A Thoughtful Educators' Guide to Interim/Benchmark Assessment

A Guide to Choosing, Developing, and Using Data from Interim/Benchmark Assessments

By Jim Gullen, Ph.D.

Michigan Assessment Consortium
Improve learning.
Increase success.

Table of Contents

A Thoughtful Educator's Guide to Interim/Benchmark Assessment

By James A. Gullen, Ph.D.

Forward

Interim/Benchmark Assessment has been receiving increased attention in public education over the past several decades. The Federal No Child Left Behind legislation gave assessment, more specifically summative assessment, a much larger role in public education due to the accountability requirements enacted in that law. In response to these requirements, greater attention has been paid to educational assessment than ever before. Clarity around what types of assessments are available for use and to what ends they can be employed is essential for an efficient educational system. The notion of a balanced assessment system as an integral part of the overall educational system is gaining wider acceptance. Interim/Benchmark Assessments are part of a balanced assessment system.

Additional legislation around educator evaluation has also provided impetus for more interim/benchmark assessment. Requirements around the monitoring of student growth in academic achievement play right into the capabilities of a well-crafted interim/benchmark assessment. Unfortunately, in too many instances, interim/benchmark assessments are chosen and used with less than desirable results.

This guide is written with the aim to help the reader understand interim/benchmark assessments in the broader context of a balanced assessment system. Further, it addresses some of the characteristics of interim/benchmark assessments and how they are constructed. Additionally, several potential uses of these types of assessments are explored along with relevant issues related to each use.

If you're reading this guide, you most likely have an interest in interim/benchmark assessment probably stemming from your role in public education. Perhaps you serve on a group that needs to choose an interim/benchmark assessment for use in your school or district. Maybe you already have ties to a school that uses an interim/benchmark assessment and are not happy with the results that you are getting. In either case, this guide may provide more context and information to help you make decisions with respect to how to choose an assessment and how to use the results in a way that provide useful information to you and your colleagues.

This guide assumes that the reader has experience with and a basic understanding of educational assessment. If you've spent time in a classroom or a school, you probably have enough background knowledge to make this guide accessible. You do not need to have formal statistical or psychometric training to read this guide. If you are looking for guidance on how to

do such technical things as fit a psychometric model, this is not the guide for you. If you find that you would like more information around some of the topics discussed, you are invited to visit www.michiganassessmentconsortium.org. There you will find lots of resources that can help with more information around various topics related to all aspects of educational assessment.

The Importance of Balanced Assessment Systems

Assessment has the ability to play a crucial role in effective education. When done well, an assessment system provides all stakeholders with relevant, timely, and useful information on the educational process. When done poorly, assessment can be a waste of time and resources and in really bad situations, provide inaccurate data from which to make decisions. In public education, students, parents, teachers, administrators, and policy makers all have legitimate needs for educational assessment information, but the types of information that they each need is very different. Additionally, information about the effectiveness of what has happened in the classroom, what is currently going on in the classroom, and where instruction should go next are all necessary. If any piece of information is not available to the correct stakeholder at the appropriate time, the educational process will not be as efficient and effective as it could be.

The importance of providing necessary information, of the correct types, to all stakeholders in public education is hard to overstate. In recent years, the notion of a balanced assessment system has received increased attention in an effort to improve the instructional system.

Simply put, a balanced assessment system can be thought of as a system that provides the right assessment information to the right users at the right time.

Simply put, a balanced assessment system can be thought of as a system that provides the right assessment information to the right users at the right time. In a balanced assessment system, no user is denied relevant information for their needs and assessments that are not needed, either due to redundancy or lack of usefulness, are not administered.

Uses of Assessment Information

Assessment information needs are different between groups. The type of assessment information a student needs is very different from the information a school board member, for example, needs. A balanced assessment system will employ different types of assessments to provide the appropriate type of information to each user. Assessments in a balanced assessment system can be thought to lie along a continuum classifying their typical uses and some of their design characteristics.

On one end, are **formative assessment practices**. These types of assessment activities are very closely tied to the instructional offerings in the classroom. The information obtained from these practices are most useful to students and teachers in monitoring progress toward meeting the educational targets and determining when and how to move forward in the learning. Formative assessment practices are an integral part of monitoring a student's individual educational journey toward their educational objectives.

*On one end, are **formative assessment practices**. These types of assessment activities are very closely tied to the instructional offerings in the classroom.*

On the other end of the continuum are **summative assessments.** Summative assessments are typically administered after instructional activities have completed and students have had the opportunity to meet the educational targets. Information from these types of assessments speak to how well the educational targets have been met, or not. Typically, the provided information doesn't address specific areas of need to improve student attainment of the learning objectives as would result from formative assessment practices. Summative assessment data is often more useful to policy makers and administrators than it is for students. As summative assessments are administered after instruction and learning has taken place, they do more certifying of the learning than they do guiding the learning. In essence, when summative assessments are administered, the "train has left the station" in terms of adjusting classroom offerings. This is not to say that there aren't legitimate needs for summative data. Summative data play an important role in monitoring the educational system.

Information from [summative assessments] speak to how well the educational targets have been met, or not.

Lying between these two broad categories of assessment are **interim/benchmark (I/B) assessments**. These assessments fill the space between formative assessment practice and summative assessments. Depending on the intended use and design, interim/benchmark assessments may be able to provide formative data for use in instruction. Sometimes, interim/benchmark tests are designed to provide more summative information. Interim/benchmark assessments can be built to provide a wide range of useful information.

Interim assessments can provide a wide range of information to a variety of users.

Interim assessments can provide a wide range of information to a variety of users. Information from I/B assessments can document the degree of mastery of the learning objectives. They can also be built to provide information to students and teachers on progress made toward the mastery of learning objectives. Sometimes the purpose of an assessment is to give a prediction of how students may score on future assessments. Students who are predicted to not do well on the future test, based on the predictive assessment, have

the opportunity to receive additional educational support in an attempt to improve their knowledge and skills for the upcoming summative assessment.

Implications for Development and Validity

Assessments are crafted for specific purposes. The stated purpose of an assessment will direct some of the characteristics of the assessment. If the purpose of an assessment is to document the mastery of a large amount of material, say as in a final exam, the content that makes up the assessment must represent the broad range of material that made up that course. If the purpose of the assessment is to diagnose student misunderstandings, then an assessment will cover a narrower range of material but will cover it in more detail. If the purpose of an assessment is to track progress in mastering learning objectives, items must be included that fall all along the continuum of mastery of those objectives. Learning progressions can be very useful for these types of assessments if validated learning progressions are indeed available for the desired content.

Because assessments are built for a specific purpose, they can't be used for just any stated purpose. When an assessment is well crafted, administered appropriately, and the results are used as designed, we say that the results are valid. If we use an assessment for a purpose for which it was not designed or administer it in a way that was not intended, we say that the results and or the use of that assessment are not valid. It is important to note that assessments are not valid or invalid, per se. Rather, assessments are valid or invalid for a specific, stated, purpose. We may use an assessment in our schools for a valid reason, but if we try to use those same results for something else, we run the risk of getting invalid results for the second purpose. If an assessment is built for multiple purposes, evidence of the validity of that assessment *for each purpose* must be provided and evaluated. There will be more on this later in this document.

When an assessment is well crafted, administered appropriately, and the results are used as designed, we say that the results are valid.

Interim/Benchmark Assessments

Rationale Behind Interim/Benchmark Assessment

If formative assessment practices provide the guiding information useful for learning and instruction in the classroom, and summative assessments provide information on the degree to which the learning objectives were achieved, why do we need interim/benchmark assessments? Interim/benchmark assessments gained popularity shortly after the federal No Child Left Behind (NCLB) was enacted. Under NCLB, summative assessments took on increased importance due to the accountability provisions within the legislation. States were required to develop accountability systems that relied heavily on scores from summative assessments given to all public-school students statewide. While the development and administration of these

summative assessments fell to the states, the implications of the accountability directly impacted local districts and schools.

School systems needed a way to track student progress toward the attainment of learning objectives that occurred before the state summative assessments were administered for accountability purposes. Results from the NCLB accountability assessments came *after* learning was to have taken place, and as such were of very little use in adjusting classroom instruction to improve learning for current students. It may be possible to adjust curriculum and instruction based on summative assessment scores for future students, but not for the students whom the scores are based on.

Purposes of Interim/Benchmark Assessments

(Refer to the Appendix at end of this document for more detailed description of each of these uses and the characteristics of Interim/benchmark assessments.)

Instructional Use

Interim/benchmark assessments can be built to be administered in such a way as to **document student progress** toward meeting educational outcomes. In this way, educational activities can be tailored for students to help maximize their achievement before they have to participate in the state accountability summative assessments. This instructional use of interim/benchmark assessments is, perhaps, the most useful to students and teachers in classroom instruction as it provides information related to the content and skills that they are working to master.

Predictive Use

Taking this idea one step further, some interim/benchmark assessments were built for the specific purpose of **predicting future performance** — how students would perform (score) on the accountability assessments. Assessments of this type may be less useful to the classroom for a few reasons. First, this purpose of assessment predicts a score on a future test. In most cases, a test score is not the goal of education. It is the interpretation of a test score in relation to educational outcomes that is the goal. Further, tests designed to predict future performance on other tests make use of statistical relationships between the two tests. The test built to predict may look very little like the test for which the score is being predicted. This can lead to a lack of confidence in the use of the predictive assessment. Finally, it can be difficult to maintain an instructionally useful predictive test.

Consider this case; we have an I/B assessment that predicts future performance on the future accountability assessment quite well. Students will fall into basically two groups: those who are predicted to do well on the future assessment and those who aren't. Based on these results, students predicted to not do well will receive additional support in content and instruction before the future assessment, which will give them a higher probability of doing well on the future assessment. In the extreme, all students would end up doing well on the summative assessment. Those that were predicted to do well will do well, and those who weren't received support to increase their achievement and thus ended up doing well on the assessment. All

students did well on the test in the end. This is one reason why it is difficult to maintain a useful predictive assessment.

Evaluative Use

A third category of use for interim/benchmark assessments is evaluation. In this use, results from the assessment are used for **program evaluation purposes**. If the I/B assessment is built to monitor student progress in meeting the educational goals and results show that students are having difficulty meeting some of those goals, educational offerings related to that content deserve a closer look. Insufficient, inefficient, or ineffective curriculum or instruction aren't the only possible reasons for low scores, but in a program evaluation use, they would be looked at if there is confidence that the I/B assessment is a good (i.e. valid) assessment of the instructional goals.

It should be noted that the evaluative purpose of an interim/benchmark assessment and instructional uses of the assessment are very similar. The difference between the two is *what is being monitored or evaluated*. When an interim/benchmark assessment is used for instructional purposes, results are used to make inferences about the students. Depending on the purpose of the assessments, inferences about what content has been mastered, what is still in development (and how far along), and possibly where to go next for students are made based on test results. When the results are used for evaluative purposes, it is the content and instruction that are being looked at. Inferences about the efficacy of classroom offerings are made when I/B assessments are used for evaluative purposes.

The instructional and evaluative purposes of interim/benchmark assessments are very closely related. If an I/B assessment is built for instructional purposes, it is often not much more work to make that assessment useful for evaluative purposes. In addition to ensuring that the content of the assessment matches the content of the instructional objective, the format or context in which content is presented and questions are asked must be similar if the results are going to be used to make inferences about the instructional offerings.

In this case there is a tension between the two potential purposes. We need the format of the assessment to be close to the way that content was presented during instruction, but we don't want it so close that it only reflects rote memorization of instructional methodology. Students need to be able to apply and adapt their learning to novel situations on assessments. The more the assessment strays from the format of instruction, the more tenuous the inferences made about instructional materials are. If student performance on the assessment is high, and the format of the assessment is quite different from the instructional materials, we might be quite pleased with the results as the results are evidence that students have met the educational goals and are able to apply them in new situations that are different than the ones used during instruction. On the other hand, if results are poor on the assessment, do we attribute this to ineffective instructional offerings or did students do well with the content but the format of the assessment is so different they are not able to apply what they learned. Professional judgement and sound test construction principles need to be followed to adequately address this tension.

Characteristics of Interim/Benchmark Assessments

Although Interim/benchmark assessments may be built for different purposes, most I/B assessments will share two common characteristics; they will be administered under standardized conditions and the results from the assessment can be aggregated among groups. These two characteristics are inter-related and contribute to the usefulness of results from interim/benchmark assessments.

Standardized administration conditions

Administering a test under standardized assessment conditions simply means that each student who participates in the assessment participates under the same "ground rules". During test construction, the administration conditions of the assessment must also be determined and documented. Considerations such as format of the assessment (pencil and paper, oral), resources (calculator, thesaurus, periodic table) available for use or not, and perhaps timeframe, all need to be addressed. Decisions about the standardized assessment conditions impact the types of inferences we can make from results from using the assessment. If we allow students to use a copy of the periodic table as a resource during the assessment, we can't make valid inferences about their level of mastery of memorizing the format and content of the periodic table if that is one of the educational outcomes to be assessed on the test.

Although Interim/benchmark assessments may be built for different purposes, most I/B assessments will share two common characteristics; they will be administered under standardized conditions and the results from the assessment can be aggregated among groups.

Interim/benchmark assessments exist in the broader context of our educational system. The need for standardized assessment conditions must be balanced with other legitimate educational concerns. Students with Individualized Education Plans (IEPs) that call for certain testing accommodations will need to have their administration conditions modified to meet the requirements of the IEP if the student is going to participate in the I/B assessment. Professional judgement will have to be used to determine whether these accommodations have an impact on how that student's results are used. Consulting the 7 Principles of Universal Design may be useful during test development in mitigating this issue.

Ability to aggregate scores

The second characteristic most interim/benchmark assessments share is the ability of results to be aggregated (combined) across groups. As one of the common purposes of I/B assessment is to monitor student progress of mastery, or progress toward mastery, of learning objectives, the number or percentage of students in various categories is useful information. It is the characteristic of aggregation of scores that requires assessment conditions to be standardized. If one group of students is allowed to use calculators on a mathematics assessment and another is not, it is unlikely that their scores are comparable. This is also the reason that

formative assessment practices typically don't require strict standardized administration, the results from formative assessment activities are not designed to be aggregated or compared across groups.

Construction of Interim/Benchmark Assessments

Information from Two Types of Interim/Benchmark Assessments

As mentioned earlier, assessments are built for specific purposes. Additionally, the intended purpose of the I/B assessment will guide some aspects of its development. Interim/benchmark assessments can be thought of providing information in one of two ways; curriculum-based information or scale-based information. The type of information being provided by the I/B assessment guides how the assessment is developed.

If an interim/benchmark assessment is providing **curriculum-based information**, claims such as how many standards have been mastered are typically made based on the assessment results. Information related to how far students have come along on learning progressions may also be available depending on how the I/B assessment is constructed. In either case, it is the curriculum and the learning targets that provide the frame of reference for the scores of the assessment. This type of information is useful for standards-based reporting. Traditionally, this type of score reporting has been referred to as **criterion-referenced scoring**.

When an interim/benchmark assessment is providing **score-based claims,** a scale score is provided based the results of an assessment. This score provides a location for the student's performance along a continuum or dimension of achievement in the content area. The items on the assessment define the dimension of the assessment target. The dimension ranges from lower achievement (or ability) to higher achievement (or ability) and the student's score gives us a location that can be compared to other scores or to predefined performance standards placed on the dimension. Unlike the results provided from a curriculum-based I/B assessment, a student's score on the continuum doesn't provide specific information related to what standards have been mastered. The scale or dimension provides the frame of reference with which to interpret the student score, not the content, specifically.

Interim/Benchmark Assessment Construction for Curriculum-based Score Interpretation

If curriculum-based claims are going to be made based on results from an interim/benchmark assessments, the content and format of the assessment items are very important.

If curriculum-based claims are going to be made based on results from an interim/benchmark assessments, the content and format of the assessment items are very important. In these types of assessments, the domain-sampling model is employed. The domain is the content area, or areas, that are being assessed. Sampling refers to the selection of items and/or tasks that will constitute the assessment. In test construction, it is rare that

items are randomly selected, rather, items and tasks are selected to span the range of content in the domain. Additionally, the difficulty of the tasks and items will be chosen based on the desired information from the assessment. If results related to student progress toward meeting educational objectives are desired, items spanning a range of difficulty would be required. If the assessment is based on learning progressions, items placed at various points along the progression would be included. On the other hand, if more summative information is desired—the number of standards that students have mastered, for example—items and tasks at a level that represents mastery would be included. Less difficult items that "scaffold" to the difficulty level of mastery would not be included.

Test blueprint

A very useful tool for constructing a curriculum-based interim/benchmark assessment is a test blueprint. A test blueprint documents the **learning targets** that are being assessed (the domain) as well as **the level of cognitive complexity** and **format of the items and tasks** that are included on the assessment. One common format for a test blueprint is a grid where the rows are the learning targets and the columns represent the levels of cognitive complexity of the assessment items. In each cell, the number and format of items are presented. Useful taxonomies for the levels of cognitive complexity include Bloom's Taxonomy as well as Webb's Depth of Knowledge (DOK). There are others that can be used. What is important is that a differentiation of cognitive load is specified and documented, particularly if results are used to document progress toward mastery of standards or learning progressions.

If the information desired from the assessment is the number of standards mastered, a sufficient number of items, of appropriate difficulty, must be included. If the items on the assessment are to be rubric scored, the rubric level that constitutes mastery must be clearly defined. There is no hard and fast rule with respect to how many items are required to be included on a test to be able to infer mastery. Educational goals come in many different "grain sizes" and different sized goals would require differing numbers of items to assess. Additionally, the format of the items also plays a roll. It is not uncommon for performance tasks to be more robust than selected response items. As such, we may be able to get more information regarding student achievement with a single performance task than we would from many more selected-response items. Professional judgment is required during the development of the assessment to determine the number and types of items included on the assessment as well as how many of those items must be answered correctly and/or what performance level must be achieved.

Strand scores

An interim/benchmark assessment that is built to monitor student progress or achievement with respect to content standards will typically provide results/scores tied to multiple content standards in addition to, or in place of, an individual overall score. Often these scores are referred to as strand scores.

Strand scores are often presented in one of two ways. The first is either the fraction or percent of points earned out of the possible points for each strand. This type of score reporting appears to be straightforward both in presentation and interpretation. Statements such as "earning 80% of the points (or 4 out of 5 points) on strand A," for example, are commonly made. This interpretation is complicated if a strand is assessed by differing types of items, some selected response and a rubric-scored item for example.

A second way that strand scores can be reported are by setting a performance standard, or standards, for each strand. This type of score reporting requires more work and documentation than simply reporting percent correct, but may provide more useful information from which to make inferences about student achievement.

A test blueprint can provide very useful information regarding strand scores. The rows in the blueprint will clearly present the strands that are assessed by the I/B assessment. Additionally, the blueprint documents the numbers and types of items that are used on the assessment to assess the strands. This documents the decisions made during the domain sampling and provides summary information that can be used in establishing the validity argument for use of the interim/benchmark assessment. Combining a test blueprint with documentation on who was involved in the test construction is important information for future users of the assessment. That information along with documentation of how test construction activities were carried out provide the basis for demonstrating an assessments validity for a stated purpose. Providing information on the processes used in selecting content, determining how many items to use on the assessment, and the method used for score reporting and standard setting provide future users of the assessment valuable information about the appropriate uses of the assessment.

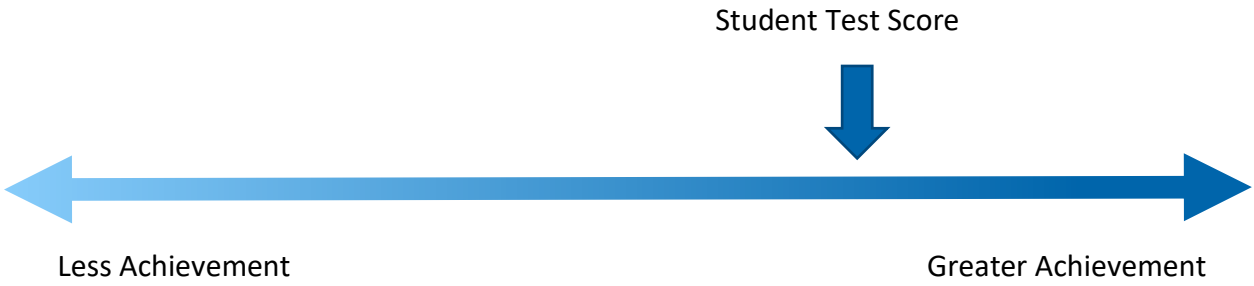
Providing information on the processes used in selecting content, determining how many items to use on the assessment, and the method used for score reporting and standard setting provide future users of the assessment valuable information about the appropriate uses of the assessment.

Interim/Benchmark Assessment Construction for Scale-based Score Interpretation

Some interim/benchmark assessments are built such that they define a specific scale. In this type of I/B assessment, a student's score places them along a continuum or dimension of performance. (See Figure 1.)

The placing of student performance along a well-defined scale is useful for when reporting or research is desired that requires math to be performed on the scores. Aggregating scores for

Figure 1: Student Performance Continuum



groups of students and comparing differences in single student scores for purposes of evaluating progress or growth all require mathematical computation on scores.

Scaling is conducted so that differences of the same magnitude represent the same difference in achievement regardless of where those differences occur along the scale. By contrast, if we look at curriculum-based reporting, the amount of achievement required to go from 5 out of 6 (5/6) points on a strand to 6 out of 6 (6/6) points on that strand might be very different than the amount of achievement required to go from 0/6 points to 1/6 points on that strand. This is a very important consideration especially for interim/benchmark assessments that are going to be used to monitor student growth.

The development of scales that support precise location of student performance and precise, interpretable, mathematical calculation are not easy to construct. Complex mathematics, large representative samples of students, and strong assumptions about the nature of the content spanned by the scale are needed. For these reasons, formal scaling is often only achievable by state testing programs or independent testing companies, who assessment large numbers of students.

In order for changes in position along the tested scale to be interpretable, the content, in most cases, must be unidimensional. That is, the scale must relate to only one thing. We interact with many unidimensional scales in our daily lives. Things like height, weight, credit card balance, age, are all unidimensional scales. In education, unidimensionality is more elusive. We may think that a social studies test is unidimensional, but what if the assessment contains items from history, political theory, and economics? In that case is the social studies test *really* unidimensional? Notice that if we are building a test to give us strand scores, we may be building a test that isn't unidimensional. (In the interest of completeness, there is ongoing research in the area of multi-dimensional item response theory (M-IRT) that allows for scaling of tests that are decidedly not unidimensional, but M-IRT is way beyond the scope of this paper.)

Finally, I/B assessments that are built for scale-based interpretation employ scaling activities that require data from test takers who actually took the test. In this sense, the results of the scaling depend, in part, on the people who take the test to provide the data used in scaling. This is NOT the same as traditional norm-referenced, percentile-rank scores where we make statements such as "Student A scored better than 75% of test takers in the norm group on this assessment." While scale scores are not providing information about performance in terms of other students' performance, the characteristics of the students in the scaling group in their performance of the items determines the nature and characteristics of the scale.

Student test scores resulting from scale-based I/B assessments are simply locations on the scale of that assessment. In that sense, the number is fairly arbitrary. The number of the scale score itself does not provide information about the number of items answered correctly or incorrectly. It only provides the student's location on the scale based on their taking of that assessment.

While the actual numbers on the scale are arbitrary, the differences between them are not. The scale is crafted such that differences between the points along the scale are all equal intervals. In theory, a difference of 5 scale score points represents that same change in achievement whether it occurs at the low end, middle, or high end of the scale. Contrast this with the strand-based reporting where we hypothesized that going from 0/6 to 1/6 (+1 gain) points on a strand represented less gain in achievement than going from 5/6 to 6/6) (+1 gain) on that strand.

Interim/Benchmark Assessments and Student Growth

We are all familiar with a notion of growth. We watch as our children grow in height and weight or we watch as our retirement or other bank accounts (hopefully) grow in value, as examples. In these instances, growth is a rather straightforward concept and calculation involving a simple subtraction. You take the current value and subtract from it the previous value and you have the amount of growth over that period of time. Pretty straightforward. Two things make this simplicity available: we are talking about unidimensional (single dimension) growth, and we have an equal-interval scale (ruler, scale, amount of money) where differences are interpreted the same wherever they occur along those scales.

In education, growth is often not quite this straightforward. As previously mentioned, content is often multidimensional. When we are looking at changes over time in things that are multidimensional, interpretation can be complicated.

When we are looking at changes over time in things that are multidimensional, interpretation can be complicated.

As an example, we know that height and weight are each unidimensional measurements; children grow taller and heavier as they age. Putting these two types of growth together gives us the notion of "bigger." Children get bigger as they age but quantifying bigger is more difficult than quantifying taller or heavier. We have indices for human "bigness," like the body-mass index (BMI), but a quick search of the literature reveals issues with BMI. Many of these issues arise from interpreting and comparing BMI values.

Academically, it is often easier to see that students are getting the equivalent of "bigger" in their content areas than it is to measure the equivalent of taller or heavier. Again this is due to the nature of academic content; it is often not unidimensional, even in a single content area. Suppose the bulk of the content in a certain grade's social studies curriculum consists of some American history, some world history, and some economics. If we look at student test scores on some interim/benchmark assessments for this curriculum and we see increasing scores, how do we interpret that? Does it represent an equal increase in knowledge in all three of these sub-areas? Was growth consolidated in one area and the others held steady? Even strand sub-scores may not help in this area. Are the strands represented equally in the assessments? Are the strands of equal difficulty for students in that particular area? These questions need to have clear answers if we are going to be able interpret the scores in a way that is useful for assessing student growth.

Curriculum-based Assessments and Growth

If an interim/benchmark assessment that is curriculum based is used to evaluate student growth, a change in the number of standards mastered may be useful. This assumes things such as the same standards being assessed on both assessments, mastery having been defined in terms of the items on each assessment, and equivalency of difficulty of the strands on each assessment. Tracking this change in strands mastered can be straightforward for individual students; three of seven standards were mastered on the first assessment and five of seven were mastered on the second, a gain of two standards mastered. It may be problematic comparing the changes in numbers of standards mastered between students or even aggregating the numbers of standards mastered across groups of students. Number of standards mastered may not be an equal interval scale and a difference of two standards mastered may not represent equivalent learning. This is similar to the issue we saw previously where a change in the number of questions answered correctly within a strand might represent different amounts of learning depending on where in that "scale" the change occurred.

Scale-based Assessments and Growth

Scale-based I/B assessments are built so that their scales are equal interval, and you can use scores on the scale to do mathematical computations like subtraction for growth, or calculating a mean or median to summarize group achievement. Scale scores are developed for the purpose of being able to be used in mathematical computations. The interpretation of the results of these computations may not be clear, however. If we see that a student grew 17 score points on our scale, how do we interpret that value? It will depend on the range of the scale as well as information on how others have been located on that scale. Measurement scales like height and weight use familiar units which makes interpretation easy. Many academic assessments use scales that aren't nearly as familiar and as such, require more work and explanation in interpreting.

Clearly, choosing an assessment to assess academic growth requires clarity on how you think of growth and its measurement, how you want to use those scores, and a deep understanding of how the assessments under consideration were crafted.

Clearly, choosing an assessment to assess academic growth requires clarity on how you think of growth and its measurement, how you want to use those scores, and a deep understanding of how the assessments under consideration were crafted.

A Last Word About Validity

Interim/benchmark assessments can be used for a variety of useful purposes in education. Tracking student progress toward mastery of learning objectives, monitoring the effectiveness of instructional offerings, and measuring student growth are all things that can be accomplished with well-crafted I/B assessments in the proper context.

The success of any interim/benchmark assessment system requires the selection and use of an assessment that has been built for the desired purpose or purposes. Clarity in what types of results are needed and how they will be used is essential information that needs to be explicated before an assessment is chosen. It is often not possible to add on additional, valid, purposes and/or uses of an assessment after the fact.

The validity, or non-validity, of a test is **only** assessed in the context of its purpose or how the results are used. Tests are **not** valid or invalid, per se. *Rather, tests are deemed valid or invalid for a specific, stated purpose*. Validity is **not** a characteristic of the test itself. Validity is a characteristic of the use, and consequences, of an assessment.

Tests are not valid or invalid, per se. Rather, tests are deemed valid or invalid for a specific, stated purpose. Validity is not a characteristic of the test itself. Validity is a characteristic of the use, and consequences, of an assessment.

Evidence must be brought to bear in establishing an assessment's validity. A validity argument must be made rather than calculating some sort of validity statistic (which doesn't exist). If an assessment makes claims about verifying student mastery of content, evidence must be brought that establishes the assessment addresses that content in sufficient breadth and depth to provide adequate evidence of mastery. In addition, the administration conditions need to be verified to be consistent with how the assessment was designed to be administered and that the uses of the results are appropriate for the purpose of the assessment.

If an interim/benchmark assessment makes claims about measuring student growth, at least in the common understanding (e.g. height and weight), then information about the scale of the assessment must be understood. First and foremost, the scale of the assessment must be demonstrated to be equal interval. That is, a difference of "N" points represents the same change in achievement regardless of where it occurs along the scale. This requires substantial technical development, complex statistical models, and strong assumptions about the nature of the content. "Number correct" or "percent correct" typically do not yield an equal interval scale when interpreted as to how much additional knowledge is represented. (Does going from 20%

correct to 25% correct represent the same increase in achievement as going from 95% to 100% correct does?)

If an interim/benchmark assessment is going to be used for multiple purposes, a validity argument needs to be made *for each proposed use or purpose*. The fact that an assessment is demonstrated to be valid for one purpose does not mean that it is necessarily valid for any other use.

About the Author

Jim Gullen, Ph.D., is a retired educator who has worked for local and intermediate school districts and the Michigan Department of Education. He serves on the boards of the Michigan Assessment Consortium (MAC) and the Michigan Educational Research Association (MERA). He holds a Bachelor of Science degree in mathematics education, and M.A. and Ph.D. degrees in educational evaluation and research, all from Wayne State University.

Resources for further learning

The Michigan Assessment Consortium offers a variety of resources to help you learn more about interim/benchmark assessments and communicate your learning with others in your network. All these resources are available at www.MichiganAssessmentConsortium.org.

Learning Points (two-page handouts for posting or printing)
www.michiganassessmentconsortium.org/aln/aln-learning-points

- Start with Purpose When Choosing Assessments
- What do we mean by Interim/Benchmark Assessments?
- Interim Assessment: What are some key characteristics?
    - Companion Chart: Purposes for and Essential Characteristics of Interim Assessment

Professional Learning

- **MAC Learning Modules—available through Michigan Virtual**
  michiganassessmentconsortium.org/almodules
    - Module 3: Developing Appropriate Assessments
    - Module 4: Selecting Appropriate Assessments
    - Module 5: Developing a High Quality Balanced Assessment System
    - Module 6: Making Meaning from Student Assessments
    - Module 7: Understanding the Technical Concepts Used in Student Assessment
    - Module 8: Using Assessment Data Well

- **An Assessment to Every Purpose, Under Heaven—**a recorded workshop and resources by Marianne Perie, University of Kansas michiganassessmentconsortium.org/event/assessment-learning-network-2017-18-event-2/

- **Measuring Student Growth: So Much More than Subtracting Two Numbers—**a recorded workshop and resources by Jim Gullen michiganassessmentconsortium.org/event/assessment-learning-network-2017-18-event-3-2/

Appendix:
Purposes for and Essential Characteristics of Interim Assessment

Assessment Category: Achievement Monitoring (PAGE 1 of 2)	
Assessment Purpose	**Essential Characteristics**
Determine how well the student has learned the material to date	■ Sufficient alignment (both breadth and depth) with the curriculum to provide accurate information about how well students are mastering the content ■ Large range of standards assessed so that all students receive an estimate of how well they have mastered the content they have achieve
Provide aggregate information on student achievement at a school or district level	■ Sufficient alignment (both breadth and depth) with the curriculum to provide accurate information about how well students are mastering the content ■ Large range of standards assessed so that all students receive an estimate of how well they have mastered the content they have achieved
Provide specific feedback on where there are gaps in a particular student's knowledge	■ Sufficient alignment (both breadth and depth) with the curriculum to provide accurate information about how well students are mastering the content ■ Large range of standards assessed so that all students receive an estimate of how well they have mastered the content they have achieved ■ Items constructed such that incorrect or incomplete responses provide useful information on student misconceptions ■ Test reports designed to highlight what students know and don't know to help students adjust their learning strategies
Diagnose and provide corrective feedback to help a group of students get on track to succeed on the summative assessment	■ Strong statistical correlation between results on the interim assessment and the summative assessment ■ Sufficient alignment (both breadth and depth) with the curriculum to provide accurate information about how well students are mastering the content ■ Large range of standards assessed so that all students receive an estimate of how well they have mastered the content they have achieved. ■ Items constructed such that incorrect or incomplete responses provide useful information on student misconceptions ■ Test reports designed to highlight what students know and don't know to help students adjust their learning strategies
Motivate and provide feedback to students about their learning	■ Large range of standards assessed so that all students receive an estimate of how well they have mastered the content they have achieved ■ Items constructed such that incorrect or incomplete responses provide useful information on student misconceptions ■ Test reports designed to highlight what students know and don't know to help students adjust their learning strategies

Reference: Perie, M., Marion, S. & Gong, B. (2009). Moving toward a comprehensive assessment system: A framework for considering interim assessments. Educational measurement: Issues and practice, 28(3) pp. 5-13

Continued next page….

Assessment Category: Achievement Monitoring (PAGE 2 OF 2)

Assessment Purpose	Essential Characteristics
Ensure that teachers are staying on track in terms of teaching the curriculum in a timely manner (i.e., pacing)	■ Sufficient alignment (both breadth and depth) with the curriculum to provide accurate information about how well students are mastering the content ■ Test administration scheduled so that it follows closely the temporal requirements of the pacing guides
Provide a more thorough analysis of the depth of students' understanding	■ Sufficient alignment (both breadth and depth) with the curriculum to provide accurate information about how well students are mastering the content ■ Large range of standards assessed so that all students receive an estimate of how well they have mastered the content they have achieved ■ Items constructed such that incorrect or incomplete responses provide useful information on student misconceptions
Determine whether students are prepared to move on to the next instructional unit	■ Test constructed so that it focuses measurement on the prerequisite skills of the next unit that are contained in the current content

Assessment Category: Prediction

Assessment Purpose	Essential Characteristics
Predict students' performance on a summative assessment	■ Strong statistical correlation between results on the interim assessment and the summative assessment Note: A test might have strong statistical correlation with a summative assessment but not have face validity, e.g., it might not look like it measures the same thing as the summative assessment
Determine whether students are on track to succeed on the summative assessment	■ Strong statistical correlation between results on the interim assessment and the summative assessment ■ Sufficient alignment (both breadth and depth) with the curriculum to provide accurate information about how well students are mastering the content ■ Test items of varying difficulty so that all students can get an estimate of how well they're mastering the content

Assessment Category: Program Evaluation

Assessment Purpose	Essential Characteristics
Determine whether one pedagogical approach is more effective in teaching the material than another	■ Adequate alignment (both breadth and depth) with the content ■ The assessment must be equally sensitive to the instructional methods of both pedagogical approaches
Provide information to help the instructor better teach the new group of students by evaluating the instruction, curriculum, and pedagogy	■ Test items of varying difficulty so that all students can get an estimate of how well they're mastering the content ■ Items constructed such that incorrect or incompete responses provide useful information on student misconceptions

Reference: Perie, M., Marion, S. & Gong, B. (2009). Moving toward a comprehensive assessment system: A framework for considering interim assessments. Educational measurement: Issues and practice, 28(3) pp. 5-13