Guidance for Administering NWEA MAP/MPG Assessments When Results are Used for High-Stakes Purposes

The state of accountability in K–12 education has shifted, moving from a focus on school-level accountability to an increased emphasis on accountability for teachers and students. As a result, some schools are now using results from the NWEA™ Measures of Academic Progress® (MAP®) and MAP for Primary Grades (MPG) interim assessments for any or all of the following: as a component of their teacher evaluation system; to determine whether a student advances to the next grade; and/or as an indicator of student readiness for certain programs or interventions (such as special education or gifted and talented programs). With assessments increasingly used for high-stakes purposes, guidance is needed about how best to protect the integrity of the testing process and the reliability and validity of student assessment results.

NWEA offers these guidelines based on our current research and experience with school districts using our tests for high-stakes purposes. NWEA conducts regular research in this area, and we may refine or redefine these guidelines as better information becomes available.

Early Termination of Test Events and/or Retesting

There are a number of situations when it may be appropriate to terminate a student's test session early and have that student retest, or to suggest the retesting of a student after a testing session is complete. If a student gets sick during the test, appears to be intentionally trying to finish as quickly as possible, or appears to be simply guessing on the test, these are all situations where pausing or terminating the test prior to completion would be warranted.

In general, preventing invalid tests is preferable to retesting students after a bad testing experience. If students become ill, or have some form of emergency during the test, it is best to terminate the test prior to its completion. If students are rushing through the assessment or seem to simply be guessing, the teacher and/or proctor should intervene with the student; if the student does not respond, the test should be terminated before completion. Specific guidance is provided below for a variety of situations.

Regardless of the reason, teachers, principals, and even students are under significant pressure to perform well on these high-stakes tests. This pressure could result in situations where students are retested simply because student scores were different than what was expected or desired. To avoid this type of retesting practice, NWEA has created the following recommendations.

1. Establish a Written Policy on Early Termination/Retesting Guidelines that Applies at Every Term

In some high-stakes testing circumstances, there is some risk that educators may retest students simply to advance their self-interest in receiving a higher score for a classroom or school. To mitigate this risk, prior to the first round of testing, schools or districts should develop a written policy that clearly summarizes guidelines for when a test should be terminated early or a student should be retested. While not every possible scenario could be addressed in this set of guidelines, the general rules could be defined, as could the process by which retesting should be approved if the situation falls outside the scenarios included in this policy.

The general principle is that retesting is justified when situations occur that may impact the accuracy of test results. Some of the situations that may be considered in this written policy include:

- a student becomes ill 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 actually reading those items;
- a student shows a "substantial" decline in score, as defined by the school or district, between the current and previous testing period (see Item 2 below).

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. This would likely reduce instances of retesting students who have produced valid and reliable scores without cause.

Retesting policies should be applied consistently at every testing term.

2. Define What a "Substantial" Decline in RIT Score Between Two Test Events Entails.

A large decline in test scores between two administrations can be an indicator of an invalid test. In testing, there are more factors that can produce a deflated score (a score that is less than the student's real achievement) than an inflated score. We say that unhappy accidents are more likely than happy accidents in testing. Because of this fact, there are circumstances in which schools may consider retesting students if they show an unusually large drop in a test score in relation to the prior term.

NWEA does not have a formal description of what would be considered a "substantial" decline in a RIT score between test events (such as between fall and spring). However, in general, a decline of greater than 10 RIT points from a prior test event may be indicative of low student effort on the current test, or some other factor that caused the student to perform more poorly than would be expected. Thus, it may be reasonable to define "substantial" as any time a student's RIT score declines by 10 or more RIT points between test events. The definition of what is considered substantial should be defined by the district and is something that could be agreed upon at the beginning of the school year and included in the school's formal written policy on retesting.

By defining "substantial" in the written policy, situations can be avoided where a student is retested simply because one of the student's test scores, likely his or her end-of-year score, seemed atypical and resulted in a decline in test performance.

The definition of "substantial" should be applied at every term. For example, a student whose RIT score dropped by 10 RIT points from the prior spring term to the fall should be retested, just as a student whose RIT score dropped by 10 RIT points from fall to spring in the same school year would be retested.

3. Require a Rationale, in Writing, for Why a Test was Terminated or Why a Student was Retested.

In general, if a student needs to be retested, or if a test event was terminated, the rationale for why this occurred should be documented—in writing—at the time it occurs. The documentation should be collected by the school principal, district level administrators, and/or the assessment coordinator of a school or district. This provides school leaders with the ability to track which students were retested and for what reasons.

Documenting instances of early termination and retesting can be useful for two primary reasons:

- Because the reason for the termination/retest was documented, this process will protect the teacher from accusations of test manipulation in the event that a student's test performance is questioned.
- It will better assure the integrity of the testing process because there is clear transparency and accountability surrounding all retesting decisions.

Test Duration

The number of minutes it takes for a student to complete a test can be an indicator of whether a student gave appropriate effort during the testing process. The proctor and teacher should be monitoring testing closely and intervening when they see students progressing too rapidly through a test. Our current research indicates that tests completed in less than 10 minutes are unlikely to return an accurate estimate of student performance. The research also suggests that MAP or MPG test sessions that are shorter than 15-20 minutes in duration can be associated with inaccurate estimates of performance, though this may not be the case for every student who completes a test quickly.

Conversely, students can also take a long time to take a MAP test. The difference in measured RIT scores due to extra sustained effort will often be within the standard error of measure. When this happens, there is little value obtained from the extra time spent. Typical test durations vary based on the grade and season. In the fall, early elementary students generally average near 30 minutes, whereas middle into high school students average a bit over 50 minutes. In the spring, the average times are a few minutes longer. As a rule of thumb, no more than a few percent of students typically have durations longer than double the averages above. If students take notably longer than these averages, you should reflect on your current practices and consider whether to guide the test durations to more reasonable levels. For example, you may want to reinforce to students that MAP is seeking their instructional level, so it will ask questions to which they do and do not know the correct answer. Coach students to give their best effort, but move on if it is fairly clear that they don't know the answer to the presented item.

Because of this, a district's written policy could include a statement about when a student should be retested based on the amount of time he or she spends on the test. If such a statement is included, the following guidance should be considered.

1. **The test duration policy should be enforced at all terms, not just in the spring.**
    - An abnormally short test duration generally results in a score that underestimates a student's actual performance, and that may ultimately impact the amount of growth shown by the student.

Students with underestimated fall scores are likely to show inflated growth in the spring. Those with underestimated spring scores would show lower growth between fall and spring than would normally be expected. An abnormally long duration provides little additional measurement or instructional value and can negatively impact testing schedules and instructional time. Thus, to protect the integrity of the testing process and the accuracy of student testing data, schools or districts should include how many minutes on a test is necessary for the test to be considered valid. Our advice would be to set a minimum and maximum duration that teachers agree is reasonable and enforce that standard for every term. We also suggest that these standards contain enough flexibility to allow for implementation of accommodations or modifications as directed by a student's Individualized Education Plan.

2. **Schools should consider differences in test duration between the fall and spring test administrations.**
   - For a growth score to be an accurate measure of student progress, it is essential that the conditions in which MAP was administered be consistent between terms. If a student's fall test is significantly shorter than the spring test, that means that testing conditions were not consistent and that negatively impacts the validity of a student's growth score. It's particularly problematic when conditions are different for groups of students. For example, if an entire classroom of students completes fall tests in significantly shorter time than their spring tests, it calls into question the validity of the growth scores for the entire class. And in some high-stakes circumstances, it could suggest that there is an effort to game the results. Even if students take longer than the predetermined short test duration time in both the fall and spring, significant differences in test duration could still have an impact on the amount of growth a student shows over the course of the year. For example, if a student took 30 minutes to test in the fall, but then took 80 minutes to complete the spring test, the amount of growth that student shows is likely going to be greater than if the student had taken approximately the same amount of time on each test. As such, steps should be taken to ensure that students have sufficient time to complete MAP assessments in both the fall and spring, so that observed student growth is an accurate reflection of the amount of learning that occurred over the course of the year.
3. **Language regarding the need for consistency in testing conditions and test duration should also be included in a written policy.**
   - We would also suggest periodically monitoring testing condition and duration data and having conversations with appropriate school and district personnel if problems or issues are identified.

Proctoring

As the stakes around testing have increased, incidents of systematic cheating on assessments have been discovered and have received extensive coverage by the media.[1] Because of this, it is important to implement testing policies and procedures that not only prevent cheating, but more importantly, protect teachers and students from unwarranted challenges about their results.

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, providing testing conditions that are conducive to good performance, and actively monitoring testing to prevent problems. We would encourage districts using MAP or MPG to participate regularly in proctor training to ensure that at every term, proctoring practices that help maintain the integrity of the testing process are implemented.

Proctoring best practices should include the following steps.

- We recommend that both a teacher and an additional proctor monitor student testing.
- There are several reasons why a teacher should serve as the primary proctor during the testing of his or her students. The teacher is the most aware of the learning needs of his or her students, and is likely able to keep his or her students focused on the testing process better than different teachers, aides, or other instructional personnel. Research also suggests that students perform better when their teacher is present during testing.
- When results from the MAP or MPG assessment are used for a high-stakes purpose, it is good practice also to have a second proctor 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 those students being tested. In many schools, the testing coordinator could serve as the second proctor.
- The second proctor protects the integrity of testing results and protects teachers from false accusations of cheating.
- Manipulating the testing process can significantly undermine the accuracy of student test results and negatively influence decisions made based on these results. Having a second proctor in the room protects the integrity of the testing process and all subsequent testing data by reducing the likelihood that test manipulation will occur.
- Having a second proctor in the room should also help protect the teachers and students being evaluated. Teachers whose students show strong growth will likely have positive end-of-year evaluations as a result of the performance of their students. Because a neutral observer was present during the testing process, it is less likely that the performance of the students (and the performance of the teacher) will be challenged or questioned. Even if it is, the teacher can defend the performance of his or her students because they were monitored by an impartial proctor while they tested.

Additional Considerations

1. **If aggregated student test results (especially student growth) are used for high-stakes purposes, schools should ensure that all students are tested at all terms.**
   - If some students in a given group (class, grade, school, etc.) do not test in the fall or spring (especially those students who may not show high levels of growth), then end-of-year summaries of student performance would not accurately reflect how student performance changed for all students in the group over the course of the year. Because of this, schools should make certain that all students are tested in both the fall and spring terms. If students are not tested, teachers should document the reason why these students did not test.
2. **Students should be tested at a similar point in each testing term to ensure accurate growth comparisons.**
   - MAP and MPG measure growth by measuring achievement at two different points in time and calculating the difference. To put the measured growth in context, it is compared to the NWEA growth norms for students who tested in the same grade, subject, starting achievement level, and had the same number of instructional weeks. However, MAP does not use actual instructional weeks for each student; it uses values a partner selects as best representing their testing windows. Because the number of actual weeks

of instruction matter in how much a student learns, having a student test early in one term and late in another will cause more or less actual learning between tests to occur. This impacts the comparisons made to the calculated normative growth. To illustrate the potential impact, assume both the fall and spring test windows are five weeks long and the week selected for growth comparisons is the middle week in both windows (week 4 and 32 – 28 weeks of instruction). Also assume a 4th grade student whose achievement is typical is learning at a typical rate, and therefore, is shown as making typical growth when tested in these assumed weeks. If this student tests during the first week of each window or the last week of each window, the student's growth percentile is essentially the same. However, if the student has 24 weeks of instruction because the student is tested during the last week of the fall window (week 6) and the first week of the spring window (week 30), even though the student is learning at the same rate, they will learn less because of less instruction. When compared to the national growth norms based on the selected weeks, this student's growth percentile declines approximately ten percentile points. Therefore, it is recommended that once a testing schedule is established within a school for a testing term, a similar schedule should be used consistently term to term. Students who test in reading early in the testing window should continue to test in reading early is subsequent windows.

3. Students should test at a time of day that allows them to perform at a high level.

- We recommend that schools do not administer tests at times during the day when students have a limited amount of time to complete their tests, or when they may have difficulty concentrating on testing (such as right before lunch). Schools should be sensitive to the time of day when students test and should administer tests at a time when the students are going to be focused and will not have to rush to complete the test.

Summary

We have provided these recommendations to give school leaders and test administrators guidance about key issues that should be considered when using NWEA MAP or MPG test results as a factor in high-stakes decisions about educators and/or students. These recommendations should also be viewed as important testing practices even if assessment results are not used for high-stakes purposes. These recommendations will, in general, help to improve the overall reliability and validity of student test scores.

The guidance in this document is by no means comprehensive. Rather, our aim is to highlight what we believe are the major areas that merit consideration by schools or districts using MAP or MPG test results for high-stakes purposes. In such instances, increased attention needs to be given to the testing process to ensure that the testing data accurately reflect changes or improvements in student achievement and growth.

In summary, NWEA recommends that schools or districts should strongly consider implementing three broad policies.

1. Schools or districts should develop a written policy at the start of the year that clearly outlines expectations for teachers and students throughout the testing process.
2. These policies should be understood by all teachers prior to the first test administration, allowing teachers to have the opportunity to ask questions and seek clarifications about these new testing policies (which may be different than when the NWEA assessments were used in a low-stakes capacity).
3. Consistency is the key; these policies should be enforced at all test administration periods and should be the same for all teachers in the school.

These recommendations will help to maintain the integrity of the testing process, and they should provide teachers with protection and support in the event that their student's test results are made publicly available and subjected to additional scrutiny. And perhaps most importantly, the implementation of these recommendations should ensure that student achievement and growth data are as accurate as possible so that these data can provide valuable information to educators as they continue to help all students learn.

References:

1. For example, see http://usatoday30.usatoday.com/news/education/2011-03-06-school-testing_N.htm

Partnering to Help All Kids Learn® | NWEA.org | 503.624.1951 | 121 NW Everett St., Portland, OR 97209

January 2017 | KAP247