MetaMetrics®

POSITION PAPER

Assessment Features That Make a Difference When Examining Growth

By: Eleanor E. Sanford-Moore, Ph.D.

Today, assessments are used for a variety of purposes and at a variety of educational levels where the goal is to align the assessments with curriculum and instruction to support learning. In 2001, the Committee on the Foundations of Assessment (supported by the National Science Foundation) concluded that

> [a] *vision for the future is that assessments at all levels—from classroom to state—will work together in a system that is comprehensive, coherent, and continuous.* In such a system, assessments would provide a variety of evidence to support educational decision making. Assessment at all levels would be linked back to the same underlying model of student learning and would provide indications of student growth over time. (National Research Council, p. 9)

Since the Committee meeting and reporting in 2001, there has been much conversation about how the design of assessment systems can be enhanced and/or modified to best meet the demands of the 21st Century. These conversations often include the concepts of "through-course" or "through-year" assessment systems. Beginning in 2018, the United States Department of Education developed the Innovative Assessment Demonstration Authority (IADA) to allow a number of states to create an innovative assessment system that could be used for accountability purposes. In 2020, the Council of Chief State School Officers (CCSSO) offered a new Balanced Assessment System state collaborative to help leaders support balanced assessment systems in their states.

States and districts are always looking for ways to streamline the assessment process, to include interim/formative assessments at the state level, to provide educators with instructional value, and to develop meaningful reports for all stakeholder groups. These trends are resulting in questions about how current and disparate assessments can be integrated to produce a more cohesive assessment system. With the increased use of the Lexile and Quantile scales that have been incorporated into the reporting of numerous assessments, MetaMetrics® is committed to researching and helping to explore ways to enhance meaningful measurement.

To support this ever-changing assessment landscape, MetaMetrics has been investigating the relationship between "assessments for learning" and "assessments of learning". Many researchers have investigated the features of these assessments that need to be considered prior to implementation. This paper describes some of these features that may impact the monitoring of growth in reading and mathematics.

What is a balanced assessment system?

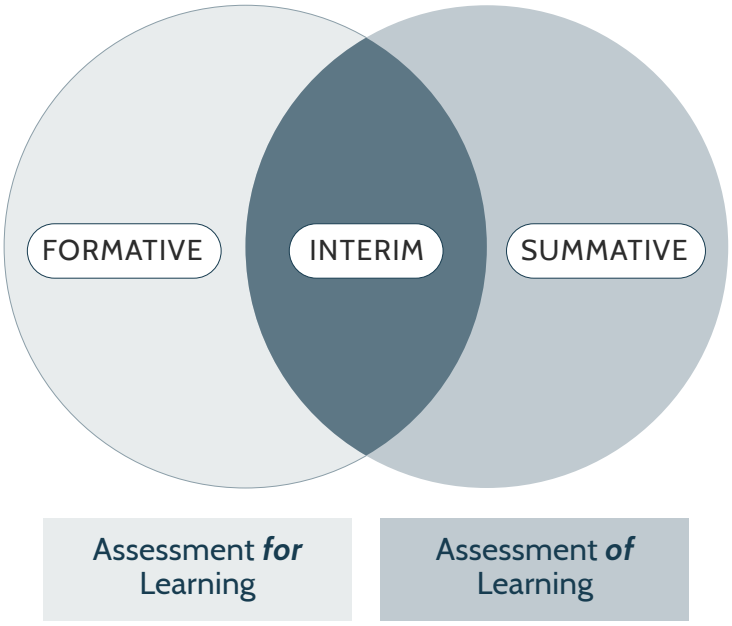
Since the Committee on the Foundations of Assessment convened in 2001, much has been written about assessment development and "balanced assessment systems", but Marion and his colleagues (2019) state that there are actually "few examples of well-functioning assessment systems" (p. 3). Today, most students complete a variety of tests during the school year where the purpose is "*for* learning" or "*of* learning". And these tests employ various scales, offer a variety of reports, and may be aligned with the state and district curricular frameworks – all of which can make comparing results from one test to another a challenge. Below are three ways to think about categorizing these tests:

1. **Summative assessments** – annual assessments typically at the end of the school year focused on long-term goals to evaluate student learning; often described as "assessment *of* learning";

2. **Formative assessments** – ongoing assessments focused on short-term goals and objectives to provide feedback and adjust instruction; often described as "assessment *for* learning"; and

3. **Interim assessments** – intermittent assessments focused on short and intermediate goals and objectives to evaluate progress and inform decision making; often described as "assessment *for* learning" and "assessment *of* learning". Interim assessments may be less instructionally relevant than formative assessments because the timing of the administration is controlled outside of the classroom.

Here is an example of how these three categories of student assessment interrelate.



(For more information about balanced assessment systems and assessment components, see Black and Wiliam, 1998; Redfield, Roeber, and Stiggins, 2008; ETS, 2018; and Marion, et.al, 2019.)

How should interim/formative assessments be evaluated?

Numerous researchers have identified and described interim/formative assessment features that need to be considered when purchasing (Herman and Baker, 2005; see also, Dadey and Gong, 2017; Herman, 2017; and Perie et al., 2007). Herman and Baker (2005) identify six criteria that can be relevant to both categories of assessments:

- *Alignment* – the assessment should reflect (1) the purpose for using the assessment (e.g., are you looking for diagnostic information about a particular student's weaknesses, are you looking for information to inform instruction), (2) the district and/or school standards and learning goals, and (3) the specific curriculum sequence. In addition, the assessment should include high quality items that assess important standards/learning objectives.

- *Diagnostic value* – the assessment should provide actionable information to describe a student's strengths and weaknesses and curriculum and/or program strengths and weaknesses, and to help to identify the sources of student difficulty.

- *Fairness* – the assessment should provide all students with a means to present what and how much they have learned and the presentation should not be impeded by extraneous factors such as stereotypes or content familiar to only some groups. The assessment should also provide accommodations consistent with instructional practices where appropriate to assess the construct being measured.

- *Technical quality* – the assessment should have adequate research and documentation related to (1) the consistency and stability of scores and subscores, (2) the precision of scores and subscores, and (3) the relationship of scores and subscores to other measures assessing the same construct.

- *Utility* – the assessment should provide useful information that can be used by educators at a variety of levels.

- *Feasibility* – the assessment should be useful given both the financial and human costs to administer the assessment and interpret and use the results.

This list is a good starting point to evaluate assessments in relation to the specific needs of a district or school. But what if we want to do more than immediately use the results from the administration of an assessment to inform instruction or monitor progress? What if we want to measure growth across the school year with those interim/formative assessments and the summative assessment that is already in place?

How is growth in reading and mathematics measured?

As described by Williamson, the measurement of growth is based on the "assumption that it is the same thing on each occasion even though its magnitude might differ over occasions" (2006, p. 3). So, first, we need a common construct of reading or mathematics. And then, we need a common measurement scale. The Lexile/Quantile scales can be the common scales needed. Lexile® reading measures and Quantile® measures are reported from most interim and summative assessments administered in K-12 schools today (see, for Lexile reading measures: https://metametricsinc.com/products/standardized-assessments/; and for Quantile measures: https://metametricsinc.com/products/standardized-assessments-2/). Measuring growth with Lexile reading measures and Quantile measures can be quite common from one assessment. But, to measure growth across interim and summative assessments, we need to examine what features of the interim and summative assessments impact what is being measured. Each test is slightly different in terms of scope and sequence, administration, scaling, and reporting, and integration; and, while often small, those differences are not accounted for by just using Lexile reading measures and Quantile measures.

How is growth on a summative assessment different from growth on an interim/formative assessment?

MetaMetrics has been actively investigating the relationship between interim/formative and summative assessment results and growth measured by the two types of assessments since 2017. Part of the research has been to examine the features of various interim/formative and summative assessments and to explore what impact those features have on the measurement of reading and mathematics growth.

The first feature we looked at was alignment of the interim/formative and summative assessments to content standards. Do they assess the same standards and learning objectives? Summative assessments are designed to measure the skills, concepts, and objectives associated with a specific grade. We need to look at content alignment from two perspectives: (1) do the items on the interim/formative assessment measure the same skills, concepts, and objectives as the items on the summative assessment, and (2) is each student administered interim/formative assessment items that measure the same skills, concepts, and objectives as on the summative assessment.

The first perspective on alignment can be evaluated by conducting an alignment study. Do the Grade 7 mathematics items on the interim/formative assessment align with the Grade 7 curricular framework developed by the state department that was used as a blueprint for the summative assessment? Do the Grade 5 reading items on the interim/formative assessment align with the Grade 5 ELA curricular framework which includes reading, writing, listening, and research?

The second perspective on alignment investigates how the assessment is administered. If the interim/formative assessment is a computer adaptive assessment (CAT) with a separate item bank for the students

in each grade, then the content of the interim/formative assessment in terms of scope and sequence is most likely aligned with the content of the summative assessment. But, if there is only one item bank that covers the scope and sequence from Kindergarten through Grade 8 (or higher), then, for a particular student the content of the interim/formative assessment may or may not be aligned with the content of the summative assessment. One Grade 7 student may be administered Grade 7 content while another student may be administered Grade 3 content (this has often been called "functional-level" testing or "out-of-level" testing or "instructional-level" testing).

MetaMetrics' latest research study included data collected between SY2013-2014 and SY2020-2021 and examined the reading and mathematics growth of approximately 38,000 students across grades K-12 on interim/formative and summative assessments. The Lexile and Quantile scales provided a mechanism to look across four summative assessments and three interim/formative assessments. These summative and interim/formative assessments varied in terms of the assessment design (fixed forms versus CAT), the standardization of the administration (specific test window or as needed), whether the assessment was specifically integrated into an instructional program, etc. We looked at the ranges of scores at each grade level on each type of assessment given within 22 days of each other for a matched sample of students. While the mean scores were very similar, the ranges of the scores were not.

	Grade	Interim/Formative Assessment Range	Summative Assessment Range
Reading	3	-220L to 1330L	85L to 1095L
	6	-60L to 1755L	285L to 1400L
	11	-220L to 1940L	690L to 1635L
Math	3	-580Q to 1155Q	40Q to 1150Q
	6	-245Q to 1875Q	155Q to 1455Q
	11	-20Q to 2120Q	590Q to 1665Q

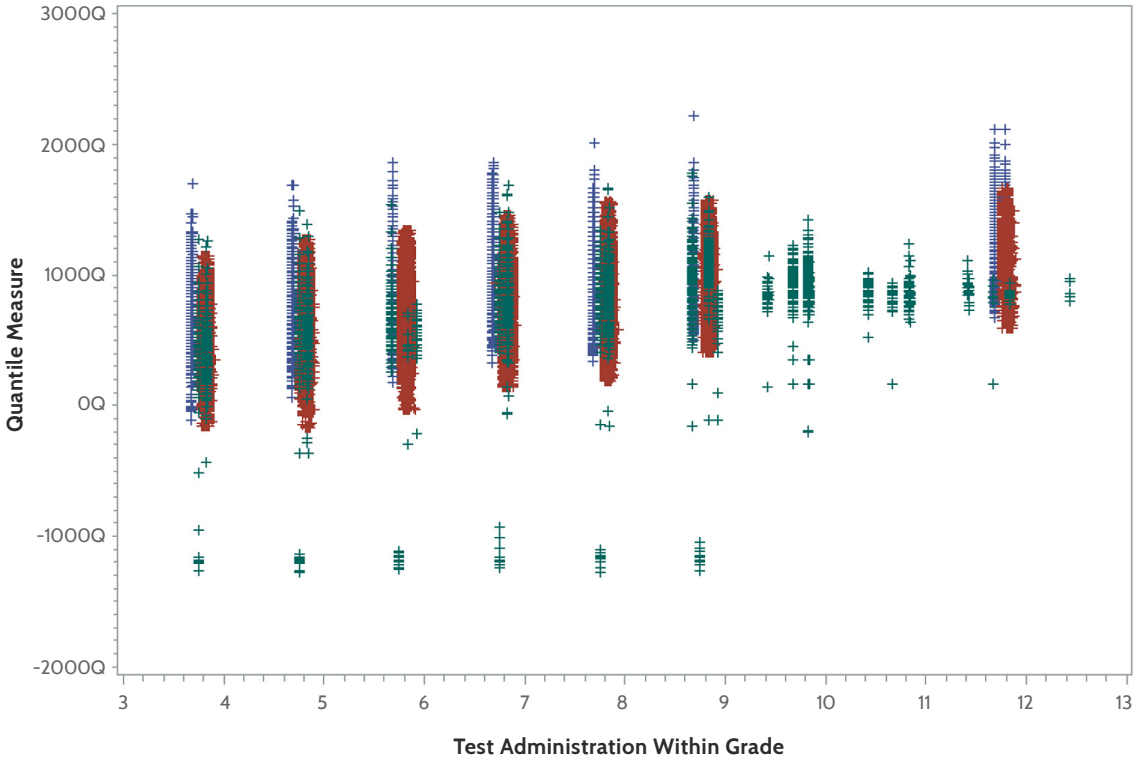
The interim/formative assessment score ranges are much larger than the summative assessment score ranges. For example, the Grade 3 reading interim/formative assessment scores span a range of 1550L points, whereas the Grade 3 summative assessment scores only span a range of 1010L points. And, this increase in range can lead to the interim/formative assessments providing more instructionally relevant information — which should be of no surprise to anyone. But, to measure growth across interim/formative and summative assessments, we need the assessments to be aligned from both perspectives.

Next, we need to examine the purposes for administering and the uses of the data from interim/formative and summative assessments. Summative assessments are focused on long-term goals and are typically administered to evaluate student learning after an extended period of instruction. Conversely, interim/formative assessments are focused on short- and intermediate-term goals and are typically administered to evaluate student learning and to collect diagnostic information to inform instruction. Given the

difference in focus in relation to instruction, students may do well on an interim/formative assessment during one administration window and not as well on the interim/formative assessment during the next administration window depending on how well they understood the instructional content. Examining growth using each of the interim/formative assessments as a baseline may give very inconsistent results for a student.

Then, we need to examine the actual administration protocols of the interim/formative and summative assessments and investigate how that may impact the examination of growth. Summative assessments are typically administered to all students within a specific administration window. As you can see in the graph in *Figure 1* of Quantile measures from summative assessments, students in Grades 3 through 8 generally took the assessments in the spring across all three summative assessments (blue, green, and red data), whereas high school students took the assessments in the fall or in the spring (likely courses taken under a "block scheduling" paradigm; blue and green data). For each of the summative assessments, the administration is fairly consistent for all students.

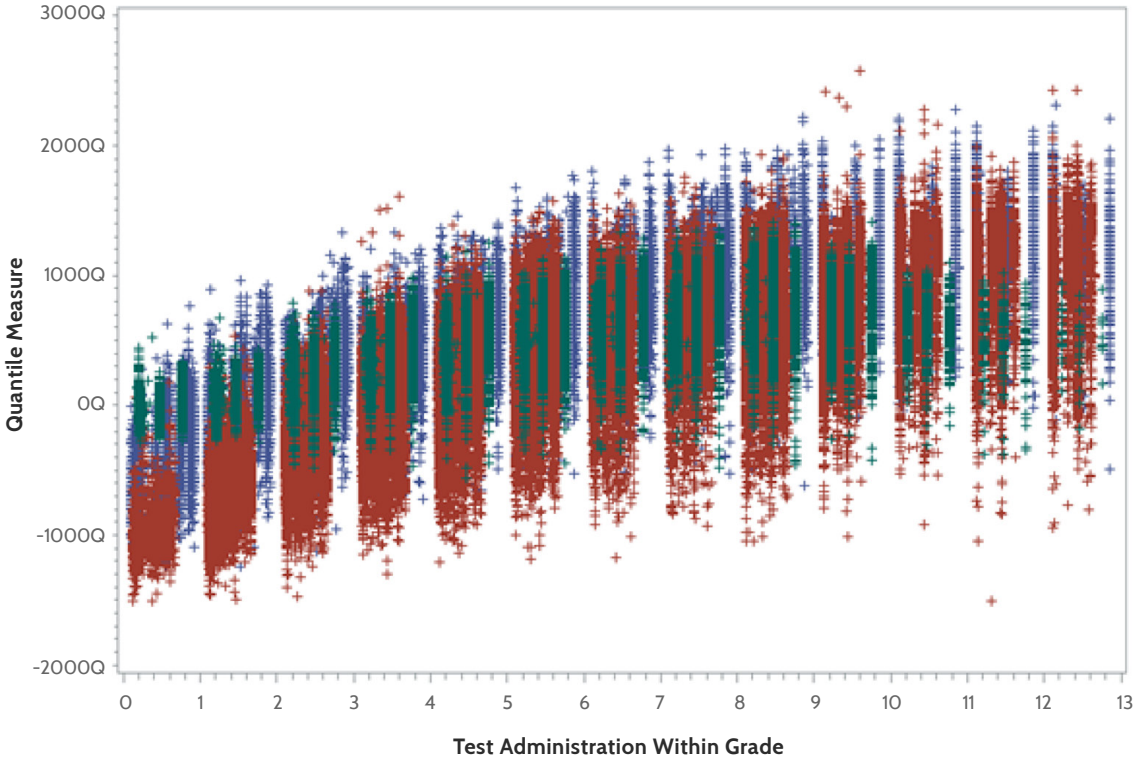
Figure 1. Distribution of summative assessment data (reported as Quantile measures), by test window within the school year.



One question we investigated was "Is the interim/formative assessment administered on a fixed date for all students like the summative assessment?" or "Is the interim/formative assessment administered when each individual student reaches a certain point in the instructional program?" The amount

of growth observed will vary based on both the length of time and the amount of content covered between administrations. As you can see in the graph in *Figure 2* of Quantile measures by interim/formative assessment date from our research, some interim/formative assessments are given on a very specific schedule (typically three times per year; blue and green data) and other assessments are given continuously throughout the school year (red data). While continuous administration may be more instructionally relevant, it is very different from how summative assessments are administered.

Figure 2. Distribution of interim/formative assessment data (reported as Quantile measures), by test window within the school year.



Other things to be investigated include what stakes are attached to the results. Are all students being tested or only certain students being tested for diagnostic purposes? The amount of growth for students at the lower end of the distribution may be very different from the amount of growth typically observed by students at the middle and upper end of the distribution. Finally, these features of the administrations of interim/formative and summative assessments will likely impact student motivation which has been observed to affect interim test scores (Finn, 2015). When students are not motivated to do well on an assessment, the results may be biased downward and underestimate student ability. Still, at the heart of things, we may be examining the amount of growth a student makes using a baseline interim/formative assessment where the student was unmotivated.

Conclusion

The Lexile and Quantile scales put student ability and complexity of reading materials or mathematical skills and concepts on a common scale. This allows Lexile and Quantile measures to be used with the various categories of assessments to describe and interpret results within a common, consistent framework. A Lexile reading measure reported from one assessment has the same interpretation (the level of text a student should likely be able to read successfully and comprehend) as a Lexile reading measure reported from another assessment. But, if we want to compare results from multiple assessments, we first need to evaluate each assessment within an assessment system and understand its purpose — "assessment for learning" or "assessment of learning". This evaluation can become even more complicated when a district is using multiple interim/formative assessments. Comparisons will need to be made between each interim/formative assessment and the summative assessment and then across the various interim/formative assessments. The results of the evaluation across and between assessments should be used to understand how closely aligned the interim/formative assessments are with the summative assessment and, therefore, used to interpret trends and patterns of growth.

The Lexile and Quantile scales can be linked with assessment scales or be the actual scales of assessments. This may provide a means for understanding how similar or different various assessments are and the extent to which they measure growth in student reading and mathematics ability consistently – comprehensively, coherently, and continuously. Further research on student growth using longitudinal datasets across multiple interim/formative and summative assessments and, most importantly, contextual information, is the first step in this direction.

References

Black, P. & Wiliam, D. (1998). *Assessment and classroom learning. Assessment in Education: Principles, Policy & Practice,* 5(1), 7-74. DOI: 10.1080/0969595980050102

Dadey, N. & Gong, B. (2017, April). *Using interim assessments in place of summative assessments? Consideration of an ESSA option*. Washington, DC: Council of Chief State School Officers.

Educational Testing Service (ETS). (2018). Understanding balanced assessment systems: Integrating assessment in a way that works for students and their families, the school, the district, and the state. Princeton, NJ: Author.

Finn, B. (2015). *Measuring motivation in low-stakes assessments* (Research Report, RR-15-19). Princeton, NJ: Educational Testing Service. doi:10.1002/ets2.12067

Herman, J. (2017). *Interim assessments in brief*. Los Angeles, CA: The Regents of the University of California.

Herman, J. and Baker, E. (2005). Making benchmark testing work. *Educational Leadership*, 63(3), 48-54.

Marion, S., Thompson, J., Evans, C., Martineau, J., & Dadey, N. (2019, April 30). *The challenges and opportunities of balanced systems of assessment: A policy brief.* Dover, NH: Center for Assessment (National Center for the Improvement of Educational Assessment).

National Research Council. (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academies Press.

Perie, M., Marion, S., Gong, B., & Wurtzel, J. (2007). *The role of interim assessments in a comprehensive assessment system: A policy brief.* Washington, DC: Aspen Institute. Retrieved from https://www.achieve.org/publications/role-interim-assessments-comprehensive-assessment-system

Redfield, D., Roeber, E. & Stiggins, R. (2008). Building balanced assessment systems to guide educational improvement. A background paper for the keynote panel presentation at the National Conference on Student Assessment, June 15, 2008, Orlando, FL.

Williamson, G.L. (2006). *What is expected growth?* An occasional white paper by MetaMetrics. Durham, NC: MetaMetrics.

METAMETRICS is an award-winning education technology organization that offers the only scientifically valid, universal scales for measuring silent and oral reading and listening (Lexile®) and math (Quantile®) with plans to develop measures for writing. The Lexile and Quantile Frameworks measure student ability and the complexity of the content they encounter. Lexile and Quantile measures and related technologies link assessment to instruction and provide next steps for students of all ages and abilities. The measures also provide valuable insights about students' potential for growth. MetaMetrics' measures, products and services are licensed to dozens of education product companies to help achieve that growth. For 35 years, MetaMetrics' work has been increasingly recognized for its research-based approach to improving learning. For more information, visit **MetaMetricsInc.com.**

MetaMetrics®

METAMETRICS®, the METAMETRICS® logo, LEXILE®, the LEXILE® logo, QUANTILE®, and the QUANTILE® logo are trademarks of MetaMetrics, Inc., and are registered in the United States and abroad. The trademarks and names of other companies and products mentioned herein are the property of their respective owners. Copyright © 2022 MetaMetrics, Inc. All rights reserved.

MM0378W