NWEA
Northwest Evaluation Association

The Importance of Data Comparability

Interim Assessment in Context

The second in a multi-part series about the importance of interim assessment.

APRIL 2014

Interim assessments offer many advantages, including the ability to gather and compare data that's collected over time—both within a single year and over the course of multiple years. An interim assessment that provides accurate longitudinal data benefits students, teachers, and administrators in different ways, but chief among them is the ability to make meaningful comparisons.

- Comparable data allow for a longitudinal perspective on student learning. This helps a **teacher and student** establish reasonable growth targets and provides context to understand a student's current achievement status in terms of growth. *Are students growing? What areas are seeing the most growth, and in which areas has growth seemed to plateau?* This can provide a teacher with information on where to focus instructional energy and class time. Some of these questions can only be answered by having data that go back in time, or longitudinal data.
- Longitudinal data give **school building administrators** the ability to identify learning trends across groups of students, and those students can be flexibly grouped by subject, grade, or classroom teacher depending on what the school building administrators are analyzing. These trends may indicate that certain curricula, programs, or pedagogical approaches are more successful than others.
- **District-level administrators** are able to see growth trends district-wide, providing them with relevant data points around allocation of instructional resources, staffing, technology, and professional development. Longitudinal data allow for informed growth projections and predictive functions. This helps district level administrators know whether their students are on track for meeting progress goals; and if not, it gives them time to do something about it.

When data are compared between groups and over time, the stability of those data becomes a virtue. In this article, we'll compare and contrast that important factor as we examine the three kinds of data comparability: horizontal, vertical, and longitudinal.

HORIZONTAL COMPARABILITY

Before data from different systems can be combined, compared, or aggregated, the data elements in all systems must be the same. They must:

- represent the same entity or attribute with the same definition
- be collected with a consistent method

When this doesn't happen, comparability suffers. Under *No Child Left Behind*, each state had to define proficiency on their individual state summative assessment scales; they then committed to every student achieving that measurement of proficiency by 2014. To date, however, the Department of Education hasn't been able to compare the performance of students across states because neither the scales nor the cut points, nor the assessments are common to all states. **In this case the data element, "proficient," is useless for comparison beyond state borders.**

VERTICAL COMPARABILITY

Data elements which claim to represent the same entity or attribute should:

- have the same definition
- be calculated in the same way in different parts of vertical data systems

What does this look like in real life? Let's say a Local Education Agency (LEA) teacher contract requires tracking of employee professional development activities in Continuing Education Units (CEUs) that represent one contact hour. In the region's state system, though, a CEU might represent completion of an approved course, regardless of the number of contact hours involved. For reporting purposes, the LEA will be required to have a second data element that is comparable to the state CEU.

LONGITUDINAL COMPARABILITY

The meanings of data elements can drift over time or they can be intentionally redefined. If the data are to be compared or aggregated over time, though, it's important to know when changes or drift have occurred. In April 1995, the College Board re-centered the scores on the SAT because student performance had shifted. Establishing 500 as the mean score—the midpoint on the 200-800 scale—made it easier for schools to interpret the scores. When the re-centering occurred, the College Board notified school districts and colleges throughout the nation that they couldn't compare students' SAT scores *after* the re-centering to the same scores achieved *before* the re-centering.

After the initial change, the College Board created conversion formulas to help schools adjust the old scores. By using the formulas, schools can compare old scores with the re-centered scores.

MAKING THE TRANSITION TO A NEW SET OF STATE STANDARDS

When a state adopts a new set of academic standards, many changes happen to curriculum, instruction, and assessment. New academic standards require a new state summative assessment with new cut scores for proficiency. This shift to a new summative assessment brings acute pressure on vertical comparability and longitudinal comparability within a state's assessment system.

For **vertical comparability**, all of the assessments within the system that predict proficiency on the state summative will need to be re-equated to accommodate the new proficiency benchmark. For **longitudinal comparability**, a break in data will happen as educators transition from their former assessment (based on old standards) to their state's new test (typically based on more rigorous standards). This break in data makes longitudinal comparability problematic, if not outright impossible.

During the transition period, this data break can be quite disruptive to students. Not only is their growth history interrupted, their performance may appear to have fallen precipitously. This fall is illusory—the standards are likely more challenging and the proficiency bar has been set higher. A higher bar on more challenging standards means that the students could be performing just as they were previously while getting different achievement results. If student performance appears to be plummeting, the lack of context, the new curriculum, and the expectations on teachers and students all combine to make for a heady, sometimes turbulent transition period.

A data break when transitioning from old to new standards can be quite disruptive to students. Not only is their growth history interrupted, their performance may appear to have fallen precipitously. This fall is illusory—the standards are likely more challenging and the proficiency bar has been set higher.

Understanding what parts of the assessment system are not changing—what offers stability, and how that stability can provide an anchor—helps immensely during this kind of systemic transition.

NORMING FOR INTERIM: WHY IT'S IMPORTANT AND WHAT IT CAN TELL EDUCATORS

Measures of Academic Progress® (MAP®) and MAP for Primary Grades (MPG) computer adaptive interim assessments from NWEA™ provide stability that allows educators to know how much students are growing and whether they are on track with their learning before, during, and after the transition period. NWEA has a stable scale and over thirty years of valid, comparable assessment data.

Educators use interim assessment such as MAP and MPG to gather information around many important educational questions.

- Are my students progressing toward their learning goals?
- How are my students performing compared to how they were performing at this time last year?
- How are my students performing compared to their peers?
- Are my students showing optimal growth?
- How does my students' growth compare to the growth of their peers?
- Are my students on track for achieving proficiency at the end of the year?

Answering some of these questions involves comparing student performance to that of other students who share certain demographic similarities,

such as grade-level. To accomplish this, assessment providers create norms that represent the aggregate responses of a representative group of students.

NWEA norm construction. At NWEA, we create our norms from a nationally representative sample. Educators can use our norms to compare the performance of individual students or a class to that of the national sample. This comparative analysis provides one kind of data point that helps educators understand student performance in a larger context.

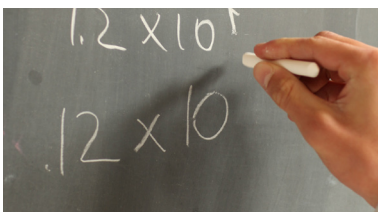
We also carefully construct NWEA norms to be independent of any specific state test. Most tests would need to have new norms calculated when the test is redesigned, or realigned, because norms are tied to answers to an existing test. Because we score MAP using Item Response Theory (IRT), however, and because we calibrate test items to a stable scale, MAP doesn't require new norms when alignments are created to new standards. Educators who use MAP always have important contexts for data interpretation and evaluation.

How we align our item pools to content in standards. MAP tests are based on pools of items that span RIT (for **R**asch Un**it**) ranges and goal areas and are aligned to standards in the sense they only cover content in the standards. The only effect a new set of standards has on items is a redefining of the scope and contents of the item pool. To the degree that new standards add or subtract content from previous standards, the items in the pool "aligned" to the new standards will differ. In any psychometric sense, any two MAP pools are equivalent and yield the same results, as would two different yard sticks if one were plastic and the other wood. This equivalence means we don't need to create new norms. Educators can compare student scores even when standards (and thus MAP tests) change—and even when a student moves and takes MAP in a different state. Growth trends persist.

Why our norms change with student population and performance. At NWEA a RIT is a RIT is a RIT, regardless of the standards or other criteria being measured. Accordingly, as long as the scale is stable, our norms won't be affected by new alignments. What will have an impact on the norms is the population of students taking the test—which is why we conduct new norming studies every three years. Because norms refer to the test-taking population in the year when the norms were calculated, norms change as student performance changes. If a student achieves the 56th percentile this year, he or she isn't being compared to students taking the test this year—the comparison is to the students tested in the year of the most recent norming study.

By design, MAP doesn't require new norms when alignments are created to new standards. Educators can compare student scores even when standards (and thus MAP tests) change—and even when a student moves and takes MAP in a different state.

Exclusive to NWEA: our growth norms. Most assessment companies provide achievement norms for their assessments, but only we provide growth norms as well. From those norms, and using other information about the student, educators can make growth projections. The norms themselves, however, don't constitute a projection. If one were making a growth projection for a student, one would use the norms and that student's past growth patterns together, calculating the student's annual growth pattern as compared to mean growth for his or her grade. Then one would project the growth for the coming year based on the comparison between actual growth in the past and mean growth in the norms. When teachers use the mean growth as "typical" and set goals based on that alone, they are ignoring some very important student-centric data.



THE IMPORTANCE OF A STABLE SCALE

Because our RIT scale has decades of stability, we can provide comparable growth and status data across 30 years and across all 50 states. This level of comparability and stability permits educators to use NWEA assessments as a bridge between prior and current standards. Even as data from old and new assessments become useless for longitudinal analysis, MAP and MPG constitute a consistent measuring device. States and the assessment consortia creating state tests aren't making an attempt to create comparability between the old tests and new tests, and to do so would be futile. Accordingly, educators and the public must rely on a third party to create the unbroken data stream that will identify whether—and how—the implementation of higher standards, revised curricula, and new assessments is changing student performance.

WHAT COMPARABLE DATA INDICATE ABOUT STUDENT LEARNING

Overall, data comparability lets teachers, administrators, parents, and students make important connections, recognize growth patterns and trends, develop achievable growth projections, and compare groups of students. Maintaining assessment data so that it can be compared vertically, horizontally, and longitudinally is one of the challenges that comes with such a data-rich culture. Data doesn't take care of itself, but with careful stewardship it can be a lever for improved student learning outcomes.

This article is the second in a multi-part series. In the next part, we'll explore the importance of multiple measures to inform instruction.

NWEA
Northwest Evaluation Association
Partnering to help all kids learn®

Founded by educators nearly 40 years ago, Northwest Evaluation Association™ (NWEA) is a global not-for-profit educational services organization known for our flagship interim assessment, Measures of Academic Progress (MAP). More than 6,800 partners in U.S. school districts, education agencies, and international schools trust us to offer pre-kindergarten through grade 12 assessments that accurately measure student growth and learning needs, professional development that fosters educators' ability to accelerate student learning, and research that supports assessment validity and data interpretation. To better inform instruction and maximize every learner's academic growth, educators currently use NWEA assessments and items with nearly 10 million students.