

ANALYSIS OF ASSESSMENTS PILOTED FOR USE IN EDUCATOR EVALUATION

**Prepared for the DC Office of the State Superintendent of Education
March 2013**

TABLE OF CONTENTS

TABLE OF CONTENTS.....	2
I. REPORT OVERVIEW	3
II. PROJECT INTRODUCTION	4
III. MATRIX OF ASSESSMENTS PILOTED BY LEAS.....	6
IV. A GUIDING FRAMEWORK FOR SELECTING STUDENT ASSESSMENTS FOR USE IN TEACHER EVALUATIONS.....	13
V. RECOMMENDATIONS TO SUPPORT THE DEVELOPMENT OF STUDENT GROWTH MODELS.....	19
VI. SUMMARY	21
VII. GLOSSARY	22

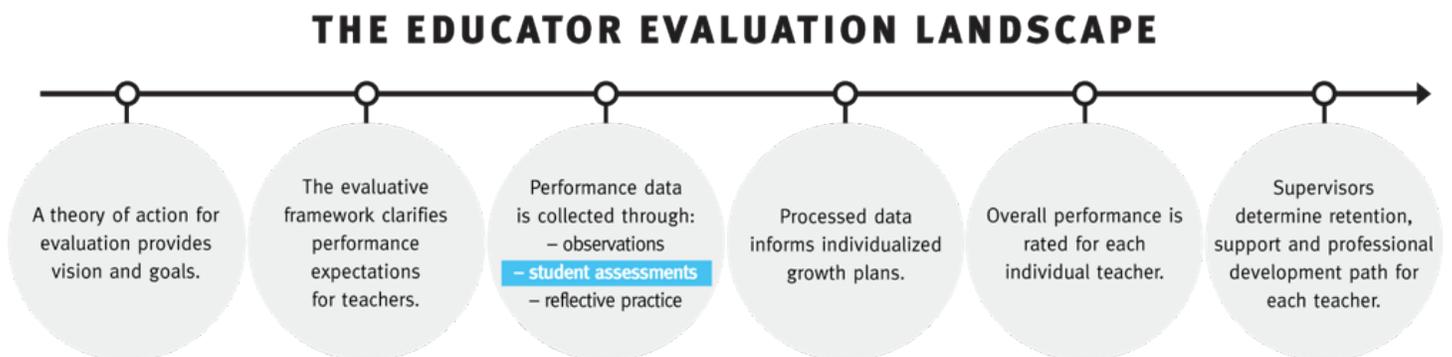
I. REPORT OVERVIEW

Teacher evaluation systems across the country are being revamped or created anew to provide school leaders and teachers with a more comprehensive view of teacher performance. The data from these systems are used to better inform decisions around professional development, retention, and promotion.

UPD has developed the accompanying report for the Office of the State Superintendent of Education (OSSE), and for all Race to the Top local education agencies (LEAs) piloting assessments in non-tested grades and subjects, in order to assist OSSE in providing technical assistance to LEAs as they incorporate student assessment data within teacher evaluations. The report provides summary technical information about all assessments piloted by Race to the Top LEAs across DC in select grades and subjects not tested by the DC-CAS. The objective of this report is to provide both OSSE and LEAs with meaningful and actionable feedback on the suitability of the assessments selected for use in informing educator evaluation systems.

In order to frame the pages that follow, it is important to understand that student assessment data is just one of several pieces of evidence collected and used to inform educator evaluation and support plans. The process of selecting student assessments whose data is appropriate to collect and use as an indicator for educator evaluation is one that requires careful thought to ensure alignment and suitability. Looking at the technical aspects of an assessment, which is what we examine in this report, is only one piece of the bigger puzzle of selecting the most appropriate student assessments for use in educator evaluation systems. Figure 1, below, illustrates where the collection of student assessment data fits in to the broader landscape of educator evaluation.

Figure 1.



II. PROJECT INTRODUCTION

Overview

DC OSSE has contracted with UPD Consulting to review the results from the Race to the Top (RTTT) pilot assessments and offer technical assistance to LEAs in using assessments in evaluation systems. All LEAs participating in RTTT were required to pilot an assessment in at least one grade and subject not currently tested by the DC-CAS. LEAs will ultimately incorporate the results from these assessments into their educator evaluation systems as the student achievement component.

In August 2011, OSSE provided LEAs with technical guidance on assessments that encouraged LEAs to select reliable, valid, and bias-free assessments of high technical quality. During the summer of 2012, UPD analyzed the assessments piloted by LEAs and gathered the documented reliability and validity statistics on these assessments.

UPD also inventoried the evaluation systems used by each participating LEA and analyzed the suitability of the chosen assessments for use in educator evaluation systems. Additionally, UPD provided a detailed item – level analysis for LEAs who provided UPD with student-level assessment results.

Objectives

The objectives of this report are:

1. To provide basic analysis on the assessments, including
 - a. Description of the LEA's evaluation model and chosen assessments;
 - b. Summary information about all assessments used to measure student achievement, including their documented reliability and validity statistics; and
 - c. Item-level analysis for LEAs providing summary level student results.

2. To provide OSSE with meaningful and actionable feedback on the suitability of assessments LEAs have chosen for use in their educator evaluation system.

Methodology

To find third-party evidence of technical adequacy of the assessments, we retrieved reliability and validity information from the following sources: (1) The Measurements Yearbook, (2) Tests in Print VIII, (3) Tests in Print Test Publisher Catalogue, (4) Educational and Psychological Measurement, or (5) Internet (e.g., ERIC Test Locator). We then looked at technical manuals from the publishing companies, and consulted with assessment experts from the Center for Assessments and Tembo Consulting.



**ANALYSIS ON ASSESSMENTS PILOTED FOR USE IN EDUCATOR
EVALUATION SYSTEMS**

III. MATRIX OF ASSESSMENTS PILOTED BY LEAS

How to read this section: *The following section contains a summary matrix of the technical information for all assessments piloted by RTTT LEAs in grades and subjects not tested by the DC-CAS.*

Basic summary information includes: grades/ages tested; subjects/domains tested; type of test; how test is administered; and time for test (in minutes).

Reliability information includes the types of reliability tested, as referenced by either third-party research or in the publisher's technical manual. To enable LEAs to see all reliability tests performed – consistent with the technical guidance provided to RTTT LEAs by OSSE and WestEd – we have listed all types of reliability tests found for each assessment. However, the reliability estimate provided (strong, medium, weak, inconclusive) ONLY refers to internal consistency reliability tests, which are commonly considered the gold standard of reliability and the most important for determining an assessment's technical adequacy. Reliability estimates are categorized as strong (coefficient above .9), medium (coefficient .8-.89) and weak, (coefficient below .8) or inconclusive (coefficient undetermined). Common rules of thumb are that these indices should be in the vicinity of .8 and above for group decision and .90 and above for individual decision.¹

Similarly, validity information includes all types of validity tested – consistent with the technical guidance provided to RTTT LEAs by OSSE and WestEd – as referenced by either third-party research or in the publisher's technical manual. However, because validity is specific to the purpose of the assessment – does the assessment measure what it was intended to measure – we have NOT provided validity estimates in this summary. For example, while it may be interesting to know that the assessment was analyzed concurrent validity (how well did the assessment correlate with another assessment?) or predictive validity (how well did the assessment predict performance on another assessment?), these statistics are not as helpful in determining whether or not the LEA should use the assessment to measure student learning and effective teaching.

We have included whether or not the assessment was examined for bias – either by differential item functioning (DIF) analysis or a review panel of experts. Note that bias is present when there are differences in how students from particular subgroups of the same ability perform because of irrelevant difficulties in assessment items.

¹ Herman, et.al. *Guidance for Developing and Selecting Assessments of Student Growth for Use in Teacher Evaluation Systems* (2011).

FIELD ITEM RESPONSES:

Grades/Ages	Subjects/Domains	Purpose of Assessment	Type of Items	Administered	Time	Scored	Reliability Estimates	Bias Estimate	Groups For Which Bias Examined
All Grades/Ages	English/Reading/Literacy	Screening/ Diagnostic	Multiple Choice	Computer Adaptive	> 30 min	Computer	Strong (+.9)	Low to no bias found	African American
Early Childhood	Mathematics	Interim	Constructed Response	Pencil & Paper	30-60 min	Local	Medium (.7- .89)	Presence of bias	Asian
Elementary	Social Studies	Summative	Observation	Teacher Observation	> 60 min	Publisher	Weak (<.7)	Not available	Hispanic
Middle School	Science		Picture-based	Visual/Verbal			Inconclusive		Native American
High School	Early Childhood Development Early Childhood Vocabulary Other		Other	Other					Gender Ethnicity (unspecified)

Assessment Summaries:

Name of Assessment	Grades/Ages	Subjects/Domains	Purpose of Assessment	Type of Items	Administered	Time	Scored	Reliability Estimates	Bias Estimate	Groups For Which Bias Examined
ACCUPLACER: Computerized Placement Tests	High School	English/Reading/Literacy	Screening/Diagnostic	Multiple Choice	Computer Adaptive	< 30	Computer	Medium-High	Low to no bias	African American
		Mathematics		Constructed Response						Hispanic Asian Native American Gender
Achievement Network	Elementary Middle School High School	English/Reading/Literacy Mathematics	Interim	Multiple Choice	Pencil & Paper	30-60	Unknown	Medium-High	Not available	Not available
Core Knowledge Pre-school Assessment Tool (CK-PAT)	Early childhood	English/Reading/Literacy	Screening/Diagnostic	Picture-Based	Teacher Administered/ Observation	< 30	Local	Not available	Not available	Not available
		Mathematics	Interim	Verbal						
		Science	Summative	Observation						
		Early Childhood Development								
		Other								
Edison Sixth Edition	Elementary	English/Reading/Literacy	Screening/Diagnostic	Varied	Paper & Pencil	< 30	Local	High	Not available	Not available

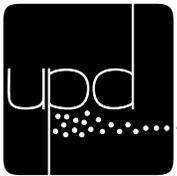
DIBELS Next	Elementary	English/Reading/Literacy	Screening/Diagnostic	Varied	Teacher Administered Paper & Pencil Handheld Mobile	< 30	Local	Medium-High	Not available	Not available
Discovery Education	Elementary	English/Reading/Literacy	Screening/Diagnostic		Computer Adaptive					
	Middle School	Mathematics	Interim	Multiple choice	Pencil & Paper	30-60		Low-Medium	Not available	Not available
	High School		Predictive							
Early Learning Accomplishment Profile	Early childhood	Early Childhood Development	Screening/Diagnostic	Picture-Based						
			Interim	Verbal	Teacher Administered/Observation	30-60	Local	High	Not available	Not available
Gates-MacGinitie Reading Tests®, Fourth Edition, Forms S and T	Elementary		Screening/Diagnostic	Picture-Based			Publisher			African American
	Middle School		Interim	Verbal			Local			Asian
	High School	English/Reading/Literacy	Summative	Observation	Pencil & Paper	> 60		Medium-High	Low to no bias	Hispanic Native American Gender
Learning Accomplishment Profile-3 (LAP-3)	Early childhood	Early Childhood Development	Screening/Diagnostic	Picture-Based						
			Interim	Verbal	Teacher Administered/Observation	> 60	Local	High	Not available	Not available
mCLASS®:CIRCLE™	Early childhood	English/Reading/Literacy	Screening/Diagnostic	Picture-Based	Teacher Administered/Observation	< 30	Local	High	Not available	Not available
			Summative	Observation						

		Mathematics			Verbal						
		Early Childhood Development			Observation						
mCLASS:3D – Text Reading & Comprehension	Elementary	English/Reading/Literacy	Screening/Diagnostic	Verbal	Teacher Administered/Observation	< 30	Local	Not available	Not available	Not available	
Peabody Picture Vocabulary Test, Fourth Edition	All Grades/Ages <i>(Primarily Early Childhood)</i>	English/Reading/Literacy Early Childhood Development	Screening/Diagnostic	Picture - Based	Teacher Administered/Observation	< 30	Local	Not available	Not available	Not available	
Phonological Awareness Literacy Screening PreK	Early childhood	English/Reading/Literacy	Screening/Diagnostic	Varied	Teacher Administered/Observation	< 30	Local	Not available	Not available	Not available	
Stanford Achievement Test	Elementary Middle School High School	English / Reading / Literacy Mathematics Social Studies Science	Interim Summative	Multiple choice	Pencil & Paper	> 60	Local Publisher	Medium-High	Not available	Not available	
Scantron Performance Series	Elementary Middle School High School	English/Reading/Literacy Mathematics Science	Screening/Diagnostic Summative	Multiple choice	Computer Adaptive	30-60	Computer	Medium-High	Low to no bias	Not available	
Teaching Strategies Gold	Early childhood	Early Childhood	Interim	Observation-based	Teacher Administered/Observation	No specific duration	Local	High	Low to no bias	African American	

		English/Reading/Literacy								Hispanic
		Mathematics								Gender
	Elementary	English/Reading/Literacy		Multiple Choice						Ethnicity (unspecified)
TerraNova, Third Edition	Middle School	Mathematics	Summative	Constructed Response	Pencil & Paper	> 60	Publisher	Low-High	Low to no bias	Gender
	High School	Social Studies Science								
Test of Early Mathematics Ability (TEMA), Third Edition	Early childhood	Mathematics	Screening/Diagnostic	Picture - Based	Pencil & Paper	30-60	Local	High	Low to no bias	Ethnicity (unspecified)
	Elementary		Interim							Gender
Test of Preschool Early Literacy (TOPEL)	Early childhood Elementary	English/Reading/Literacy	Interim	Picture - Based Verbal	Teacher Administered/Observation	< 30	Local	Medium-High	Presence of bias	African American Hispanic Gender

Assessments with no available psychometric information:

Psychometric information was not found for the following assessments: Brigance Developmental Inventory; Every Child Ready; Fountas & Pinnell; Iowa Test of Basic Skills; Scantron Achievement Series; Slosson Diagnostic Screener.



**ANALYSIS ON ASSESSMENTS PILOTED FOR USE IN EDUCATOR
EVALUATION SYSTEMS**

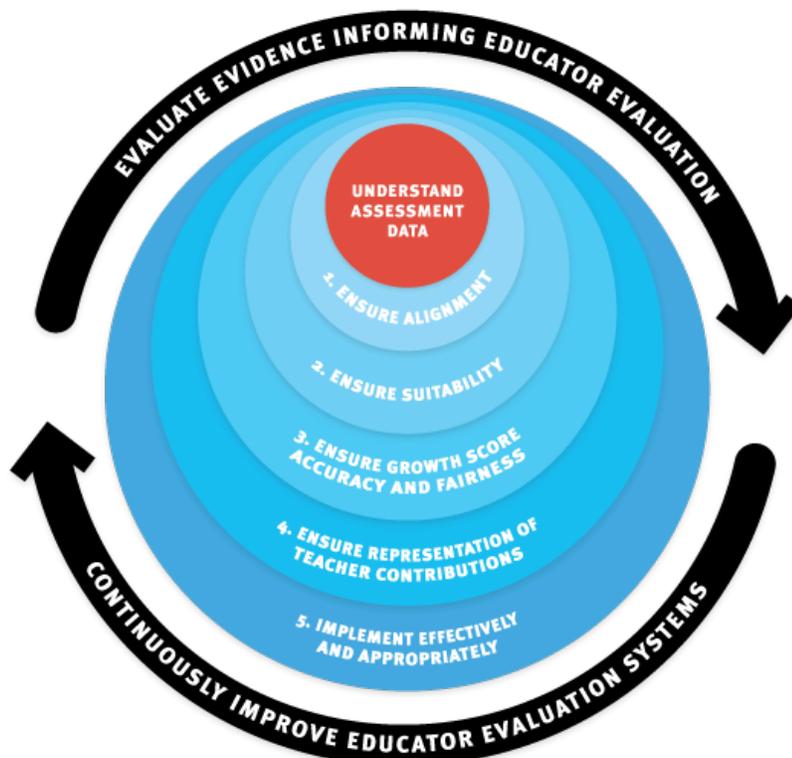
IV. A GUIDING FRAMEWORK FOR SELECTING STUDENT ASSESSMENTS FOR USE IN TEACHER EVALUATIONS

Across the country, states and districts are tackling the challenging questions associated with how to use the results of assessments of student learning as a component of teacher evaluation systems. Carefully designed and validated assessments of student learning can provide valuable evidence of teacher quality. Researchers investigating the psychometric properties of an assessment can make inferences about an assessment’s ability to measure what it claims to measure (validity) and an assessment’s ability to yield stable and consistent results (reliability). Most of the research on the validity of assessments makes inferences about whether or not students have learned what the assessment claims to measure. However, what researchers have not done, for the most part, is taken the next step to validate the use of this data to make claims about the effectiveness of the instructional practice.

Determining whether an assessment is appropriate for use in measuring student learning *and* as a reflection of teacher practice requires a nuanced, reflective process. The prerequisite for this process is highlighted in red to call out its importance; building an understanding of the assessment data currently collected must be the starting point. Please refer to Section VII of this report, titled “Understand Assessment Data,” for a description of the kinds of questions we recommend answering when engaging in a dialogue about assessment data.

Figure 2, below, illustrates the process we recommend for continuously evaluating the evidence informing educator evaluation, including student assessment data, and continuously improving educator evaluation systems.

Figure 2.



Pre-requisite Step: Understand Assessment Data

Before LEAs can dive into answering the questions posed in our *Guiding Framework for Selecting Assessments for Teacher Evaluation*, the first step must be to consider the assessment data with which you are working.

While the framework itself will be useful for LEAs as they inventory and select pilot assessments to inform teacher evaluation, we also acknowledge that LEAs have already selected some assessments to pilot in untested grades. It is imperative that LEAs designate time for building an internal, collective understanding of the data captured by their current pilot assessments.

This prerequisite step is about practitioners – teachers, curriculum developers, testing coordinators, school leaders – sitting down together to ask questions about what the scores collected really mean, what the scores collected reveal about student learning, and what the basis would be of any comparison made about student performance.

The following questions provide guidance around how to engage in this dialogue:

- What types of scores are generated by the piloted assessments?
- What does our current pilot assessment data tell us about student learning?
- Do we know what our scores really mean in terms of student growth?
- How do we know what counts as a “good” score on this assessment? What is our basis of comparison from one score to the next?
- Does the data support our assumptions about student growth and teacher performance?
- Does the data support sufficiently differentiated conclusions about teacher performance?

Step 1: Ensure Alignment

Once there is an understanding of the assessment data itself, we propose that the first step to selecting student assessments is a process of ensuring alignment. Figure 3, below, provides four key questions that LEAs must answer when considering alignment.

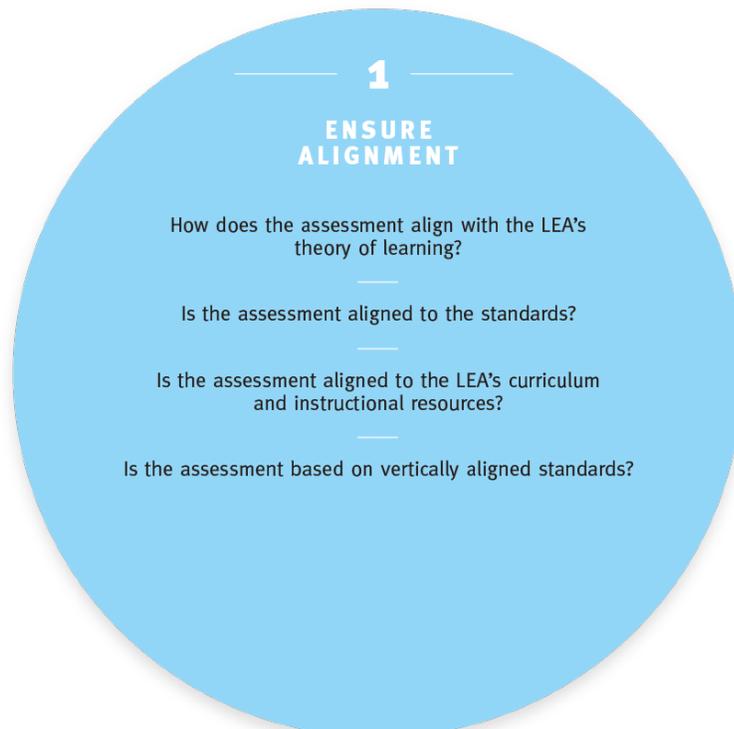
First and foremost, does the assessment align with the LEA's theory of learning? In other words, if an LEA has a project-based learning approach, and all teachers provide project-based instruction, does the selected assessment allow for self-constructed responses by students? If the LEA is using an assessment with only multiple-choice questions, are there other assessments that can be used to compliment the LEA's theory of learning?

The next series of questions asks if the assessment is aligned to the standards, the curriculum and the instructional resources in use. In other words, what is the expected instructional trajectory for students so they are able to exhibit the stated learning outcomes? Do these outcomes align with your instructional materials and with what is being assessed? Verifying this alignment will ensure that the assessment covers the content and skills that students are expected to acquire and demonstrate in a particular course.

This series of questions begs another series of questions about the degree to which the instructional resources, curriculum and standards are aligned. It also begs questions about the fidelity with which instructional resources are implemented in classrooms.

The fourth question below asks about vertical alignment. Vertically aligned standards describe the progression of how students' knowledge and skills in a given subject matter are expected to develop over the course of time, from one grade to the next (Herman et al).

Figure 3.



Step 2: Ensure Suitability

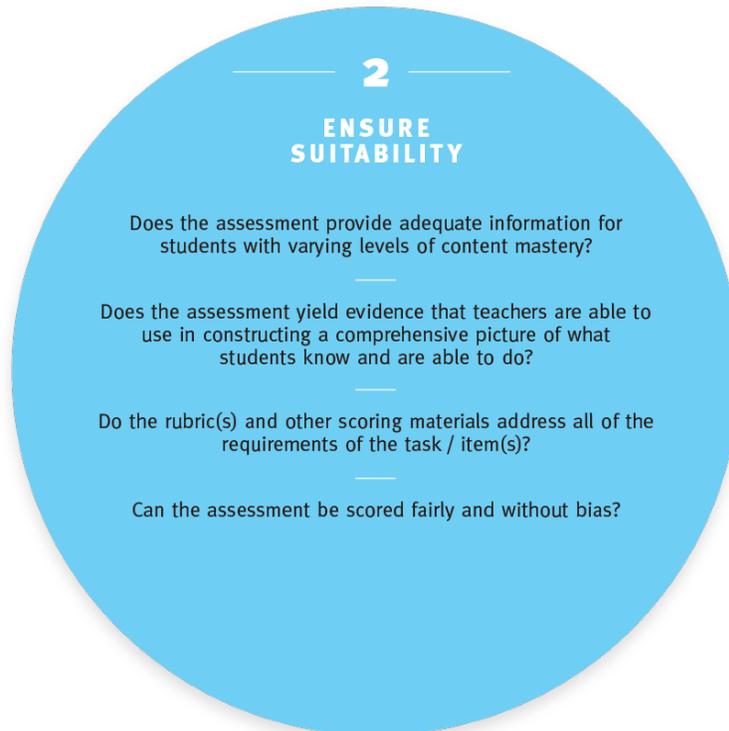
After answering the series of questions about alignment outlined in Step 1, the next step we propose to take in selecting student assessments is to investigate the suitability of the assessment. Figure 4, below, provides four key questions that LEAs should answer when considering suitability.

A high quality assessment is one that yields reliable and meaningful information about what students know and are able to do, and is scored using clear guidelines and criteria. LEAs must examine whether or not the designated assessment provides adequate information about students' competencies as demonstrated by their responses on the assessment. In other words, the assessment data should provide equally rich information for students who demonstrate low, mid-range and high levels of content mastery. Data collected from the assessment should support, and expand upon, teachers' pictures of what students know and are able to do. The assessment should be designed in a way that is accessible and fair for all students.

Does the assessment provide information, when paired with other student performance data, give the teacher a comprehensive view of a student's skills and content knowledge? For example, if you are an ELA teacher, the assessment may demonstrate that a student has strong writing skills. The assessment might not, however, reveal evidence of a student's ability to speak in front of an audience. In order to develop a comprehensive view of the student's ability to communicate, the ELA teacher must consolidate and interpret data from a range of sources.

The assessment itself must be able to be scored fairly and without bias. One indicator of fair and unbiased scoring is the likelihood of different raters arriving at the same score for a given response. Scoring materials must also address the requirements of the task. The scoring categories should be clearly defined and coherent across the range of performance levels. If students are scored for observation-based tasks, scoring must clearly indicate acceptable student responses.

Figure 4.



Step 3: Ensure Growth Score Accuracy and Fairness

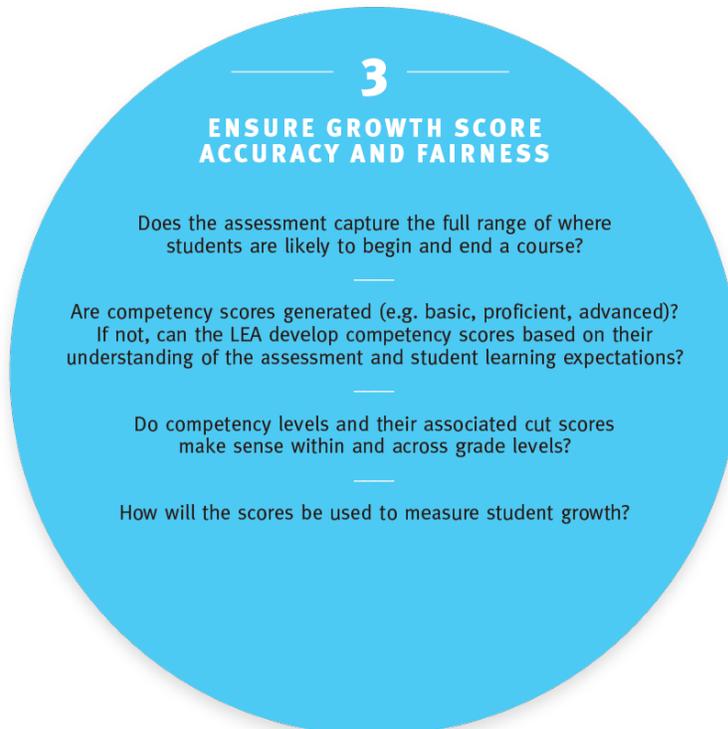
The third step in selecting student assessments is to ensure that the growth score captured is accurate and fair. Figure 5, below, provides four key questions that LEAs should consider to ensure growth score accuracy and fairness.

First, in order to portray students' learning and growth over time, the scores yielded should represent the range for where students fall at the beginning, and then at the end, of a school year. In some cases, the assessment generates a score that indicates a student's skill and content knowledge as basic, proficient, or advanced based on their performance on the assessment. If the assessment does not generate a competency score, LEAs may be able to develop their own bands of basic, proficient, or advanced based on the percentage of items scored correctly on the assessment. Again, this depends on the LEA's understanding of the questions, and the depth of knowledge covered by the questions, on the assessment.

Cut scores for defining proficiency levels must make sense both within and across grade levels. For example, to be classified as "advanced" on a 7th grade science exam should require deeper analysis and more understanding than an "advanced" classification on a 6th grade science exam.

Finally, teachers should know how scores are used to measure student growth. In some cases, growth may be calculated based on the amount of progress a student demonstrates from the start of the school year to the end of the school year. In other cases, growth might be explained by the percentage of content mastered by the end of the school year. Defining adequate growth via assessment scores should occur after carefully reviewing the assessment and reviewing the school's overall academic goals.

Figure 5.



Step 4: Ensure Representation of Teacher Contributions

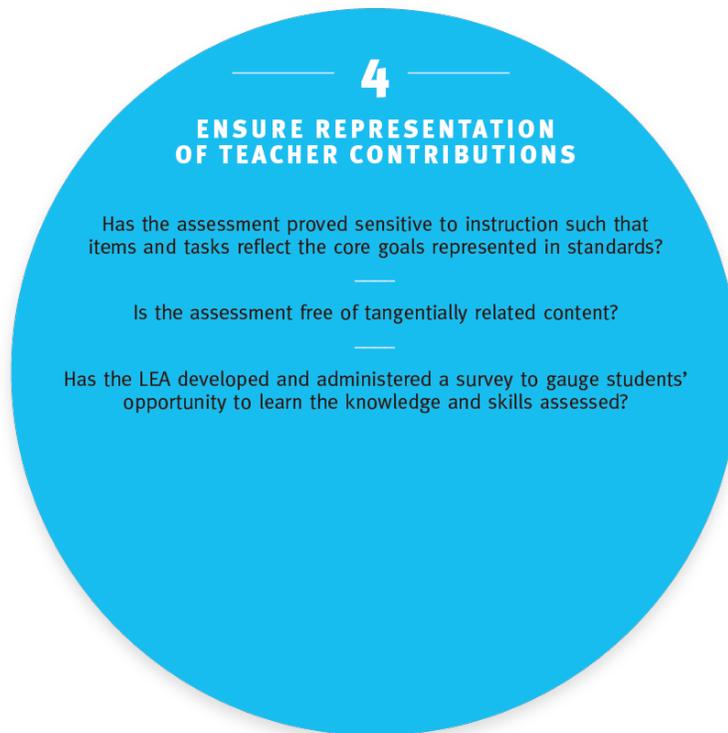
Once an LEA has decided that an assessment is suitable for measuring student learning, it can start the process of gauging whether or not the assessment is suitable for evaluating student growth as a component of teacher evaluation. Figure 6, below, outlines three questions that LEAs must answer to ensure representation of teacher contributions.

To answer these questions, LEAs must examine the assessment’s ability to reflect a teacher’s contribution to student learning. The assessment should be analyzed for its sensitivity to teacher instruction so that students receiving poor instruction and students receiving well-crafted and well-delivered instruction do not demonstrate the same learning competencies.

The assessment should be free of tangentially related content. In other words, it should evaluate students’ performance on the targeted learning goals represented in the standards and curriculum, and should not evaluate students on content that is not covered, or only tangentially related.

One suggestion made by both the Gates Foundation in their *Measures of Effective Teaching* study, and in the paper titled “Guidance for Developing and Selecting Assessments of Student Growth for Use in Teacher Evaluation Systems” written by Joan Herman and her colleagues, is to administer a survey to gauge student perceptions about the quality of their learning experiences. Administering a survey to gauge students’ perceptions about the effectiveness of the teaching to which they are exposed in concert with test administration is one method for ensuring that scores represent individual teachers’ contributions, and can be used as a tool for interpreting students’ performance.

Figure 6.



Step 5: Implement Effectively and Appropriately

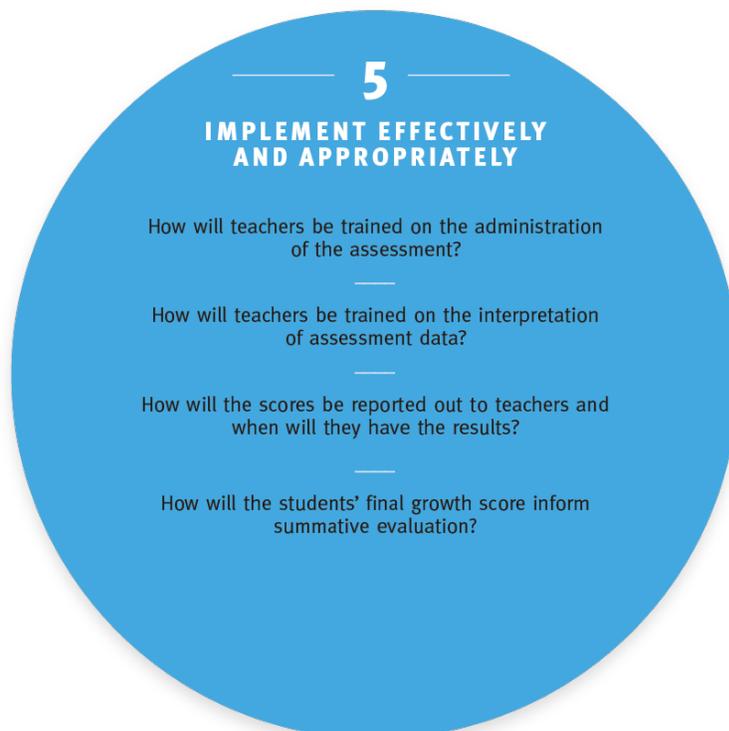
The final step we propose LEAs take in selecting student assessments to inform educator evaluation is to implement assessments effectively and appropriately. Figure 7, below, outlines the four questions that LEAs should answer in order to use an assessment in a manner that ensures fairness and integrity among their staff.

First, teachers must be trained on the administration of the assessment, including, but not limited to, any guidelines for the amount of time allotted for students to complete the assessment and accommodations made for students with special needs. Teachers must also receive training on the interpretation of assessment data. Once teachers receive their assessment data, they should know how to interpret it. What does the data tell a teacher about their students' learning? Can this data be used to inform future instruction?

Teachers should also know when they will see their students' results and how the scores will be reported. Will teachers look at aggregate class data or will they only have access to individual student data? Should teachers expect the school's data personnel to distribute reports or can teachers access the data on their own?

Finally, what will student scores mean for evaluation ratings for teachers in non-tested grades and subjects? In some cases, student growth scores contribute to 30% of a teacher's overall evaluation score while another LEA may decide that student growth scores account for 50% of a teacher's evaluation score. A teacher might have to have a certain percentage of all of their students attain one year of growth to be deemed 'effective' for the growth component of their evaluation or an LEA might decide that one year's growth is not ambitious enough given the school's performance on the exam, but instead suggest 1.25 years of growth as adequate. In either case, we are not making a recommendation, but instead, urging LEAs to do a thorough review and analysis before finalizing growth expectations *and* share these expectations with instructional staff.

Figure 7.



V. RECOMMENDATIONS TO SUPPORT THE DEVELOPMENT OF STUDENT GROWTH MODELS

The process for collecting student data and developing student growth calculations to attribute student growth to teacher performance can only be accomplished after an LEA has established that the assessment is an appropriate and fair measure of student learning. In order for an assessment to yield growth scores, LEAs must first determine that the assessment (1) is technically reliable and valid² (2) aligned to the LEA's theory of learning (3) aligned to the LEA's standards and curriculum (4) contains suitable scoring components (5) has an accurate and fair process for determining growth scores (6) is sensitive to teacher contributions to student learning and (7) can be implemented effectively and appropriately.

We make the following recommendations in an effort to support the development of student growth models in non-tested grades and subjects.

Recommendation #1: Engage in a Dialogue about Data

Before LEAs can dive into answering the questions we pose in our *Guiding Framework for Selecting Assessments for Teacher Evaluation*, they must first consider the assessment data with which they are working. We recommend that LEAs engage in an internal dialogue about what they have learned from their pilot assessment results. The following questions can guide this dialogue:

- What types of scores are generated by the piloted assessments?
- What does our current pilot assessment data tell us about student learning?
- Do we know what our scores really mean in terms of student growth?
- How do we know what counts as a “good” score on this assessment? What is our basis of comparison from one score to the next?
- Does the data support our assumptions about student growth and teacher performance?
- Does the data support sufficiently differentiated conclusions about teacher performance?

Recommendation #2: Engage in a Dialogue about Assessment Viability

Just as we suggest that LEAs designate subject matter and special population experts to review and compare test items to standards, we also suggest that LEAs designate subject matter and teacher experts to engage in a similar process for determining whether their assessments can be used to measure and attribute student growth to teacher performance. The questions within the *Guiding Framework for Selecting Assessments for Teacher Evaluation* will be useful as LEAs select assessments to measure student learning and inform teacher evaluation. The process for determining assessment viability involves establishing the degree to which the assessments are vertically aligned so multiple scores can be used as a “yardstick” of progress that runs from the lowest to highest skill and development levels³, analyzing the precision and reliability of using growth measures to identify real differences among teachers based on student growth, and testing out different assumptions behind performance levels and cut scores through the analysis of empirical data.

² Please note that when we make reference to the technical “validity” of an assessment we are not referring to a characteristic of the instrument. Rather, by “validity” we are referring to the inferences one can make based on the results the instrument yields, and to the ability of the instrument to measure what it claims to measure.

³ Guidance for Developing and Selecting Assessments of Student Growth for Use in Teacher Evaluation Systems (Extended Version), Herman et al (2011)

The process for investigating the viability of a particular assessment tool for the purposes of measuring and attributing student growth is captured in Step 3 of our proposed *Guiding Framework for Selecting Assessments for Teacher Evaluation*.

Recommendation #3: Build a Validity Study

Once LEAs are engaged in the work of using assessment data to develop student growth calculations to attribute to teacher performance, we recommend developing a process for checking the quality of the evidence collected.

Our recommendation is to try an approach, administer a pre and post assessment, and then analyze the data collected and see where teachers land. Test the system to see if the conclusions LEAs draw about teacher performance align with the rest of the evidence they have collected. In other words, can LEAs make inferences from the assessment data about teacher performance that align with the inferences they make through classroom observations, student surveys, and teachers' own reflective work. While these are only the initial steps of a validity study, they will get LEAs on the path of checking the quality of the evidence they have collected.

VI. SUMMARY

Developing a comprehensive teacher evaluation system to fit the unique needs of your LEA is a complicated process. Such a process requires that you have a clear vision for the intentions of the newly developed system. It also requires that you identify the pieces of information that will be most valuable for creating a system that both supports and evaluates teacher performance, and that you continuously improve the system based on the feedback and data received.

Student assessment can provide invaluable information about what students are learning. While assessments may display evidence of being statistically valid and reliable, we realize that these data points are only useful if the assessment also provides the LEA with the information it needs to measure student learning in its unique context. To that end, we strongly suggest that LEAs take the time to consider the value of the data yielded by their selected pilot assessments. By engaging in the process outlined in Section V, *A Guiding Framework for Selecting Student Assessments for Use in Teacher Evaluations*, LEAs can determine if the selected assessments are indeed meeting their needs for measuring student learning and informing teacher evaluations. Following this process, our goal is that LEAs will have the information necessary to make a number of decisions, including, but not limited to, the option of re-evaluating their curriculum, re-considering the assessments used to measure student learning, re-setting the growth targets for the students and receiving additional targeted support from OSSE to modify the teacher evaluation systems at work.

VII. GLOSSARY

Reliability Terms

Reliability – Consistency of measurement overall and at various points of the score scale. Assessments should yield similar results over time with similar populations in similar circumstances.

Internal consistency – Determines whether items that propose to measure the same construct produce similar results. Internal consistency is usually measured with Cronbach's alpha, a statistic calculated from the pairwise correlations between items. Internal consistency ranges between zero and one. Coefficients above .8 are generally considered good.

Standard error of measurement (SEM) – Measurement error on an assessment.

Test-retest – A method of estimating test reliability in which the same assessment is given to the same group of research participants on two different occasions (separated by days, weeks, or months). The results from the two tests are then correlated to see if the test is stable over time.

Split-half – A measure of reliability where a test is split in two and are scored separately. The score of one half of test are compared to the score of the remaining half to test the reliability.

Scorer/Interrater – Measures the degree of agreement between persons scoring a subjective test (like an essay exam) or rating an individual. This type of reliability is most often used when scorers have to observe and rate the actions of participants in a study.

Alternate form – A method of reliability in which two forms of the same test are created, with items slightly varied, to determine stability.

Validity Terms

Validity –The test measures the desired performance and appropriate inferences can be drawn from the results. The assessment accurately reflects the learning it was designed to measure.

Construct validity – The test is measuring the target skills and content. (Example: An example could be a doctor testing the effectiveness of painkillers on chronic back sufferers. construct validity would test whether the doctor actually was measuring pain and not numbness, discomfort, anxiety or any other factor.

Criterion validity – The assessment has the expected relationships with other measures of the same construct.

Concurrent validity – Type of criterion validity. There is evidence that an assessment correlates well with another assessment that has previously been validated. (Example: The written drivers test is a replacement for driving around with an observer until you show you know the rules).

Predictive validity – Type of criterion validity. There is evidence that the assessment predicts future performance. (Example: Your GRE score (taken now) predicts how well you will do in grad school)

Consequential validity – There is evidence that adverse consequences are minimal.

Bias Terms

Bias – A situation that occurs when items systematically measure differently for different ethnic, gender, or age groups. Test developers reduce bias by analyzing item data, then identifying and discarding items that appear to be biased.

Differential item functioning (DIF) analysis – Item-level analysis to determine if individuals that have the same ability but belong to different groups (commonly gender, race/ethnicity) have a different probability of success on an item.

Bias review panels – Item-level analysis by expert panel to determine any presence of bias.

Scoring Terms

Norm – Referenced – Scores on a norm – referenced assessment compare the performance of a tested student to the performance of a predetermined population. Norm – referenced assessments compare students to one another.

Criterion – Referenced – Scores on a criterion – referenced assessment reveal whether or not the tested student performs well or poorly on a given task; criterion – referenced scores do not reveal anything about how the tested student performs compared to other test takers.

Percentile Rank – The percentile rank of a score is interpreted as the percentages of students tested in the norm group who scored below the score of interest. Percentile ranks are normally distributed, or distributed on a bell curve.

Grade Equivalent – A grade equivalent score expresses the grade level of students tested based on their raw score. Raw scores can be converted to grade equivalent scores.

Raw – A raw score is the original test data before it is transformed into a percentile rank or a standard score. A raw score that can be used for statistical purposes, for example, reflects the number of correctly answered test items. Raw scores are converted to Standard Scores.