop your own written policies.

Section 5: Proctoring

[9] The percentiles referenced in these examples come from the *Average MAP Growth Test Durations* document. To estimate the percentile associated with a test duration use the following link to access an on-line tool that does the calculations for you.
https://public.tableau.com/profile/jfcnwea?/vizhome/NWEAcollegescorecard/Dashboard1#!/vizhome/MAPGrowthDurations/Percentiles

The primary responsibility for good testing conditions lies with the proctor and the teacher. Part of that responsibility includes motivating students to do their best, establishing testing conditions conducive to good performance, and actively monitoring testing to prevent problems. When the stakes around test results are high, and especially when test results are used as part of teacher or school evaluation, it is important for everyone's sake for testing procedures, policies, and practices to be as transparent as possible.

NWEA encourages districts using MAP Growth to participate regularly in proctor training to ensure that proctoring practices maintain the integrity of the testing process. Proctoring best practices should include the following steps:

1. A teacher should generally be present during testing because he or she is the most aware of the learning needs of his or her students and can help keep students focused on the testing process.

2. When results from the MAP Growth assessment are used for a high-stakes purpose, it is good practice to also have a second proctor (someone in addition to the teacher) in the room to help oversee the testing process. The second proctor should be someone who does not have direct investment in the performance of the students being tested. In many schools, the testing coordinator could serve as the second proctor. The second proctor shares responsibility for protecting the integrity of testing results and protects teachers from accusations of cheating.

Section 6: Unusual Testing Practices

Most of the examples[10] we offer in this document reflect actual questions or cases around test integrity that have arisen in the past few years. We would emphasize that instances of unsupported test administration practices tend to be rare and that the vast majority of educators are doing their best to administer the MAP Growth assessments appropriately.

6.1. Pausing tests for purposes of generating a new question

The purpose of the pause function on the MAP Growth test is to permit the test to be temporarily suspended if a student needs a break, or if time does not permit the test to be completed until a later time. To maintain test security, a new item is presented to a student after a test event is paused. Given this, it is possible that a test could be paused so that a student who may not know the answer to an item is presented with a new item that they might be more likely to answer correctly. This practice is unsupported and could affect the validity of the assessment. As an adaptive test, MAP Growth was designed to adjust the difficulty of each item based on student answers to previous items. It was designed with the expectation that students would be given challenging items that they were likely to get wrong about half of the

[10] Since the last publication of this document in April 2019, some new issues have arisen which we address in this section.

time, and with the expectation that students would attempt those items. If the pause function is used to skip items that are perceived to be too difficult, then item responses that are essential to the accuracy of the measure are not considered in assessing the student's achievement. This could make the resultant score invalid.

Control of the pause feature is the responsibility of the teacher and proctor. Students should not be allowed to pause the test on their own. Here are some examples of supported and unsupported use of the pause feature:

- Roberta asks the proctor to pause the test so she can use the restroom. The teacher pauses the test and resumes on her return. This is a supported use of pausing.
- Ms. Green believes that students are seeing some items on the MAP Growth tests that have not been taught. She tells her students that they may pause the test up to 4 times if they believe an item is too hard. This practice would be in violation of testing procedure and is not supported.

6.2. Recording of items

MAP Growth is a secure assessment. NWEA's nationally representative student achievement and growth norms (Thum and Hauser, 2015) are both premised on the condition that the test is taken on a "level playing field" – that is, educators have not received access to test items, and students have not been intentionally exposed to items prior to the test. If a class of students is advantaged because teachers or students have recorded or transcribed questions and pre-prepared students with the answers, that is not a supported conduct.

Here are two examples that help differentiate between supported and unsupported activity.

- In observing math tests, Ms. Casey sees that her students are struggling on questions that focus on measuring circles. She realizes that she didn't spend much classroom time on this topic and decides she needs to develop some new lessons on this subject. She also devises some of her own practice problems to assess whether students understand the concept. This is a supported activity.
- In administering math tests, Mr. Hanson tells students to write down questions they do not know so that they can cover those questions after the test. This is not a supported activity.

6.3. Providing assistance to students

MAP Growth is a measure of student achievement and the results are intended to reflect the student's work.[11] If an educator offers students any form of assistance with answers on the assessment, it compromises the integrity of the test result because the result is no longer solely reflective of the student's achievement level. Here we distinguish between offering assistance

[11] Standard 6.6 of the *Standards for Educational and Psychological Testing* states: "In any evaluation of work products (e.g., portfolios) steps should be taken to ensure the product reflects the test taker's own work…(p. 117)"

with answers, which is always unsupported by NWEA, and accommodations which are approved and intended to remove barriers that prevent students from demonstrating their learning (see Accessibility and accommodations features and FAQ). Here are examples that help differentiate between supported and unsupported activity.

- Ms. Johnson sees that Robert is rushing through questions on the reading test. She stops Robert and encourages him to take his time to answer each question carefully. This is a supported activity.
- Manuel calls Mr. Peterson over and tells him that he is stuck on a question. Mr. Peterson tells Manuel that it is OK to take his best guess and move on. This is a supported activity.
- Ms. Gardner wanders the classroom during MAP testing and sees that Imelda has selected the wrong answer to a question. She encourages Imelda to double check her answer. This is not a supported activity.
- Mr. Hardin tells students at the beginning of the test to raise their hands when they are unsure of an answer so he can check their work. This is not a supported activity.

6.4. Modifying test administration dates

The MAP Growth norms are adjusted to match the number of instructional weeks that the students received between assessments, so the norm against which student growth is compared reflects the actual amount of instructional time between test events.

Schools should try to schedule testing so that the actual amount of instructional time students received between tests matches the norming weeks they use for reporting purposes. For example, if a school uses the 4th instructional week in the fall and the 32nd instructional week in the spring, then the school should try its best to have students test near the identified instructional weeks. This ensures achievement norms that are used to contextualize student achievement reflect when in the year students test, and that the growth norms used to contextualize student growth reflect that 28 weeks of instruction (in this example) occurred between test events.

In some cases, educators may have administered tests at the very beginning of the fall testing period and the very end of the spring testing period to potentially maximize interpretations of growth relative to a fixed norming period. This is not a supported testing practice and can have a large impact on the validity of results. Here is an illustration of the problem.

- Mr. Burn's has taught 5th grade math to his class for the school year. His class is scheduled to take the MAP Growth test at the beginning of each testing window just as they did in the fall. It's now the scheduled time for his class to take the spring MAP Growth test. Week 4 and 32 (28 weeks of instruction) are the weeks set into the MAP Growth reporting system used to calculate the normative achievement and growth of all students in the school. His school has established five-week test windows at each term, with the middle week of that window occurring at the 4th and

32nd weeks. He believes there are a few students in his class who did not learn at the same pace as others. He is confident that if he provides these few students with four extra weeks of instruction and then tests them at the end of the spring test window (the 34th week instead of in week 30 as his class is scheduled), their measured growth might equal others in his class. One of those students started the year with a 205 RIT. He ended the year with a 216 RIT when tested at the end of spring window. Given the instructional weeks established in the reporting system (Weeks 4 and 32), his normative growth is reported as growth at the 56th percentile. When the actual amount of instruction (four extra weeks, from the 30th to the 34th week) is accounted for, the student's 11 RIT point growth is actually growth at the 48th percentile. Wanting his students to learn the same amount as the other students is a worthy goal and more powerful instruction for them is appropriate. However, given that Mr. Burn's median growth percentile is used for part of his performance review, modifying the instructional weeks used in the measurement of growth for a certain group of students is not a supported use as the extra instruction can artificially inflate their normative performance relative to others tested in his classroom and the school.

6.5. Using MAP Screening tests as a practice test or to familiarize educators or students with items

MAP Screening tests are intended to be used as an assessment to help make an initial placement of students or for a quick check of student performance during the school year. Because MAP Growth Tests and MAP Screening tests are secured assessments and draw from the same item pool, each should only be administered for its intended purposes. It is not a supported use of MAP Screening tests since MAP Growth warmup tests are available and designed for this purpose. Also, the MAP Screening tests should not be used to familiarize a class with the test content. Teaching standards aligned knowledge and skills will ensure students are familiar with the test content since each test asks questions aligned to the standards at a student's achievement level.

Here are some examples of supported and unsupported use of MAP Screening Tests:

- Ms. Barnard is working with ten students who are behind for their grade and wants to see if the interventions she is using are helping them make progress. She administers the MAP Screening Test in January as a quick progress check. This is a supported use of MAP Screening Tests.
- Mr. Anders has been assigned a middle-school social studies class. He would like to get a reading level on his students so he can make sure his course materials are properly adapted to his students. He administers the MAP Screening Test for this purpose. This is a supported use of MAP Screening Tests.
- Ms. Black administers the MAP Screening Test as a "warmup" for her students the week before she administers the spring MAP Growth assessment. She did not administer

MAP Screening Test to any students in fall. This is not a supported use of MAP Screening Tests.

6.6 Ensure That All Students Are Testing at All Terms

If some students in a group (e.g., class, grade, or school) do not test in the fall or spring (especially students who may not show high levels of growth), then end-of-year summaries of student performance will not accurately reflect how student performance changed for all students in the group over the course of the year. Therefore, schools should make sure that all enrolled students are tested each season. If students are not tested, teachers should document the reason why these students did not test.

6.7 Other Unusual Testing Practices

In general, the more the testing practices align with what is "routine," the more accurate the student results. Some unusual practices that may occur could include testing occurring on days and times outside normal school hours, or teachers removing students from the testing room to test them privately, or using accommodations, like text-to-speech for all students, to students without documented needs. We encourage schools to monitor their testing practices and include updated direction in their written guidance that addresses any of these or other unusual testing practices that are observed in their schools.

Section 7: Summary

These recommendations provide school leaders and test administrators with guidance about key issues that should be addressed when using MAP Growth test results from NWEA, particularly when used in high-stakes decisions about schools, educators, or students. These recommendations help to improve the overall reliability and validity of student test scores.

In summary, NWEA recommends that schools or districts should strongly consider the following:

1. Develop written policies and procedures prior to the start of the year that outline the professional obligations of teachers and clearly lay out the expectations for teachers and students throughout the testing process.
2. Socialize all procedures with all teachers prior to the first test administration, allowing teachers the opportunity to seek clarifications about the testing procedures, which may be different than when the NWEA assessments were used in a low-stakes capacity.
3. Expect consistency – Monitor test administration practices and data, such as test durations and student engagement, and enforce the procedural requirements at all test administration periods across all teachers and schools.

The implementation of these recommendations should ensure that student achievement and growth data are as valid and reliable as possible so that these data can provide valuable information to educators as they continue to help all students learn.

Section 8: References

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, Joint Committee on Standards for Educational and Psychological Testing (U.S.). (2014). Standards for educational and psychological testing. Washington, DC: AERA.

Aviv, R. (2014, July). Wrong Answer. *The New Yorker.*

Cottman v The State (2017). Georgia Court of Appeals. Case: A17A1050 in https://caselaw.findlaw.com/ga-court-of-appeals/1870483.html

McCray, V. (2019, March 20). Atlanta Public Schools to continue test cheating victims aid though results are hard to see. *Atlanta Journal Constitution.*

NWEA. (2018). Average MAP Growth test durations. https://www.nwea.org/content/uploads/2018/08/Average-MAP-Growth-Test-Durations.pdf

NWEA. (2018). How much time do students take to complete NWEA assessments? https://public.tableau.com/profile/jfcnwea?/vizhome/NWEAcollegescorecard/Dashboard1#!/vizhome/MAPGrowthDurations/Percentiles

Thum, Y.M. & Hauser, C. (2015). *RIT Scale Norms.* Portland, OR: NWEA.

Wise, S., Kuhfeld, M., Soland, J. (2019). The effects of effort monitoring with proctor notification on test-taking engagement, test performance, and validity. *Applied Measurement in Education 32*(2), 183-192. (brief)

Wise, S. (2019). The relationship between test-taking disengagement and performance on MAP Growth retests. (The Collaborative for Student Growth at NWEA Working Paper).

Appendix A: Sample Sections of Policies and Procedures

Board Policy Excerpt

Student Testing and Assessment Program
The District student assessment program provides information for determining individual student achievement and instructional needs; curriculum and instruction effectiveness; and school performance measured against District, State, and National norms.

The Superintendent or designee shall manage the student assessment program that, at a minimum:

1. Administers the State assessments to all students and/or any other appropriate assessment methods and instruments, including norm and criterion-referenced achievement tests, aptitude tests, proficiency tests, and teacher-developed tests.
2. Informs students of the timelines and procedures applicable to their participation in every state and local assessment.
3. Provides each student's parents/guardians with the results or scores of each State and local assessments and an evaluation of the student's progress.
4. Ensures staff use professional testing practices.

Overall student assessment data on tests required by State law will be aggregated by the District and reported, along with other assessment information, on the District's annual report card.

Procedure Excerpt

Test Administration/Re-testing of Students

Repeatedly retesting students can have many negative impacts. This is in addition to the ongoing dialogue around over-testing. We must do our best to protect students as well as the integrity of the testing process. All proctors should try to prevent the need to retest by providing students with the necessary information and a stable test environment before testing begins. The following are justification for retesting a student:

- a student's RIT score dropped 10+ RIT points from his or her last testing event;
- the student provided rapid guesses on 30% or more of items;
- the student's NWEA MAP profile reveals that greater than 10% of items were marked as rapid guesses and the estimated impact on the student's RIT score was significant (specific definitions of "significant" should be determined by the district);
- the student took 20 minutes or less to complete the assessment;
- the student exhibited test anxiety;
- the student became ill during the test; or
- there was a significant disruption or interruption (e.g. fire alarm, etc.).

Please note that this list is not exhaustive. If any of the above behaviors are observed while a student test is in progress, it is preferable to pause or suspend the test, and resume the test through the proctor menu at a later point when the student is ready/able to engage with the test.

A student cannot be retested without prior approval of an administrator. In addition to administrator approval, schools are required to collect documentation of the reason for retesting. In such a case, the following information is collected: school, test name, student name and ID, the date and RIT score of the test deemed invalid, and the reason for invalidation. This information should also be sent via email to the District MAP Coordinator.

Retest Request Form

NWEA Retest Request

Student Name: _____ **Grade**: _____

Teacher: _____

MAP Growth Test to be Retested: _____

- ☐ student showed a "substantial" decline in score between the current and previous testing period (more than 10 points);
- ☐ student rapidly guessed on greater than 10% of test items and the estimated impact on RIT score was significant, or student rapidly guessed on 30% of items. Estimated impact was _____ RIT points;
- ☐ student rushed to complete the test items or was observed responding without actually reading the items;
- ☐ student refused to take or complete the test or became overly anxious;
- ☐ student became ill during the test;
- ☐ there was a significant disruption or interruption;
- ☐ Other: _____

Required Signatures

Proctor: _____

School Administrator: _____

Parent: _____

Parent signature required if student retests more than once in an academic year

Thank you for helping us maintain the integrity of our testing process!