Invitational Research Symposium on
Through-Course
Summative Assessments

PICKING UP THE PIECES: AGGREGATING RESULTS FROM THROUGH-COURSE ASSESSMENTS

Lauress L. Wise

HumRRO

March 2011

Center for K – 12 Assessment
& Performance Management at ETS

Picking up the Pieces: Aggregating Results From Through-Course Assessments

Lauress L. Wise
HumRRO

Executive Summary

Both the SMARTER Balanced Assessment Consortium (SBAC) and the Partnership for Assessment of Readiness for College and Careers (PARCC) are developing assessments that will be used in many different states, and both are planning to implement systems of through-course assessments. Each of the consortia is designing assessments to be administered at different points of the school year and considering how to combine results across these through-course assessments into overall summative measures of individual student proficiency and growth. This paper explores alternative methods for aggregating through-course assessment results.

Simulations of different models for student learning and different methods of aggregating through-course assessment results illustrate several important concerns. For one thing, measurement error may limit uses and interpretations of individual student results. Also because of the likely magnitude of measurement error, giving students multiple opportunities to take the same test and then assigning the highest score is likely to seriously overstate student achievement levels. At the same time, simply adding up results from the different assessments is likely to significantly understate end-of-year achievement and growth if significant learning occurs on topics after the point at which they are tested.

Two methods are shown to provide good estimates of student proficiency and annual growth and to offer some advantages in comparison to end-of-year testing. For topics and skills that are taught and learned at a particular point in the school year, through-course assessments

matched to when particular topics are taught would support simple addition of results across topics. For topics and skills that are improved continually throughout the school year, a method involving projections to end-of-year proficiency would provide reasonable estimates.

The results presented in the full paper are meant to suggest issues that warrant more specific investigation. Research using forms of the actual assessments as they are developed is needed to check assumptions about models of student learning and the appropriateness of specific score aggregation methods. Research will also be needed on how through-course assessment results will be used, both for improving instruction and for accountability, and on the impact of through-course assessments on instructional practices.

Recommendations

The consortia are still in preliminary stages of designing through-course assessments and planning the way results from these assessments will be used. The analyses reported in this paper are intended to stimulate careful attention to how students learn during the year and suggest that uses of through-course assessments should be built around proven models of student learning. Several specific recommendations are offered to aid the consortia in consideration of these issues.

Recommendation 1

Be very cautious in promoting or supporting uses of individual student results. Even with highly reliable tests, there will be significant measurement error in estimates of student proficiency at any one time and in measure of growth relative to some prior point of assessment. Research, likely using a test-retest design, will be needed to demonstrate that within- and between-student differences are real and not just a result of measurement error.

Recommendation 2

Methods used for aggregating results from through-course assessments to estimate end-of-year proficiency or annual growth should be based on proven models of how students learn

the material that is being tested. Research, such as that outlined above, is needed to demonstrate relationships between time of instruction and student mastery of targeted knowledge and skills. As shown in this paper, mid-year results can significantly underestimate or, in some cases, overestimate end-of-year status and growth if the method for aggregation is not consistent with how students actually learn.

Recommendation 3

An end-of-unit testing model, with simple addition of results from each through-course assessment is appropriate if most or all student learning on topics covered by each assessment occurs in the period immediately preceding the assessment. Developers should also be clear whether the target is measuring maximal performance during the year or status and growth at the end of the full year of instruction.

Recommendation 4

A projection model, where results from each through-course assessment are used to predict end-of-year proficiency or growth is needed where student learning on topics covered by each assessment is continuous throughout the school year. For this approach, research will be needed to determine how to weight results from each assessment to provide the most accurate estimate of end-of-year proficiency and growth.

Recommendation 5

Short-term research is needed to monitor the different ways, some possibly unintended, that through-course assessment results are used. For example, the timing of instruction or of the assessments may be altered in a way that actually detracts from learning for some or all students. Materials and guidance will be needed to promote positive uses and eliminate uses and interpretations that might have negative consequences.

Recommendation 6

Longer-term research is needed to gauge the impact of through-course assessments on instruction and on improvements to student learning. Through-course assessments are part of a theory of action intended to lead to significantly increased levels of student proficiency and, by the end of high school, to readiness for college and careers. Specific assumptions of the theory of action should be checked as a step to establishing and improving the effectiveness of the assessments for achieving their intended ends.

Picking up the Pieces: Aggregating Results From Through-Course Assessments

Lauress L. Wise
HumRRO

Context

With the slow speed of our current economic recovery, Americans are being forced to confront the concrete reality of global competition for products, services, and most of all, jobs. Multinational companies are increasingly shifting jobs overseas to work forces that are not only less expensive but, according to the latest results from the Program for International Student Assessment (PISA) and Trends in International Mathematics and Science Study (TIMSS), also better educated. While many of us are still sleeping through this wake-up call, many are not. We have shifted with surprising rapidity from a K–12 system with state-by-state expectations that were often not tied to what students really need to be ready for college and work to an emerging consensus on a common set of high standards for student achievement that have been adopted by nearly all states.

Now we are engaged in two major efforts to develop common measures of student progress toward college and career readiness by the end of high school. These measures are essential to monitoring and evaluating progress in moving to the high level of achievement that students need and deserve. The measures will shine a bright light to help us identify programs and systems that are particularly effective and also those that are not. Many states have committed to using student achievement results from these new assessments in evaluating teachers as well as districts, schools, and programs.

The new assessments being developed by the Partnership for Assessment of Readiness for College and Careers (PARCC) and the SMARTER Balanced Assessment Consortium (SBAC)

will be aligned to the new common standards for student achievement. Both of the consortia also plan to introduce new features to improve the usefulness of assessment results for the wide variety of intended instructional and accountability purposes. Key among these new features is supplementing a single end-of-year assessment with a system of through-course assessments.

Descriptions of Through-Course Assessments

Details of the content and use of the through-course assessments have yet to be worked out. One model being considered by PARCC would include three quarterly assessments and a final comprehensive assessment. The first two quarterly assessments would each be administered in a single class period and would include one or two focused tasks designed to assess a small number of key standards or competencies. The third quarter would be administered over several class periods and would be designed to measure skills not easily assessed with multiple choice or short answer questions. Presumably, some weighted combination of scores from the final and each of the quarterly (through-course) assessments would be used in assessing each student's level of proficiency.

Another model being considered by SBAC would divide the material covered by the end-of-course assessment into three or four parts. An adaptive assessment including perhaps 20 to 40 machine-scored multiple choice or short answer items and possibly one or two tasks that could not be immediately scored would be developed for each part. Schools could decide when to administer each part, and opportunities may be available for students to retest. Proficient performance on each part could then be used as an alternative to evidence from the full-year assessment.

The design and use of through-course assessments requires answering two key questions. The first is how to decide what content to cover in each of the different assessments. Will each assessment cover a different part of the curriculum? Or might the assessments be somewhat cumulative, with each one covering a new piece of the curriculum and also covering

the content included in assessments administered earlier in the year? Or will the assessments be essentially parallel forms covering the entire set of targeted content standards? There may be concerns that the sequencing of material to be tested will essentially force a common curriculum, a step many states may not be ready to take. On the other hand, there may be concerns that a better articulated model of within-year learning is exactly what is needed for significantly increasing student learning.

The second question is how results from each of the through-course assessments will be combined to give an overall measure of the status and growth of individual students as well as of classes and schools of students. Will results from assessments administered later in the year count more heavily? If so, how will the relative weights of assessment results be determined? The main idea of this paper is that methods for aggregating results from throughout the school year must be based on validated models of how students learn the content covered by these tests.

Types and Uses of Through-Course Measures

A key tension in the design and use of the common assessment systems is the many different ways in which we expect the results to be used. Three quite different uses are described here. Each is important, but each may place different demands on the design of the assessments, particularly the summative uses of the through-course components.

Status measures. Most current state assessments are designed to answer the basic question of whether students are performing at expected levels. Status measures are needed to answer key policy questions such as whether our overall investment in education is sufficient or whether programs and instruction in particular schools are good enough. Note, however, that status measures do not provide direct information on the source of student learning. Students may have already mastered most or all of the required skills in prior years or significant learning may be taking place outside of the classroom. Thus, status measures are not ideal for comparing the effectiveness schools, programs, and even teachers that serve different populations of students.

Growth measures. Fairness in accountability requires recognition of the fact that students vary in levels of prior learning. Schools and teachers cannot be accountable for prior deficits in learning and should not be given excessive credit for advanced learning prior to coming to the school or classroom. Growth measures are needed to assess how much students have learned during the year. A key question for through-course assessments is whether and how prior-year achievement levels will be taken into account in interpreting results from each of the current-year through-course assessments.

Note too that not all learning occurs in the classroom. Assessments cannot easily differentiate between learning that occurs as a result of classroom instruction and learning that comes from experiences outside the classroom. There is considerable debate about the extent to which schools and teachers should be accountable for or credited with learning that occurs outside of the classroom, although many do feel that schools should be responsible for promoting and building on learning that occurs in other venues.

Diagnostic measures. Assessments require a significant investment of time and effort, both on the part of the students who take them and the teachers and other school officials who administer them. Through-course assessments are likely to increase time requirements for taking and administering the assessments. It is reasonable to expect that instructionally useful information about individual students will be provided as a return on this investment. Most commonly, we expect some information on which standards the student has or has not met, or at least on relative strengths or weaknesses across different areas of the curriculum. Current end-of-year assessments often include subscores that are neither normed nor standards-based and are also not very reliable. To the extent that through-course assessments cover different and more targeted portions of the curriculum, they have the potential for providing more reliable measures of mastery of each of these different parts than is currently the case with a single end-of-year assessment.

Potential Advantages of Through-Course Assessment Systems

The systems of through-course assessments being considered by both the PARCC and the SBAC offer two key advantages over current end-of-year assessments in meeting the multiple goals and uses demanded of assessment results. First, the increased testing time will almost surely lead to more reliable information about the status and growth of individual students. As noted below, assessment results for individual students typically contain a margin of error that is large (e.g., one third of a standard deviation). If testing time were increased by a factor of four, we would expect the standard error of individual student measures aggregated across the different assessments to be cut in half.

The second advantage offered by through-course assessments is that they can provide more timely data, allowing diagnostic information to be used before students move on to the next grade or class. Testing right after instruction in particular topics or skills could help to identify deficits that need remediation prior to moving on to more advanced topics or skills.

Concerns With Through-Course Assessment Systems

Apart from general concerns with too much testing, there are several more specific concerns about the use of through-course assessments as part of summative measures used in accountability. One concern is that testing earlier in the year will understate the effectiveness of a full year of instruction. Some topics may not yet have been taught and mastery of topics that have been taught may be further increased through reinforcing activities.

Another concern with summative uses of through-course assessments is that they may create too much pressure to follow a prescribed ordering of the curriculum and reduce opportunities for trying out and evaluating different ways of teaching essential skills. A related concern is that the prescribed order may not work best for all students, creating tensions between maximizing accountability scores and doing what is best for particular students.

General Methods for Aggregation of Results From Through-Course Assessments

Content experts will debate what topics are best covered by through-course assessments administered at different times during the school year. The focus of this paper is on how results from the different through-course assessments might be combined into an overall summative measure. Wise (2010) presented several models for aggregating through-course assessment results to yield overall summative measures. Several of these models, believed to be under consideration by one or both of the consortia, are described here.

Multiple Opportunities to Test

The first approach to through-course assessment is simply to allow students to take a full form of the same test at several points during the year. The student is assigned the highest score earned across these multiple opportunities. This approach does provide early indications of student strengths and weaknesses and an opportunity to track progress through the year. It also supports tracking progress for students who learn at different rates in comparison to an approach that tests different topics at specific times of the year. It does not, however, offer increased reliability over a single assessment. If anything, taking the highest of several scores increases the likelihood of a positive measurement bias.

End-of-unit model. A second model for aggregation is to treat each of the through-course assessments as assessing status or growth over one or more discrete units of instruction. An appropriate summative measure for the year or course as a whole is obtained by simply adding scores across the different end-of-unit tests as if they were different sections of the same test. This approach offers increased reliability in comparison to a single end-of-year assessment covering the full range of instruction for the year. It is also possible that students will demonstrate higher levels of proficiency on material that has just been taught in comparison to results from assessments later in the year.

Skill-growth model. In some cases, instruction may be viewed as focused on development and enhancement of a set of complex skills that are taught continuously

throughout the year and, in most cases, across years as well. Reading comprehension may be a good example. This same skill is assessed across a number of years, using texts of increasing complexity and requiring increasingly sophisticated analyses of these texts. Assessment of mastery of these skills throughout the year could be diagnostically useful. The use of mid-year assessment results in forming an overall summative measure is less clear. One approach is to use each mid-year result to predict end-of-year status and then weight results from each through-course assessment according to how accurately end-of-year status is predicted. In a simple linear example, *growth* (current score level minus prior-year score level) halfway through the year could be doubled to predict full-year growth. This prediction would then be weighted more than predictions from the first quarter but less than predictions from the third or final quarter.

Hybrid aggregation model. A more sophisticated aggregation model involves the use of subscores for different skills or areas of knowledge. Scores covering discrete areas of knowledge could be summed across assessments following an end-of-unit model. Scores covering more complex skills could be aggregated as weighted predictions of end-of-year status as in the skill-growth models. A hybrid model would likely be needed to cover a mathematics curriculum that included both discrete concepts taught in separate units and also more complex skills, such as problem solving or mathematical reasoning, that are taught throughout the year.

Does the Aggregation Model Matter?

The primary results reported here address the question of whether the choice of an aggregation model really matters. The approach taken was to (a) simulate individual student growth under alternative models for student learning, (b) simulate end-of-quarter test scores for individual students under each learning model, and then (c) examine the accuracy with which the summative scores from the different methods for aggregating the quarterly assessment results estimate the simulated values for true growth under each learning model.

Simulated Models of Student Learning

A key point of this paper is that we need a deeper understanding of how students learn before we can evaluate alternative ways of assessing that learning. Mathematical models of how students learn are not new. Atkinson, Bower, and Crothers (1965) provided examples of models for several types of learning. The simulations reported here examined four different models for student learning during the year. While empirical evidence has yet to be gathered regarding the degree to which these models match the learning of the common core skills measured by the new assessments, there is good reason to believe that each model matches the learning of some topics or skills and not others. The four models are described as follows:

One-time learning. This model assumes that there is little or no learning for a topic until it is taught and then there is no further learning after the topic has been mastered. Under this model, average student growth is one grade level in the quarter in which the topic is taught and zero in the preceding and following quarters. It is further assumed that about one fourth of the topics to be mastered are taught each quarter.

One-time learning with forgetting. This model assumes that students master a topic in the quarter in which it is taught, but there is some probability that mastery is lost through forgetting in a subsequent quarter. For illustration, we assume that students gain an average of 1.15 grade levels in the quarter in which the topic is taught but decline an average of .1 grade levels in each subsequent quarter. Thus the average annual gain for a topic taught in the first quarter is 1.15 - .1 - .1 -.1 = .85 grade levels, while the gain for a topic taught in the fourth quarter is 1.15 grade levels. These gain and loss values lead to an expected gain of 1.0 grade levels when averaging across topics taught in each of the four quarters.

One-time learning with reinforcement. This model assumes that students gain initial mastery of a topic in the quarter in which it is taught and then mastery improves a bit more in each following quarter as the topic or skill is reinforced by subsequent instruction. For illustration, we assume that students gain an average of .85 grade levels in the quarter in which

the topic is taught and increase .1 grade levels in each subsequent quarter. Thus the average annual gain for a topic taught in the first quarter is .85 + .1 + .1 +.1 = 1.15 grade levels, while the gain for a topic taught in the fourth quarter is just .85 grade levels. These gain and loss values lead to an expected gain of 1.0 grade levels when averaging across topics taught in each of the four quarters.

Continuous learning. Under this model, student learning of a topic or skill proceeds at a relatively even pace throughout the school year. This model is most plausible for complex skills that are practiced throughout the year or broad areas of knowledge (e.g., vocabulary in early grades) that are learned a little at a time over the year. In the simulations, it is assumed that average student growth is .25 grade levels each quarter of the school year.

Distribution of Simulated Growth Under Each Learning Model

We generated simulated quarterly and annual growth values for 400,000 students under each of the four learning models. Table 1 shows the means and standard deviations of simulated *true growth scores* under each learning model. The growth values are in annual growth units, with 1.0 representing typical (or expected) annual growth. The standard deviation of cumulative growth for the year was set to .61 which, with a normal distribution of growth scores, means that about five percent of the students would actually have negative growth for the year. Empirical data are needed to provide more precise fits to growth distributions under each of the learning models. As shown in Table 1, simulated growth means and standard deviations met the same overall target for the year, averaged across material taught in different quarters.

Table 1. Mean and Standard Deviations of Simulated Growth Scores

Learning model	Quarter content is taught	Simulated cumulative growth at the end of each quarter: 1st quarter Mean	1st quarter SD	2nd quarter Mean	2nd quarter SD	3rd quarter Mean	3rd quarter SD	4th quarter Mean	4th quarter SD
One-time learning	1st	1.00	0.53	1.00	0.56	1.00	0.58	1.00	0.61
	2nd	0.00	0.17	1.00	0.56	1.00	0.58	1.00	0.61
	3rd	0.00	0.17	0.00	0.24	1.00	0.58	1.00	0.61
	4th	0.00	0.17	0.00	0.24	0.00	0.30	1.00	0.61
	Average	**0.25**	**0.53**	**0.50**	**0.66**	**0.75**	**0.68**	**1.00**	**0.61**
One-time learning with forgetting	1st	1.15	0.52	1.05	0.54	0.95	0.57	0.85	0.60
	2nd	0.00	0.17	1.15	0.54	1.05	0.57	0.95	0.60
	3rd	0.00	0.17	0.00	0.24	1.15	0.57	1.05	0.60
	4th	0.00	0.17	0.00	0.25	0.00	0.30	1.15	0.60
	Average	**0.29**	**0.58**	**0.55**	**0.69**	**0.79**	**0.69**	**1.00**	**0.61**
One-time learning with reinforcement	1st	0.85	0.52	0.95	0.54	1.05	0.57	1.15	0.60
	2nd	0.00	0.17	0.85	0.54	0.95	0.57	1.05	0.60
	3rd	0.00	0.17	0.00	0.24	0.85	0.57	0.95	0.60
	4th	0.00	0.17	0.00	0.24	0.00	0.30	0.85	0.60
	Average	**0.21**	**0.48**	**0.45**	**0.62**	**0.71**	**0.66**	**1.00**	**0.61**
Continuous Learning	**All**	**0.25**	**0.30**	**0.50**	**0.43**	**0.75**	**0.53**	**1.00**	**0.61**

An Ugly Truth About the Measurement of Growth

The measurement of change from one time to the next is problematic (Harris, 1963). Even with highly reliable measures at each point in time, considerable measurement error in difference scores is likely to occur (Webster & Bereiter, 1963). Table 2 shows the standard error of measurement (in standard deviation units) for each of two tests as a function of the reliability of these tests. Standard errors are also shown for differences between scores on these two tests (growth) and for average differences assuming a class size of 30 or a school size of 300.

As shown in Table 2, even with highly reliable tests (coefficient alpha = .95) at each point in time the measurement error of an individual growth score is about one-third of a standard deviation. Wu (2010) recently reported average annual student growth rates ranging from .3 to .5 standard deviations. Thus, average growth is not much bigger than the standard error of the growth measure, even with highly reliable measures and confidence bounds for students with average growth that would include both no growth at all and double the average growth.

When we consider average growth for a classroom or school, our ability to distinguish average growth from no growth is much better. The consortia intend many different uses for growth measures generated from the new assessments. Some uses, such as evaluating programs, schools, or possibly even individual teachers based on average growth for moderate to large sample of students, should be easy to support. Other uses, such as reporting individual student progress to students and their parents or taking different actions based on individual student growth measures will be much more difficult to support given the likely uncertainty in individual growth scores.

Sophisticated statistical models for measuring change have been proposed (Lord, 1963; Meredith, 1991). A simple score difference model is used here to illustrate the impact of different methods of aggregation on estimates of growth. Other models (e.g., regression-based models) are possible but lack transparency and have not been shown to greatly improve accuracy.

Table 2. Standard Error of Measurement of Growth Scores in Standard Deviation Units as a Function of the Reliability of the Measures at Each Time

	Standard error of measurement			
Reliability	Each test	Average growth		
		N = 1	*N* = 30	*N* = 300
0.95	0.22	0.32	0.06	0.02
0.90	0.32	0.45	0.08	0.03
0.85	0.39	0.55	0.10	0.03
0.80	0.45	0.63	0.12	0.04
0.75	0.50	0.71	0.13	0.04

The two consortia are considering somewhat different through-course measures. SBAC proposes mostly machine-scored questions administered adaptively to increase accuracy throughout the score range. PARCC is considering assessments that include a small number of tasks each. Prior research (Shavelson, Baxter, & Gao, 1993) has shown significant student-by-task interactions for performance-type assessments, suggesting that results based on a small number of tasks might vary considerably as a function of the tasks selected. The difference between the two approaches illustrates a classic reliability-validity tradeoff. The choice is between measuring with great accuracy something that is not quite the high order skill we intend versus measuring the targeted skills, but with less accuracy. As with most tradeoffs, a balance is needed.

Simulating Measures to Estimate Growth

The main focus of these simulations is alternatives for estimating annual growth. Results are expressed in units where average (or expected) annual growth is 1.0 with a standard deviation of .61. An effect size of .33 is assumed for average annual growth, meaning that the

standard deviation of prior year scores, against which growth is measured, is about 3.0 annual growth units. We assumed a measurement reliability of .95 for end-of-year tests given in the prior and current year. This translated into a standard error of measurement of .67 growth units for prior year scores. By the end of the current year, the standard deviation of student scores had increased to 3.35 and the standard error of measurement became .75. With .95 reliabilities for each test and assuming uncorrelated measurement errors, the standard error of the growth scores (difference between prior and current end-of-year scores) is 1.00 growth units.

As an alternative to a single end-of-year test, we modeled four quarterly tests. We assumed these tests might not be quite as long as an end-of-year test and so simulated the tests to have a reliability of .90, which translated into a standard error of the estimate of growth (quarterly score minus prior year score) of 1.20.

For both the end-of-year tests and the quarterly tests, we simulated estimated or observed growth scores by adding a normally distributed random variable to the true simulated cumulative growth scores generated for each learning model as described above. The standard deviations of the random errors were equal to the measurement error just described (1.0 for the end-of-year test and 1.2 for the quarterly tests).

We looked at four ways of combining the quarterly test scores and compared the resulting composite to results from a single end-of-year assessment. The four aggregation models were as follows:

1. *Simple average:* We simulated averaging four estimated growth scores that either covered the entire annual content (regardless of when it was taught) or random samples of content that were not aligned to when the material was taught.
2. *Maximum score:* We took the highest of the four scores, again modeling the situation whether either the entire content was covered each time or subsets of

content covered by each assessment were not related to when the material was taught.

3. *Matched score:* For each of the one-time learning models, we simulated the situation where the content of each quarterly test matched what was taught in that quarter. This is the true end-of-unit model for aggregation.
4. *Projected scores:* We converted each quarterly score to an estimate of annual growth by multiplying the first quarter score by 4, the second quarter score by 2, the third quarter score by 1.33, and the fourth quarter score by 1.0. We then weighted the four resulting estimates to approximate regression weights for optimal prediction of the true annual growth score. The resulting weights were 1, 4, 9, and 17 for the four projected quarterly scores. Note that the combination of projection and estimation weights results in effective weights of 4, 8, 12, and 17, which is nearly proportional to the amount of instruction time prior to assessment.

After computing each of the four composites for the simulated students under each of the four learning models, we computed two measures of error of estimation. The first was the error in estimating the simulated true annual growth value. The other was the difference between the maximum of the quarterly cumulative growth scores and the composite. This second measure was intended to reflect the belief of some that students should be given credit for learning something, even if they later forgot it. The end-of-unit measures are specifically designed to be a better measure of what students knew immediately after instruction in a topic or skill.

Table 3 shows the mean and standard deviation of the estimation errors for each of the aggregation methods under each of the four learning models. Several important conclusions may be drawn from these simulated results:

1. *The end-of-year assessment model performed as expected* with average errors of 0.0 (no bias) and error standard deviations of 1.0 under each of the four learning

models. The maximum (during the year) growth scores are slightly underestimated by the end-of-year scores, particularly for the one-time learning with forgetting model.

2. *Simple averaging significantly underestimates annual growth*. Unless test content is closely aligned with when material is taught, early estimates of growth are much lower than eventual annual growth. In these simulations, annual growth is underestimated by more than a third (.37) for the continuous learning model and by nearly a half (.45) for the one-time learning models. The standard deviation of the estimation errors was somewhat smaller compared to the end-of-year assessments (roughly .7 compared to 1.0), but that advantage disappeared when the mean bias was added in.
3. *Taking the maximum across quarterly scores very seriously overestimates actual growth*. Estimated growth with this aggregation method is nearly double actual growth (1.9 compared to 1.0). The maximum score method also overestimates the maximum cumulative quarterly growth by nearly as much.
4. *The matched score method (end-of-unit tests) works quite well under each of the one-time learning models*. There was no mean bias and the standard deviation of the errors was less than .8 compared to 1.0 for the end-of-year tests. As expected, the matched score method, which involves testing at the end of each quarterly unit, does a better job of estimating the maximum cumulative quarterly growth values compared to end-of-year testing. Here, too, there is essentially no bias. In addition, the error standard deviations are just under .8 compared to end-of-year values of 1.0.
5. *The projected score method provides estimates that are slightly better than estimates from the end-of-year test.* It is the only other method that produces unbiased estimates of annual growth under the continuous learning model. The

standard deviation of the estimates is .9 compared to 1.0 for the end-of-year model, demonstrating a small return on the investment of additional testing time.

Table 3. Mean and Standard Deviations of Estimation Errors for Each Aggregation Model Under Each Learning Model

Aggregation method	Learning model							
	Continuous learning		One-time learning		One-time learning with forgetting		One-time learning with reinforcement	
	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	SD
	Error in estimating end-of-year growth							
End-of-year test	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00
Average score	-0.37	0.66	-0.45	0.74	-0.45	0.78	-0.45	0.70
Maximum score	0.94	0.90	0.91	0.94	0.94	0.97	0.89	0.91
Matched score	n/a	n/a	0.00	0.79	0.00	0.78	0.00	0.79
Projected score	0.00	0.90	-0.09	0.96	-0.08	0.98	-0.10	0.93
	Error in estimating maximum of cumulative quarterly growth							
End-of-year test	-0.05	1.01	-0.08	1.01	-0.15	1.03	-0.04	1.01
Average score	-0.43	0.64	-0.53	0.71	-0.60	0.71	-0.49	0.68
Maximum score	0.89	0.88	0.84	0.92	0.79	0.93	0.85	0.90
Matched score	n/a	n/a	-0.01	0.78	-0.01	0.78	-0.02	0.77
Projected score	-0.05	0.89	-0.17	0.94	-0.23	0.94	-0.14	0.93

Further Research Needs

A great many details remain to be specified about how through-course assessments will be designed, developed, implemented, and used. Since there are no current examples of how such systems might function efficiently and effectively, further research is needed. Some ideas for research on design and use of through-course assessments are described here.

Research on Assessment Design

Two studies to help in designing through-course assessments are suggested. It is highly likely that both of the consortia are already engaged in some form of this research. The emphasis here is on achieving a better understanding of how and when content targeted for a specific grade is taught as a means of identifying the most appropriate ways to aggregate scores from the through-course assessments. An initial more qualitative study should be followed by an empirical study using developmental forms of the new assessments.

Research on test content. A first key area of research concerns how best to organize the assessment of mastery of the content standards assigned to a particular grade or course. The research would involve examining existing curricula and asking experts to walk back the standards to the points at which they are taught. To support appropriate aggregation, it will be important for experts to distinguish between topics or skills that are taught at particular points in the curriculum and topics or skills that are learned and practiced more or less continuously throughout the year. Simple aggregation of end-of-unit assessments would be appropriate for the former topics and skills while projection estimates may be needed for the latter.

A related area for research concerns the development and refinement of within-year learning progressions. Larger-scale projections are implied by the grade-by-grade content standards that lead up to readiness for college and careers by the end of high school. Through-course assessments must be designed around models of more micro-level, within-year learning progressions. Existing research on the effectiveness of different instructional sequencing should be reviewed and new research added to fill in our understanding of effective sequencing.

Through-course assessments are likely to drive instructional sequencing decisions, and it is important that resulting changes lead to improved effectiveness.

Research on learning models and aggregation methods. After initial through-course modules are designed, empirical research is needed to calibrate and validate models of learning that will determine methods of aggregation. This research will involve administering each through-course assessment at different times. Most specifically, administering some tests immediately after instruction and also at the end of the year will provide data on the degree to which learning for a topic or skill continues to improve or possibly decline after initial instruction. If performance continues to improve, results from earlier assessments will need to be adjusted to provide a better assessment of end-of-year status. Adjustments might also be appropriate if performance declines after initial instruction depending on whether the target is end-of-year rather than maximal performance.

The consortia each involve a large number of states, most of whom will be eager to try out the new assessments. It should be possible to administer different forms of each through-course assessment at different times of the year and track how performance varies by time and how this relationship varies across different state curricula. The key question for topics that are taught at a particular time is how much additional learning or forgetting occurs between the time the topic is taught and the end of the year. The key question for topics or skills that are taught throughout the year is how well does performance at each point in time predict (project onto) end-of-year performance on this skill or topic. Answers to these questions can be used to check and calibrate a specific learning model, which will, in turn, indicate the most appropriate method of aggregating scores.

Research on Assessment Use and Impact

As the through-course assessments are developed, it will be important to conduct research on how results from these assessments will be used and the impact of these uses on

curriculum and instruction. Steps may be required to avoid inappropriate uses or interpretations of test results and to avoid unintended negative consequences.

Research on use of through-course assessments. It may be that states, districts, or schools will be given some flexibility as to when to administer each available through-course assessment. This will likely be the case if significant differences are evident in instructional sequencing across participating districts and states and if beliefs about particular sequencing strategies are firmly held. In this case, it will be important to conduct an operational tryout, monitor decisions about when each assessment will be administered, and survey decision-makers to identify key reasons for choosing earlier or later administration dates. Note that if administration dates vary significantly, it may be necessary to adjust projections to end-of-year proficiency as a function of administration date. This adjustment would be necessary to maintain unbiased estimates of annual growth and also to be sure that decisions about administration dates would not be influenced by perceived advantage. For example, schools might assume that they would get higher summative scores if they tested their students as late in the year as possible.

Another important area of research concerns how scores from each through-course assessment are used. Districts, schools, and teachers should be surveyed to determine the extent to which they are using score results to evaluate curriculum or programs, as part of teacher evaluation, or to modify instruction for individual students. It will be important to see that uses of test results reflect an appropriate appreciation of measurement error. Results from research on test score use would be used to develop or improve training and information materials that describe strengths and limitations of different possible uses of the test scores.

Research on the impact of through-course assessments. Over a more extended period, it will be important to observe changes in curriculum and pedagogy that are attributed to results from through-course assessments. Qualitative research will be needed to identify the nature and reasons for instructional changes. This research should be followed by a more

quantitative analysis of the extent to which these changes lead to improved student achievement, both in the current grade or course and also in subsequent grades or courses.

Summary

Both the SBAC and the PARCC are developing assessments that will be used in many different states. Both consortia are planning to implement systems of through-course assessments, assessments administered at different points of the school year. Consideration is being given as to how to combine results across these through-course assessments into an overall summative measure of individual student achievement and growth. This paper explored alternative methods for aggregating through-course assessment results.

Simulations of different models for student learning and different methods of aggregating through-course assessment results illustrated several important concerns. For one thing, giving students multiple opportunities to take the same test and then assigning the highest score without accounting for measurement error is likely to seriously overstate student achievement levels. At the same time, simply adding up results from the different assessments is likely to significantly understate end-of-year achievement and growth if significant learning occurs on topics after the point at which they are tested.

Two methods were shown to provide good estimates of student status and annual growth and to offer some advantages in comparison to end-of-year testing. For topics and skills that are taught and learned at a particular point in the school year, end-of-unit testing would support effective aggregation of results across topics. For topics and skills that are improved continually throughout the school year, a method involving projections to end-of-year status would provide reasonable estimates.

The results presented here are meant to be suggestive. Research using forms of the actual assessments as they are developed is needed to check assumptions about models of student learning and the appropriateness of specific score aggregation methods. Research will also be needed on how through-course assessment results will be used, both for improving

instruction and for accountability, and on the impact of through-course assessments on instructional practices.

Recommendations

The consortia are still in preliminary stages of designing through-course assessments and planning the way results from these assessments will be used. The analyses reported here are intended to stimulate careful attention to how students learn during the year and suggest that uses of through-course assessments should be built around proven models of student learning. Several specific recommendations are offered to aid the consortia in consideration of these issues.

Recommendation 1

Be very cautious in promoting or supporting uses of individual student results. Even with highly reliable tests, there will be significant measurement error in estimates of student proficiency at any one time and in measure of growth relative to some prior point of assessment. Research, likely using a test-retest design, will be needed to demonstrate that within- and between-student differences are real and not just a result of measurement error.

Recommendation 2

Methods used for aggregating results from through-course assessments to estimate end-of-year proficiency or annual growth should be based on proven models of how students learn the material that is being tested. Research, such as that outlined above, is needed to demonstrate relationships between time of instruction and student mastery of targeted knowledge and skills. As shown in this paper, mid-year results can significantly underestimate or, in some cases, overestimate end-of-year status and growth if the method for aggregation is not consistent with how students actually learn.

Recommendation 3

An end-of-unit testing model, with simple addition of results from each through-course assessment is appropriate if most or all student learning on topics covered by each assessment occurs in the period immediately preceding the assessment. Developers should also be clear whether the target is measuring maximal performance during the year or status and growth at the end of the full year of instruction.

Recommendation 4

A projection model, where results from each through-course assessment are used to predict end-of-year proficiency or growth is needed where student learning on topics covered by each assessment is continuous throughout the school year. For this approach, research will be needed to determine how to weight results from each assessment to provide the most accurate estimate of end-of-year proficiency and growth.

Recommendation 5

Short-term research is needed to monitor the different ways, some possibly unintended, that through-course assessment results are used. For example, the timing of instruction or of the assessments may be altered in a way that actually detracts from learning for some or all students. Materials and guidance will be needed to promote positive uses and eliminate uses and interpretations that might have negative consequences.

Recommendation 6

Longer-term research is needed to gauge the impact of through-course assessments on instruction and on improvements to student learning. Through-course assessments are part of a theory of action intended to lead to significantly increased levels of student proficiency and, by the end of high school, to readiness for college and careers. Specific assumptions of the theory of action should be checked as a step to establishing and improving the effectiveness of the assessments for achieving their intended ends.

References

Atkinson, R. C., Bower, G. H., & Crothers, E. J. (1965). *An introduction to mathematical learning theory*. New York, NY: John Wiley & Sons.

Harris, C. W. (1963). *Problems in measuring change*. Madison: University of Wisconsin Press.

Lord, F. M. (1963). Elementary models for measuring change. In C. Harris (Ed.), *Problems in measuring change* (pp. 21–38). Madison: University of Wisconsin Press.

Meredith, W. (1991). Latent variable models for studying differences and change. In L. M. Collins & J. L. Horn (Eds.), *Best methods for the analysis of change* (pp. 164–169). Washington, DC: American Psychological Association.

Shavelson, R. J., Baxter, G. P., & Gao, X. (1993). Sampling variability of performance assessments. *Journal of Educational Measurement*, 30, 215–232

Webster, H., & Bereiter, C. (1963). The reliability of changes measured by mental test scores. In C. Harris (Ed.), *Problems in measuring change* (pp. 39–59). Madison: University of Wisconsin Press.

Wise, L. L. (2010, April). *Aggregating summative information from different sources*. Paper presented at the National Research Council workshop on best practices for state assessment systems, Washington, DC.

Wu, M. L. (2010). Measurement, sampling, and equating errors in large-scale assessments. *Educational Measurement: Issues and Practice, 29*, 15–27.