

## CHAPTER I (Draft 7-21-2015)

### PURPOSES, USES, AND INTENDED OUTCOMES OF ASSESSMENT

#### TYPES OF ASSESSMENT AND APPROPRIATE USES (INCLUDING HIGH STAKES USES)

While there are several possible categorizations of assessment by type, this chapter reviews only one of the possible categorizations as particularly relevant to the work of the Task Force: the distinction between formative, interim, and summative assessment. Therefore, this section defines formative, interim, and summative assessment and provides guidelines on appropriate uses of data gathered from the three types of assessment. The guidelines for appropriate use are underlined to make them clearly visible. In defining formative, interim, and summative assessment, this section heavily borrows from Perie et al. (2007) and Michigan Department of Education (2013).

The term “high-stakes use” deserves some pointed attention in that an important question is for whom the stakes are high. Stakes may be high for students, individual educators such as teachers or administrators, or educational entities such as schools and districts. For students, high-stakes uses of test scores may include use of student test scores for making decisions regarding grades, grade promotion, tracking, graduation, admission to postsecondary education or training, and scholarships. For individual educators, student test scores in the educational entities they lead (classrooms, departments, schools, districts) may factor into formal periodic evaluations used in important employment decisions, or in employment decisions based on less formal use of student outcomes. In addition, students, teachers and administrator are also affected by high-stakes uses of test scores in school and district accountability, which may lead to involuntary participation in interventions intended to correct poor outcomes.

Two of several characteristics on which formative, interim, and summative assessments vary are the frequency of the cycle of assessment and the scope (and depth) of content covered by an assessment. We adopt the terms *short cycle*, *medium cycle*, and *long cycle* to indicate frequency of assessment with small-cycle generally happening daily, large cycle happening at the end of a large unit of instruction (e.g., a marking period or grade), and medium cycle happening with a frequency somewhere in between. We use the terms *narrow scope*, *medium scope*, and *wide scope* to indicate the breadth of coverage with narrow-scope covering something between a portion of a single content standard to a very small collection of content standards, wide-scope covering the content of an entire large unit of instruction (such as a marking period or grade), and medium scope covering something between the two. Narrow-scope assessments tend to go into greater depth on the limited content they cover, wide-scope assessments tend to cover content less deeply because they cover so much content, and medium-scope assessments tend to be somewhere in between.

Three broad types of assessment (formative, interim, and summative) are introduced below. The key differences between formative, interim, and summative assessment can be described in terms of frequency of the cycle of assessment and the scope (and depth) of content covered by assessment (adapted with permission from Perie, et al., 2009).

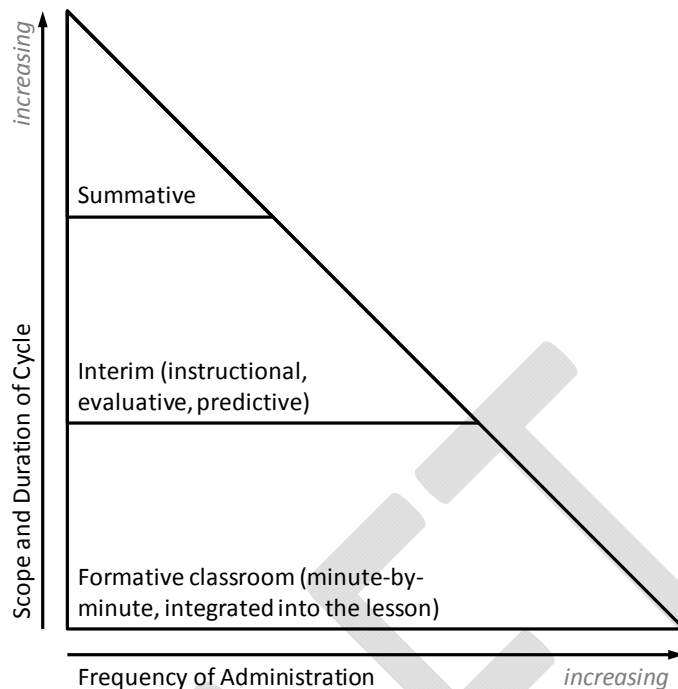


Figure 1. Frequency, Scope, and Cycle by Type of Assessment.

## Formative Assessment

Formative assessment has also been called formative instruction in that the purpose of formative assessment is to examine and adjust instruction on a moment-to-moment basis based on an understanding of where a student currently is in the process of attaining a clearly specified, small-scale intended learning target.

In 2006, state education leaders, the Council of Chief State School Officers (CCSSO), and national and international experts on formative assessment collaborated to develop a widely cited definition of formative assessment:

*Formative assessment is a process used by teachers and students during instruction that provides feedback to adjust ongoing teaching and learning to improve students' achievements of intended instructional outcomes.* (As defined by the CCSSO FAST SCASS in 2006).

The importance of this definition is that it is compatible with research showing such practices to be associated with student learning gains. At the core of the formative assessment process is that it takes place during instruction (i.e., “in the moment”) under full control of the teacher to support student learning while learning is developing. This is done through diagnosing on a very short cycle where students are in their learning, where gaps in knowledge and skill exist, and how to help students close those gaps.

Another important conclusion one can draw from this definition is that it is embedded within instruction. Instruction does not stop to engage in formative assessment. They are narrow-scope (a few seconds, a few minutes, certainly less than a class period covering from a portion of a content

standard to a very small collection of content standards) and short-cycle (they happen often within a single unit of instruction to diagnose student understanding before a unit assessment).

This definition of formative assessment makes clear that it is not a product, but a process tailored to the details of ongoing instruction and to individual students. If tasks are presented, they may vary for students depending on where they are in their learning. However, formative assessment processes often occur during regular and targeted questioning of students in small or large groups, observing students as they work in groups and/or engage in tasks, and in many other regular classroom settings. Formative assessment practices, especially when the teacher is appropriately trying to systematically engage with all her students, may be facilitated using certain technology and related tools. There is a strong view among some scholars of formative assessment that because formative assessment is tailored to the specific context of the classroom and to individual students that results cannot be meaningfully aggregated or compared. Many of these scholars also question whether, in practicing formative assessment, student knowledge and skill should even be scored.

Another conclusion that may be drawn from this definition of formative assessment is the critical importance of providing frequent feedback to individual students. Such feedback has specific characteristics that develop in each student the ability to continuously monitor the quality of their own work against a high quality target. It is this feedback to students that is the most crucial part of the formative assessment process (see Sadler, 1989).

Data gathered through formative assessment have limited to no use for evaluative or accountability purposes such as student grades, educator accountability, school/district accountability, or even public display that could allow for inappropriate comparisons. There are at least four reasons for this: (1) if carried out appropriately, the data gathered from one unit to the next, one teacher to the next, one moment to the next, and one student to the next will not be comparable; (2) students will be unlikely to participate as fully, openly, and honestly in the process if they know they are being evaluated by their teachers or peers on the basis of their responses; (3) educators will be unlikely to participate as fully, openly, and honestly in the process if they know they are being evaluated by their peers, supervisors, or their students' parents on the basis of their students' responses; and (4) the nature of the formative assessment process is likely to shift in such a way that it can no longer optimally inform instruction.

These implications create a distinct difference from summative and interim assessment (described below), which are intended to assess student achievement after an extended period of learning. Simply giving students an assessment in the classroom does not mean that the assessment is formative. Use of assessment evidence in a formative manner requires teachers to gain insights into individual student learning in relation to standards, to provide effective feedback to students about those insights, and to make instructional decisions based on those insights to guide next steps. During the formative assessment process, feedback to students and student involvement is an essential component. Teachers seek ways to involve the student in "thinking about their thinking" (metacognition) to use learning evidence to close the gap and get closer to the intended learning target.

While formative assessment is not a new idea, teachers are not typically trained in-depth on the process, especially on providing specific, effective feedback to students to engage them in metacognition. Simply putting resources and tools into teacher hands is not sufficient. Sustained professional development and material support is needed to implement sound formative assessment

practices, including the most important part of the formative assessment process: providing targeted and useful feedback to individual students.

Because the formative assessment process is intended to differ considerably from teacher to teacher, topic to topic, and student to student, there is little sense in trying to aggregate data gathered from formative assessment beyond a specific classroom. In addition, the intended nature and use of formative assessment (informing instruction on a moment-to-moment basis) is likely to be corrupted if data are used in any high-stakes manner, meaning in any way other than to inform instruction (e.g., student grades, educator or school/district accountability).

### **Summative Assessment**

Summative assessments are generally long-cycle, wide-scope assessments given once at the end of some large unit of instruction such as a marking period, course, or grade to evaluate students' performance against a defined set of content standards. The prototypical assessment conjured by the term "summative assessments" is given in a standardized manner statewide (but can also be given nationally or districtwide) and these days are usually used as part of an accountability program or to otherwise inform policy. Such summative assessments are typically the least flexible. Appropriate uses of such standardized summative assessment include school accountability, district accountability, curriculum evaluation, program evaluation, and to inform policy-makers in their high-level decision-making.

Less standardized but no less summative assessments are also found in the majority of middle- and high-school classrooms. Such assessments include exams given near the end of a large defined unit of instruction such as the large unit of instruction covered during a complete marking period. Some common examples are broad exams or projects intended to give a summary of student achievement of marking period objectives, and figure heavily in student grading. Such assessments tend to be labeled "final projects" or "final exams" in middle and high school grades, but do not have a consistent label in elementary grades. Classroom summative assessments may be created by individual teachers or by staff from one or more schools or districts working together. Appropriate uses of such summative assessments include student grading in the specific courses for which they were developed. If designed well they can also be used to adjust curriculum, programming, and instruction the next time the large unit of instruction is taught; and to serve as a post-test measure of student learning. If the assessments are well-designed and a carefully- and well-defined set of rules is in place for appropriate administration, scoring, and use of results they may also be reasonably used for accountability.

### **Interim Assessment**

Many of the assessments products currently in use that are labeled "formative," "benchmark," "diagnostic," or "predictive" actually belong in the interim assessment category in that they do not facilitate moment-to-moment targeted analysis of student learning, frequent feedback to students and teachers, and timely adjustment of instruction; nor are they one-time assessments intended to provide a broad summary of achievement of course- or grade-level learning objectives tied to specific state content standards.

Interim assessment includes assessments that fall between formative and summative assessment, including many standardized, medium-scope, medium-cycle assessments currently in wide use.

Interim assessments (1) evaluate students' knowledge and skills relative to a specific set of academic goals, typically within a limited time frame, and (2) are designed to inform decisions at both the classroom, school, or district level. Thus, they may be given at the classroom level to provide information for the teacher, but unlike formative assessment, the results of interim assessments can be meaningfully aggregated and reported at a broader level. As such, the timing of the administration is likely to be controlled by the school or district rather than by the teacher. The content and format of interim assessments is also likely to be controlled by the test developer. Therefore these assessments are considerably less instructionally-relevant than formative assessments in that decisions at the classroom level tend to be ex post facto decisions regarding post-unit remediation needs and adjustment of instruction the next time the unit is taught.

These assessments may be appropriate for a variety of uses, including student grading, predicting a student's ability to succeed on a large-scale summative assessment, evaluating a particular educational program or pedagogy, identifying potential gaps in a student's learning after a limited period of instruction has been completed, or measuring student learning over time. If interim assessments are well-designed and a carefully- and well-defined set of rules is in place for appropriate administration, scoring, and use of results, measures of student learning may also be reasonably used for school or educator accountability.

Again, less standardized interim assessments are also found in nearly all classrooms. Such assessments include teacher- or district-developed end of unit exams and projects. Common examples include typical middle- and high-school unit projects, or midterm exams or projects, and end-of-unit assessments. Even shorter-cycle classroom interim assessments are found in nearly all classrooms in such forms as daily homework, frequent quizzes, and unit projects. These assessments are intended to provide a summary of students' achievement of well-defined medium-sized groups of grade/course/credit objectives. Such assessments cannot be considered formative assessments, though they can be used in a somewhat formative manner (e.g., identifying needed remediation for both a group and individual students before a full unit of instruction has been completed).

Appropriate uses of such unit/sub-unit interim assessments tend to be limited to student grading, mid-unit adjustment of instruction, and/or post-unit evaluation of instruction and identification of remedial needs in the specific grades or courses for which they were developed. If classroom interim assessments are well-designed and a carefully- and well-defined set of rules for appropriate administration, scoring, and use of results is put in place they can also be used in measuring student learning over time for accountability purposes.

There are two other types of interim assessments currently in use beyond the "backward looking" interim assessments described above. Both are "forward-looking." One unfortunately rare type is a pre-test given before a unit of instruction to gain information about what a student and/or group of students already knows to adjust plans for instruction before beginning the unit (teachers may do these pre-instruction checks on a more frequent, formative basis, but both are rare in our experience). Such forward-looking assessments may be composed of pre-requisite content or the same content as the end-of-unit assessment. Another type of forward-looking assessment is a test is intended to predict how a student will do on a summative assessment before completing the full unit of instruction. The usefulness of this type of interim assessment is debatable in that the time spent is unlikely to provide a great deal of instructionally relevant information and there is often other information available to determine who is likely to need help to succeed on the end of year summative assessment.

## A Note on Classroom Assessment and Accountability

If considerable resources are provided to support classroom formative, interim, and summative assessment, there may be a reasonable question as to whether the funds are being invested wisely. One temptation may be to hold educators, schools, and/or districts accountable for results on classroom assessment, but such uses are inappropriate for formative and interim assessment, and great care is needed when using classroom summative assessment for such uses. Rather than holding schools and/or teachers accountable for student data gathered from classroom interim and formative assessment, the investment could be evaluated instead by

- Monitoring the *quality* of formative, interim, and summative classroom assessment practices *rather than outcomes* based on those assessments in such a way that encourages collaboration.
- Requiring teachers and administrators to attend high-quality professional development (PD) on best practices in classroom assessment.
- Monitoring the *degree and quality of administrator support* for teachers to collaborate and improve their formative, interim, and summative classroom assessment practices *rather than outcomes* based on those assessments.

If student results from formative or interim classroom assessment are used for educator or school accountability beyond the very limited uses described above, implementation is likely to be corrupted and beneficial instructional effects of the investment are likely to be lost.

### DESIRED PURPOSES, USES, AND OUTCOMES OF ASSESSMENT AS IDENTIFIED BY THE TASK FORCE

#### *Identifying High-Priority Uses and Purposes*

The Task Force considered that assessment design is always a case of optimization under constraints (Braun, in press). In other words, there may be many desirable purposes, uses, and goals for assessment. However, they may be in conflict, any given assessment and type of assessment can only serve a limited number of purposes well, and assessments always have some type of restrictions (e.g., legislative requirements, time, cost, etc...) that must be weighed in finalizing recommendations.

Task Force members initially were asked to ignore constraints for the moment, and identify their desired purposes and goals for assessment and their desired uses of assessment data. The groups noted their highest priority uses, and then reviewed the work of other groups, asking clarifying questions. After each group's highest priority uses and purposes were reviewed, each individual panelist identified their three highest priorities. The group then discussed possible patterns emerging from the activity. Task Force members were reminded that they could at any time adjust this preliminary work.

In general, Task Force members desire a Wyoming assessment (system) that is capable of serving the following broad purposes:

- Providing instructionally-useful information to teachers and students (with appropriate grain-size and timely reporting)

- Providing clear and accurate information to parents and students regarding students' achievement of and progress toward key outcomes, such as progress toward meeting grade-level standards and progress toward readiness for post-secondary education and/or training
- Providing meaningful information to support evaluation and enhancement of curriculum and programs
- Providing information to appropriately support federal and state accountability determinations

To provide more specificity to the work that Task Force completed, each member's top priority, second priority, and third priority uses and characteristics were given a score of 3 (highest priority), 2, and 1, respectively. The results of this activity are shown in Table 1 below. Any top priority uses and characteristics that were similar were consolidated. In consolidating, important differences in each contributing use/characteristics were incorporated into the description.

The Task Force's highest priority uses and characteristics were evaluated using the definitions and appropriate uses of formative, interim, and summative assessments discussed above and differences in classroom-, district-, and state-level assessment. This evaluation is reflected in additional elements added to Table 1 categorizing each type and level of assessment by degree of applicability to each high-priority use/characteristic. In each row the degree of applicability is briefly explained..

There are a few outcomes worth noting from Table 1. First, no single type of assessment (formative, interim, or summative) is applicable to all of the purposes. In fact, formative assessment is uniquely applicable to two uses/characteristics (meaning that *only* formative assessment can support the desired use/characteristic) and summative assessment is uniquely applicable to three uses/characteristics. The same is true for level of assessment, in which classroom and state assessment are each uniquely applicable to three different uses/characteristics.

Second, in some cases, formative, interim, and summative; and classroom, district, and state assessment fulfill different aspects of a use/characteristic (e.g., the top row of Table 1, among others).

Third, all types and levels of assessment are fully applicable to multiple uses/characteristics and somewhat or minimally applicable to one or more additional uses/characteristics.

These three outcomes of the Task Force's work mean that in order to accomplish the full set of uses/characteristics, a system of assessments would be required that span the range of formative, interim, and summative; and classroom, district, and state.

Table 1. *Task Force Highest Priority Uses and Characteristics.*

| Total <sup>1</sup><br>Score | Number of<br>Votes by<br>Priority |                 |                 | Desired Uses and Characteristics of Wyoming Assessment   | Applicability <sup>2</sup> |         |           |           |          |       |
|-----------------------------|-----------------------------------|-----------------|-----------------|--|----------------------------|---------|-----------|-----------|----------|-------|
|                             |                                   |                 |                 |  | Type                       |         |           | Level     |          |       |
|                             | 1 <sup>st</sup>                   | 2 <sup>nd</sup> | 3 <sup>rd</sup> |  | Formative                  | Interim | Summative | Classroom | District | State |
| 38                          | 10                                | 3               | 2               | Provide information regarding individual student achievement and growth within and across years, including readiness for the next level in a student's K-12 progression<br>- Classroom formative: narrow-scope, within-unit, daily readiness information<br>- Classroom/district/state interim: moderate-scope, within-year, unit readiness information<br>- Classroom/district/state summative: wide-scope, next-year/next-course readiness information     | ●                          | ●       | ●         | ●         | ●        | ●     |
| 27                          | 6                                 | 4               | 1               | Provide feedback on progress toward standards to inform instruction on more than a yearly basis<br>- Classroom formative: continuous, narrow-scope information informs daily instruction<br>- Classroom/district/state interim: periodic, moderate scope information informs remediation<br>- District/state summative: interim results might be rolled up for summative determinations  | ●                          | ●       | ◐         | ●         | ●        | ●     |
| 16                          | 0                                 | 5               | 6               | Allow for comparisons within the state and across states<br>- State interim: provides within-state comparability if adopted statewide<br>- State summative: provides within-state comparability<br>- State interim/summative: provides cross-state comparability if a multi-state assessment is used   | ○                          | ●       | ●         | ○         | ○        | ●     |
| 13                          | 2                                 | 2               | 3               | Provide reliable & valid data to evaluate program/curriculum effectiveness & alignment to standards<br>- District/state interim: can provide information to inform within- and between-year evaluations<br>- District/state summative: can provide information to inform between-year evaluations  | ○                          | ●       | ●         | ○         | ●        | ●     |
| 11                          | 3                                 | 1               | 0               | Be student-centered (e.g., student is not a number)<br>- Classroom formative: narrow-scope, diagnostic data to tailor instruction<br>- Classroom/district/state interim: moderate-scope, unit-based data to tailor remediation<br>- Classroom/district/state summative: wide-scope data to inform critical yearly decisions  | ●                          | ●       | ◐         | ●         | ●        | ●     |
| 8                           | 0                                 | 3               | 2               | Encourage collaboration and sharing best practices<br>- Classroom formative/interim/summative: foster teacher collaboration on teacher practices<br>- District/state interim/summative: foster teacher collaboration on using non-classroom data<br>- District/state interim/summative: foster administrator collaboration on curriculum/programming<br>- Limit use of classroom assessment for evaluation to quality of practices/support for collaboration | ●                          | ●       | ●         | ●         | ●        | ●     |
| 7                           | 1                                 | 2               | 0               | Continually inform instruction with timely feedback<br>- Classroom formative: continual, narrow-scope, diagnostic data to inform daily instruction<br>- Classroom/district/state interim: periodic, moderate-scope, unit-based data to inform remediation  | ●                          | ◐       | ○         | ●         | ◐        | ◐     |



| Total <sup>1</sup><br>Score | Number of<br>Votes by<br>Priority |   |   | Desired Uses and Characteristics of Wyoming Assessment  | Applicability <sup>2</sup> |         |           |           |          |       |
|-----------------------------|-----------------------------------|---|---|---|----------------------------|---------|-----------|-----------|----------|-------|
|                             |                                   |   |   |   | Type                       |         |           | Level     |          |       |
|                             |                                   |   |   |   | Formative                  | Interim | Summative | Classroom | District | State |
| 6                           | 1                                 | 1 | 1 | Validly inform decisions about post-secondary education/training<br><i>- State summative: likely to provide based on ties to post-secondary outcomes (onerous for a district)</i>   | ○                          | ○       | ●         | ○         | ◐        | ●     |
| 2                           | 0                                 | 0 | 2 | Consistency over time to facilitate the intended outcomes of assessment in Wyoming<br><i>- District interim/summative: stable longitudinal data can improve decision making</i><br><i>- State interim: stable longitudinal data can improve decision making</i><br><i>- State summative: likely to improve decision-making because of school/district accountability uses</i> | ○                          | ◐       | ●         | ○         | ◐        | ●     |
| X                           |                                   |   |   | Number of desired uses/characteristics with unique and full applicability   | 2                          | 0       | 3         | 3         | 0        | 3     |
|                             |                                   |   |   | Number of desired uses/characteristics with full applicability  | 4                          | 3       | 5         | 4         | 2        | 5     |
|                             |                                   |   |   | Number of desired uses/characteristics with some applicability  | 1                          | 4       | 1         | 1         | 4        | 3     |
|                             |                                   |   |   | Number of desired uses/characteristics with unlikely applicability  | 0                          | 1       | 2         | 0         | 2        | 1     |
|                             |                                   |   |   | Number of desired uses/characteristics with no applicability  | 4                          | 1       | 1         | 4         | 1        | 0     |

- Each panelist identified one characteristic as her highest priority, second highest priority, or third highest priority. These were given scores of 3, 2, and 1 respectively. The scores were summed across panelists to give a total score for each desired use/characteristic.
- , ◐, ◑, and ○ indicate desired uses or characteristics for which the type or level of assessment has full, some, minimal/unlikely, and no applicability, respectively.

## *Intended Outcomes of Implementation*

There are a few intended outcomes that can be inferred from the uses and characteristics in Table 1. In meeting #2, the Task Force did not have any edits or comments on these intended outcomes. They remain listed here as preliminary until the Task Force directly addresses this topic:

- Assist students and parents to become more engaged in their own education through a greater knowledge of (1) their strengths and weaknesses and (2) their current and likely future achievement of desirable outcomes through daily formative information and periodic interim and summative information.
- Facilitate moment-to-moment individualized instruction to maximize desirable outcomes across the continuum of existing student achievement (from far behind to far ahead) through integration of high-quality formative assessment practices into instruction for all educators.
- Improve statewide achievement and growth rates compared to previous Wyoming achievement and in comparison to other states through (1) the provision of high-quality data stable across many years to high-level policymakers to inform changes in educational policy, (2) holding schools and districts appropriately accountable for student outcomes, and (3) providing valid data to local administrators to adjust programs and curriculum to target areas of weakness.
- Improve day-to-day instruction and student grading practices by (1) encouraging and/or evaluating teacher-level collaboration on classroom assessment and integration into instruction, and (2) encourage and/or evaluating the adequacy of material state, district, and school administrative support for initial and ongoing professional development.

### **A STATEWIDE SUMMATIVE ASSESSMENT OR A COMPREHENSIVE ASSESSMENT SYSTEM?**

#### **Why the Question?**

A single assessment is incapable of fulfilling the various highest-priority purposes and uses identified by the Task Force. In order to fulfill these purposes, all three types of assessment would be necessary. However, in the same way that a pile of bricks does not make a house, a haphazard collection of assessments at the classroom, school, district, and state level would not be capable of fulfilling the identified purposes (see Coladarci, 2002).

To appropriately facilitate all of the purposes and uses would require the design and implementation of a comprehensive assessment system that coherently integrates information from various assessments at the classroom, school, district, and state levels. It is clear that the Task Force desires to respect local control, maintain the autonomy of individual educators, provide educators appropriate professional development and ongoing support. Designing a comprehensive assessment system within statutory constraints that also meets the desires listed above is a difficult and complex task, but not an impossible one. Based on these considerable difficulties and complexities, the Task Force was faced with a decision: make recommendations for a single statewide summative assessment to fulfill statutory requirements or make recommendations for a comprehensive assessment system.

The Task Force voted to explore the possibility of a comprehensive assessment system (with a few members expressing some reluctance and reserving judgment). Task Force members asked for

descriptions of how a comprehensive assessment system could work with the desired respect for local control and educator autonomy to then review the descriptions and then decide whether to move forward with a recommendation for a comprehensive assessment system.

## Theories of Action

A discussion of theories of action is important prior introducing the descriptions of how a comprehensive assessment system (or any assessment for that matter) could be designed. Over the past decade, it has become clear that implementation of educational interventions (such as a new assessment or assessment system) is unlikely to achieve intended outcomes unless a thoughtful theory of action is carefully developed. Further, a theory of action helps guide the evaluation of the system or program as well as guiding implementation. A theory of action starts with the following:

- A. A problem statement (including evidence that a problem exists and evidence and/or assumptions about why the problem exists)
- B. Interventions or other mechanisms that are intended to address the problem.
- C. Intended outcomes of the intervention

These three components can be placed in a direct line as shown in Figure 1. In this figure, the interventions are derived from the problem statement, and the intended outcomes are derived from the interventions.

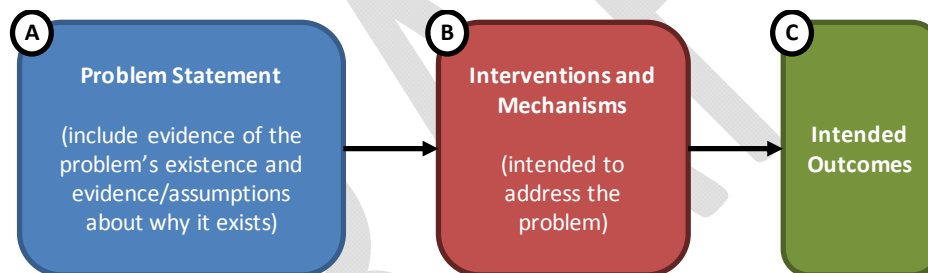


Figure 1. Basic Components of a Naïve Theory of Action.

However, explicitness about these three components is insufficient to support the attainment of intended outcomes. Intended outcomes tend to be long-term outcomes that will result from many years of implementation of interventions. While interventions are intended to ultimately have long-term effects, it is important to identify effects that could be expected to occur more quickly, in the mid-term, and the long-term. This is important because if the immediate effects or mid-term effects are not found, early or mid-course corrections to the interventions are likely necessary. To allow for this type of early and mid-course corrections, the three basic components should be expanded as follows to give the interventions a greater likelihood of long-term success:

- A. A problem statement (including evidence that a problem exists and evidence and/or assumptions about why the problem exists)
- B. Activities and mechanisms (part of the whole package of interventions) that are expected to have short-term (or proximal) effects, along with conditions necessary to achieve the intended short-term outcomes.
- C. Measures of achievement of proximal effects.
- D. Activities and mechanisms that are expected to have mid-term effects if the immediate effects have been observed, along with conditions necessary to achieve the intended mid-term effects.
- E. Measures of mid-term effects.

- F. Measures of long-term effects that are expected to occur if the short-term and mid-term effects are achieved.
- G. Identifying and monitoring of potential unintended negative consequences of the system, policy, or program.

This is in essence an activity in working backward from intended outcomes to the interventions (e.g., activities and mechanisms) and conditions (e.g., supports) that need to be put in place to facilitate changes that need to happen immediately, in the mid-term, and in the long term in order to achieve the desired outcome. It also includes a thoughtful analysis of how the intervention might go wrong and activities, mechanisms, and supports that could be put into place to minimize such unintended negative consequences. This more complete framework for a theory of action can be visualized as shown in Figure 2.

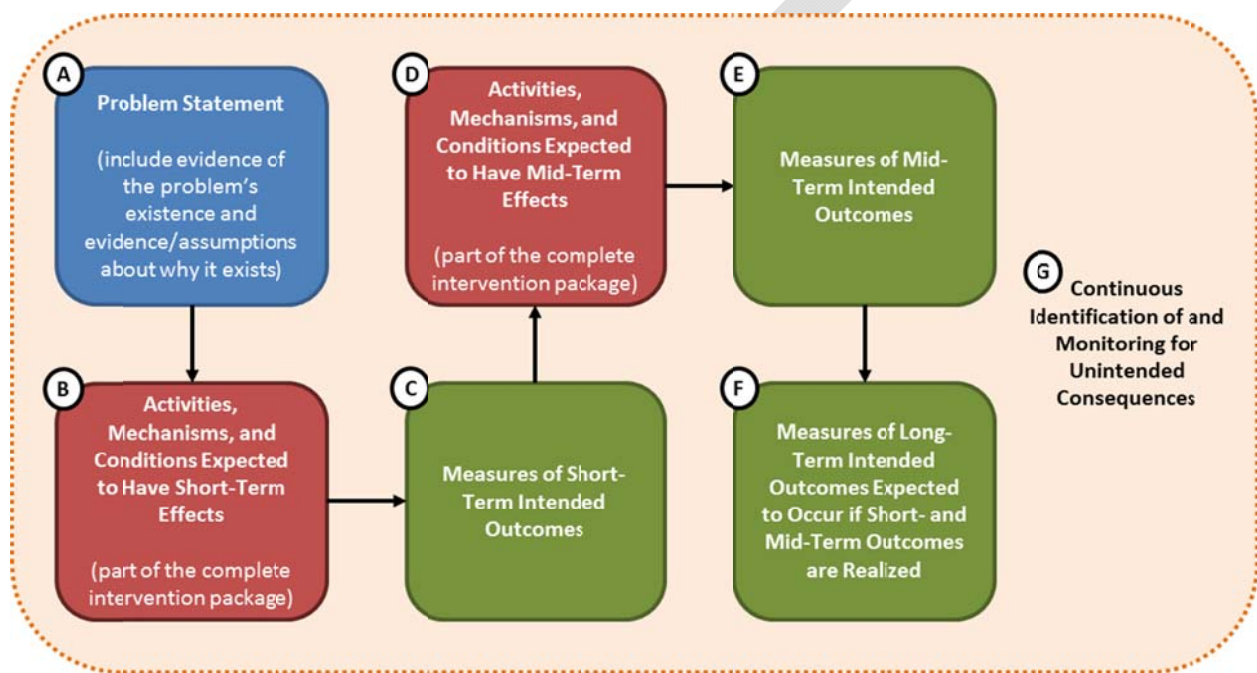


Figure 2. Basic Components of a More Complete Theory of Action.

An instructive example of an incomplete theory of action can be inferred from the implementation of the No Child Left Behind (NCLB) Act of 2001. Some portions of this theory of action were explicit in the development of the statute, and others were inferred (noted in italics) as shown below:

- A. U.S. student achievement is generally low compared to other developed countries, and there are unacceptably disparate outcomes for students of various groups such as those of particular ethnicities/races, those who come from economically disadvantaged backgrounds, males versus females in some areas, English learners, and students with disabilities. Evidence includes data from the Trends in International Math and Science Study (TIMSS), Programme for International Student Assessment (PISA), and National Assessment of Educational Progress (NAEP). *The problem exists because educators are not working hard or smart enough, particularly in economically disadvantaged areas (inferred from the activities, mechanisms, and supports listed in B&D).*
- B&D. Interventions include shining a light on low-achieving districts, schools, and groups of students; requiring schools receiving funding for economically disadvantaged students to submit school improvement plans and to participate in a system of support; and offer parents of “failing” schools free tutoring at district expense and the opportunity to send their children to “non-failing” schools in the same district at district expense.

- C&E. *District, school, and classroom educators will (for various reasons) work harder and figure out how to work smarter as documented in school improvement plans. Not measured.*
- F. Student outcomes will improve for the U.S. as a whole, and for vulnerable student groups in particular, in that 100% of all student groups will become “proficient” by the year 2014 in every grade in at least reading and mathematics in every state.
- G. *Unintended consequences were not considered.*

The lack of a carefully and thoughtfully developed comprehensive theory of action makes it unsurprising that the intended outcomes of NCLB have been realized only to a small degree rather than seeing the considerable improvement envisioned.

The challenge of developing a comprehensive theory of action for the implementation of a single statewide assessment is considerable. The challenge for a comprehensive assessment system is more considerable still. However, in order to achieve the intended outcomes, a sound and comprehensive theory of action will be necessary in order to understand how the components of a comprehensive system are intended to work together, whether the Task Force decides to recommend a single statewide assessment system or a comprehensive assessment system.

### **Characteristics on Which a Comprehensive Assessment System May Vary**

As part of the work to develop vignettes describing various potential comprehensive assessment system designs, it is important to document the various ways characteristics on which a comprehensive assessment system may vary. There are many such characteristics, and it is important to decide at what level on a continuum each characteristic should reside. Table 2 gives a listing of the various characteristics that would need to be addressed in designing a comprehensive assessment system.

Understanding Table 2 requires some explanation. Four types of assessment are given in column 1 of Table 2: classroom-level formative assessment, classroom-level summative assessment, interim assessment, and state-level summative assessment. Others could be described (e.g., non-classroom, non-state-level summative) but they would be very similar to those already in the table.

In column 2 of Table 2, potentially variable aspects of a comprehensive assessment system are listed. In the remaining columns there is room for four levels for every potentially variable aspect. Some aspects do not use all four levels, and while there could be more than four levels for some aspects, the table is limited to four levels for readability and usability. Some of the aspects vary with respect to the level of involvement of the state, and some vary in other ways.

Again, in making decisions about at what level each aspect should be placed in a comprehensive assessment system would need to be consistent with a coherent and thoughtful theory of action about how the intended outcomes will be achieved.

### **VIGNETTE DESCRIBING A COMPREHENSIVE ASSESSMENT SYSTEMS WITH VARIATIONS FOR DIFFERENT LEVELS OF LOCAL CONTROL**

Pending the outcome of the July 28-29, 2015 Task Force meeting.

## REFERENCES

- Braun, H. (ed.) (in press). *Meeting the Challenges to Measurement in an Era of Accountability*. Washington, DC: NCME
- Coladarci, T. (2002). Is It a House... Or a Pile of Bricks? Important Features of a Local Assessment System. *Phi Delta Kappan*, 83(10), pp. 772-774.
- Michigan Department of Education. (2013). Report on Options for Assessments Aligned with the Common Core State Standards. Lansing, MI: Author. Retrieved 6/20/2015 from [http://www.michigan.gov/documents/mde/Common\\_Core\\_Assessment\\_Option\\_Report\\_441322\\_7.pdf](http://www.michigan.gov/documents/mde/Common_Core_Assessment_Option_Report_441322_7.pdf).
- Perie, M., Marion, S., Gong, B., Wurtzel, J. (2007). *The Role of Interim Assessments in a Comprehensive Assessment System: A Policy Brief*. Washington, DC & Dover, NH: Achieve, Inc.; National Center for the Improvement of Educational Assessment & The Aspen Institute. Retrieved 6/20/2015 from [http://www.nciea.org/publication\\_PDFs/PolicyBriefFINAL.pdf](http://www.nciea.org/publication_PDFs/PolicyBriefFINAL.pdf).
- Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science*, 18(2), pp. 119-144.

**APPENDIX A: CHARACTERISTICS ON WHICH VARIOUS PART OF A COMPREHENSIVE ASSESSMENT SYSTEM MAY VARY**

Table A1. *Characteristics on Which the Classroom Formative Assessment Component May Vary.*

| Type of Assessment  | Variable aspect   | Exemplar Descriptions at Various Levels                                    |  |  |  |
|---|---|--|--|--|--|
|   |   | Level 1  | Level 2  | Level 3  | Level 4  |
| <b>Classroom-level Formative Assessment</b><br><br><i>Results are not used for grading and/or evaluation</i><br><br><i>Used to inform and adjust instruction on a frequent /daily basis</i> | Responsibility for Educator Professional Development (PD)/Consistency of PD   | <ul style="list-style-type: none"> <li>• School</li> </ul>                 | <ul style="list-style-type: none"> <li>• District</li> <li>• School may supplement</li> </ul>                | <ul style="list-style-type: none"> <li>• State contracts to develop high-quality PD</li> <li>• Districts opt-in to state-provided PD</li> <li>• School may supplement</li> </ul> | <ul style="list-style-type: none"> <li>• State</li> <li>• District may supplement</li> <li>• School may supplement</li> </ul>  |
|   | Responsibility for Developing & Hosting High-Quality Online Exemplars & Tools | <ul style="list-style-type: none"> <li>• School</li> </ul>                 | <ul style="list-style-type: none"> <li>• District</li> <li>• School may supplement</li> </ul>                | <ul style="list-style-type: none"> <li>• State contracts to develop</li> <li>• Districts opt-in to state-provided PD</li> <li>• School may supplement</li> </ul>                 | <ul style="list-style-type: none"> <li>• State</li> <li>• District may supplement</li> <li>• School may supplement</li> </ul>  |
|   | Personnel Supports Provided   | <ul style="list-style-type: none"> <li>• Time for collaboration</li> </ul> | <ul style="list-style-type: none"> <li>• Time for collaboration</li> <li>• Part-time school coach</li> </ul> | <ul style="list-style-type: none"> <li>• Time for collaboration</li> <li>• Part-time school coach</li> <li>• Part-/full-time district coach</li> </ul>                           | <ul style="list-style-type: none"> <li>• Time for collaboration</li> <li>• Part-time school coach</li> <li>• Part-/full-time district coach</li> <li>• Full-time state coordinator/expert</li> </ul>         |
|   | Responsibility to Fund Personnel Supports                                     | <ul style="list-style-type: none"> <li>• School</li> </ul>                 | <ul style="list-style-type: none"> <li>• District</li> <li>• School may supplement</li> </ul>                | <ul style="list-style-type: none"> <li>• State and district</li> <li>• School may supplement</li> </ul>  | <ul style="list-style-type: none"> <li>• State</li> <li>• District may supplement</li> <li>• School may supplement</li> </ul>  |
|   | Use in Accountability   | <ul style="list-style-type: none"> <li>• None</li> </ul>                   | <ul style="list-style-type: none"> <li>• Fidelity of implementation by teachers</li> </ul>                   | <ul style="list-style-type: none"> <li>• School administrator support of teachers</li> <li>• Fidelity of implementation by teachers</li> </ul>                                   | <ul style="list-style-type: none"> <li>• District administrator support of school educators</li> <li>• School administrator support of teachers</li> <li>• Fidelity of implementation by teachers</li> </ul> |

Table A2. *Characteristics on Which the Classroom Interim and Summative Components May Vary.*

| Type of Assessment  | Variable aspect   | Exemplar Descriptions at Various Levels                                  |   |  |   |
|---|---|--|---|--|---|
|   |   | Level 1  | Level 2   | Level 3  | Level 4   |
| <b>Classroom-level Interim and Summative Assessment</b><br><br><i>Results are used for grading.</i><br><br><i>Though different from formative assessment, the variable aspects are identical with the exception of use for student grades</i> | Responsibility for Educator Professional Development (PD)/Consistency of PD   | <ul style="list-style-type: none"> <li>School</li> </ul>                 | <ul style="list-style-type: none"> <li>District</li> <li>School may supplement</li> </ul>                               | <ul style="list-style-type: none"> <li>State contracts to develop high-quality PD</li> <li>Districts opt-in to state-provided PD</li> <li>School may supplement</li> </ul> | <ul style="list-style-type: none"> <li>State</li> <li>District may supplement</li> <li>School may supplement</li> </ul>   |
|   | Responsibility for Developing & Hosting High-Quality Online Exemplars & Tools | <ul style="list-style-type: none"> <li>School</li> </ul>                 | <ul style="list-style-type: none"> <li>District</li> <li>School may supplement</li> </ul>                               | <ul style="list-style-type: none"> <li>State contracts to develop</li> <li>Districts opt-in to state-provided PD</li> <li>School may supplement</li> </ul>                 | <ul style="list-style-type: none"> <li>State</li> <li>District may supplement</li> <li>School may supplement</li> </ul>   |
|   | Personnel Supports Provided   | <ul style="list-style-type: none"> <li>Time for collaboration</li> </ul> | <ul style="list-style-type: none"> <li>Time for collaboration</li> <li>Part-time school coach</li> </ul>                | <ul style="list-style-type: none"> <li>Time for collaboration</li> <li>Part-time school coach</li> <li>Part-/full-time district coach</li> </ul>                           | <ul style="list-style-type: none"> <li>Time for collaboration</li> <li>Part-time school coach</li> <li>Part-/full-time district coach</li> <li>Full-time state coordinator/expert</li> </ul>  |
|   | Responsibility to Fund Personnel Supports                                     | <ul style="list-style-type: none"> <li>School</li> </ul>                 | <ul style="list-style-type: none"> <li>District</li> <li>School may supplement</li> </ul>                               | <ul style="list-style-type: none"> <li>State and district</li> <li>School may supplement</li> </ul>  | <ul style="list-style-type: none"> <li>State</li> <li>District may supplement</li> <li>School may supplement</li> </ul>   |
|   | Use in Accountability   | <ul style="list-style-type: none"> <li><i>Student grades</i></li> </ul>  | <ul style="list-style-type: none"> <li><i>Student grades</i></li> <li>Fidelity of implementation by teachers</li> </ul> | <ul style="list-style-type: none"> <li><i>Student grades</i></li> <li>School administrator support of teachers</li> <li>Fidelity of implementation by teachers</li> </ul>  | <ul style="list-style-type: none"> <li><i>Student grades</i></li> <li>District administrator support of school educators</li> <li>School administrator support of teachers</li> <li>Fidelity of implementation by teachers</li> </ul> |



Table A3. *Characteristics on Which District Assessment Components May Vary.*

| Type of Assessment   | Variable aspect  | Exemplar Descriptions at Various Levels   |  |   |   |
|--|--|---|--|---|---|
|  |  | Level 1   | Level 2  | Level 3   | Level 4   |
| <b>District Assessment Systems</b><br><br><i>May include local assessments used for summative purposes such as student grading and program/curriculum evaluation</i> | Responsibility for Development   | <ul style="list-style-type: none"> <li>District</li> </ul>  | <ul style="list-style-type: none"> <li>District consortium</li> </ul>  | <ul style="list-style-type: none"> <li>State</li> </ul>   | <ul style="list-style-type: none"> <li>State consortium</li> </ul>  |
|  | Responsibility for (Online) Administration Tools   | <ul style="list-style-type: none"> <li>District</li> </ul>  | <ul style="list-style-type: none"> <li>District consortium</li> </ul>  | <ul style="list-style-type: none"> <li>State</li> </ul>   | <ul style="list-style-type: none"> <li>State consortium</li> </ul>  |
|  | Responsibility to Proctor  | <ul style="list-style-type: none"> <li>Teacher</li> </ul>   | <ul style="list-style-type: none"> <li>Different teacher in same school</li> </ul>   |   |   |
|  | Responsibility to Hand-score as Needed   | <ul style="list-style-type: none"> <li>Teacher</li> </ul>   | <ul style="list-style-type: none"> <li>Different teacher in same school</li> </ul>   | <ul style="list-style-type: none"> <li>Different teacher in same district, but not in same school</li> </ul>  | <ul style="list-style-type: none"> <li>Teacher in different district or contractor-provided scorer</li> </ul>   |
|  | Comparability  | <ul style="list-style-type: none"> <li>Within-district</li> </ul>   | <ul style="list-style-type: none"> <li>Within-state</li> </ul>   | <ul style="list-style-type: none"> <li>Across states</li> </ul>   |   |
|  | Specificity of coverage  | <ul style="list-style-type: none"> <li>Broad content area only</li> </ul>   | <ul style="list-style-type: none"> <li>Broad content area and sub-scores</li> </ul>  | <ul style="list-style-type: none"> <li>Broad content area (and possible sub-scores) in K-8</li> <li>Specific to a small number of common courses in HS</li> </ul>   | <ul style="list-style-type: none"> <li>Broad content area (and possible sub-scores) in K-8</li> <li>Flexible to match traditional &amp; non-traditional HS courses (e.g., integrated courses such as career/tech ed courses)</li> </ul> |
|  | Measures provided  | <ul style="list-style-type: none"> <li>Achievement score</li> </ul>   | <ul style="list-style-type: none"> <li>Achievement score</li> <li>Performance category</li> </ul>  | <ul style="list-style-type: none"> <li>Achievement score</li> <li>Performance category</li> <li>Growth score</li> </ul>   | <ul style="list-style-type: none"> <li>Achievement score</li> <li>Performance category</li> <li>Growth score</li> <li>Growth category</li> </ul>  |
|  | Use in Accountability  | <ul style="list-style-type: none"> <li>None</li> </ul>  | <ul style="list-style-type: none"> <li>Student grades</li> </ul>   | <ul style="list-style-type: none"> <li>Student grades</li> <li>As part of assessment data in school accountability</li> </ul>   | <ul style="list-style-type: none"> <li>Student grades</li> <li>Roll up across year; replace yearly summative assessment</li> <li>Likely requires the same assessment across districts</li> </ul>  |
|  | Flexibility to Match Local Instructional Units   | <ul style="list-style-type: none"> <li>State contracts to create instructional units</li> <li>Assesses only the state-defined units</li> <li>Only administered in narrow windows</li> </ul> | <ul style="list-style-type: none"> <li>State contracts to create model instructional units</li> <li>Assesses only the state-defined model units</li> <li>Administered at any time</li> </ul>   | <ul style="list-style-type: none"> <li>State contracts to create model instructional units</li> <li>Assesses state-defined units or customized units</li> <li>Administered at any time</li> </ul>   | <ul style="list-style-type: none"> <li>District, school or teacher responsibility only</li> </ul>   |
|  | Availability of Item Types   | <ul style="list-style-type: none"> <li>Multiple choice</li> </ul>   | <ul style="list-style-type: none"> <li>Multiple choice</li> <li>Constructed response</li> </ul>  | <ul style="list-style-type: none"> <li>Multiple choice</li> <li>Constructed response</li> <li>Technology enhanced</li> </ul>  | <ul style="list-style-type: none"> <li>Multiple choice</li> <li>Constructed response</li> <li>Technology enhanced</li> <li>Performance tasks</li> </ul>   |
|  | Responsibility for Auditing Local Systems (if not directly integrated with state summative assessment) | <ul style="list-style-type: none"> <li>None (inappropriate if used for accountability)</li> </ul>   | <ul style="list-style-type: none"> <li>Peers from other districts</li> </ul>   | <ul style="list-style-type: none"> <li>Peers from other districts</li> <li>Overseen by state</li> </ul>   | <ul style="list-style-type: none"> <li>Experts contracted by the State</li> </ul>   |
|  | Integration with State Summative Assessment  | <ul style="list-style-type: none"> <li>None</li> </ul>  | <ul style="list-style-type: none"> <li>Item/task templates (with exemplars) provided by state</li> <li>Templates consistent with state summative assessment</li> <li>Schools/districts opt in to any degree desired</li> <li>Schools/districts write their own tasks for templates, if opted in</li> </ul> | <ul style="list-style-type: none"> <li>Item/task templates (with exemplars) provided by state</li> <li>Templates consistent with state summative assessment</li> <li>Schools/districts write their own tasks from templates</li> <li>State helps analyze relationship of local and state tests</li> </ul> | <ul style="list-style-type: none"> <li>Items/tasks come from the same pool as state summative</li> <li>Results are reported on the same scale as state summative</li> </ul>   |

Table A4. *Characteristics on Which the State Summative Assessment Component May Vary.*

| Type of Assessment   | Variable aspect   | Exemplar Descriptions at Various Levels  |   |   |   |
|--|---|--|---|---|---|
|  |   | Level 1  | Level 2   | Level 3   | Level 4   |
| <b>State-level Summative</b><br><br><i>Used only for summative purposes such as auditing, accountability, and evaluation</i> | Frequency   | <ul style="list-style-type: none"> <li>Once yearly</li> </ul>  | <ul style="list-style-type: none"> <li>Beginning and end of year/semester To measure within-year/semester growth</li> </ul>   | <ul style="list-style-type: none"> <li>As desired through year (rolled up across occasions)</li> </ul>  |   |
|  | Comparability   | <ul style="list-style-type: none"> <li>Across state</li> </ul>   | <ul style="list-style-type: none"> <li>Across multiple states</li> </ul>  |   |   |
|  | Content Coverage  | <ul style="list-style-type: none"> <li>State content standards</li> </ul>  | <ul style="list-style-type: none"> <li>State content standards</li> <li>Additional standards</li> </ul>   |   |   |
|  | Specificity of Coverage   | <ul style="list-style-type: none"> <li>Broad content area only</li> </ul>  | <ul style="list-style-type: none"> <li>Broad content area and sub-scores</li> </ul>   | <ul style="list-style-type: none"> <li>Broad content area (and possible sub-scores) in K-8</li> <li>Specific to a small number of common courses in HS</li> </ul>             | <ul style="list-style-type: none"> <li>Broad content area (and possible sub-scores) in K-8</li> <li>Flexible to match traditional &amp; non-traditional HS courses (e.g., integrated courses such as career/tech ed courses)</li> </ul> |
|  | Measures provided   | <ul style="list-style-type: none"> <li>Achievement score</li> </ul>  | <ul style="list-style-type: none"> <li>Achievement score</li> <li>Performance category</li> </ul>   | <ul style="list-style-type: none"> <li>Achievement score</li> <li>Performance category</li> <li>Growth score</li> </ul>   | <ul style="list-style-type: none"> <li>Achievement score</li> <li>Performance category</li> <li>Growth score</li> <li>Growth category</li> </ul>  |
|  | Prediction  | <ul style="list-style-type: none"> <li>College/career readiness only</li> </ul>  | <ul style="list-style-type: none"> <li>Proficiency in the next grade only</li> </ul>  | <ul style="list-style-type: none"> <li>College/career readiness</li> <li>Proficiency in the next grade</li> </ul>   |   |
|  | Purpose   | <ul style="list-style-type: none"> <li>Audit local assessments</li> </ul>  | <ul style="list-style-type: none"> <li>Estimate local performance from sample</li> <li>Local use for program/curriculum evaluation</li> </ul>   | <ul style="list-style-type: none"> <li>Census of local performance from all students</li> <li>Local use for program/curriculum evaluation</li> </ul>                          |   |
|  | Sampling  | <ul style="list-style-type: none"> <li>Randomly assign schools to a content area</li> <li>All students take 1-2 hours of one full content area assessment</li> <li>Student test randomly drawn from entire pool</li> </ul> | <ul style="list-style-type: none"> <li>Randomly assign schools to a content area</li> <li>All students take one full content area assessment</li> </ul>   | <ul style="list-style-type: none"> <li>All students take full assessment in all content areas</li> </ul>  | <p><b>Note:</b> <u>Level 1 is inconsistent with current state and federal law.</u></p>  |
|  | Available Reports   | <ul style="list-style-type: none"> <li>State-level estimates in all content areas</li> <li>Flag inconsistent district/school reports</li> </ul>  | <ul style="list-style-type: none"> <li>State estimates in all content areas</li> <li>District estimates in available content areas</li> <li>School estimates in available content areas</li> <li>Flag inconsistent district/school reports</li> </ul> | <ul style="list-style-type: none"> <li>State census in all content areas</li> <li>District census in all content areas</li> <li>School census in all content areas</li> </ul> |   |
|  | Local (school, district) Accountability                           | <ul style="list-style-type: none"> <li>Trigger detailed audit of local systems</li> </ul>  | <ul style="list-style-type: none"> <li>Use only content areas tested in last three years</li> <li>Must always include math and ELA</li> </ul>   | <ul style="list-style-type: none"> <li>Use all content areas</li> </ul>   |   |
| Availability of Item Types   | <ul style="list-style-type: none"> <li>Multiple choice</li> </ul> | <ul style="list-style-type: none"> <li>Multiple choice</li> <li>Constructed response</li> </ul>  | <ul style="list-style-type: none"> <li>Multiple choice</li> <li>Constructed response</li> <li>Technology enhanced</li> </ul>  | <ul style="list-style-type: none"> <li>Multiple choice</li> <li>Constructed response</li> <li>Technology enhanced</li> <li>Performance tasks</li> </ul>                       |   |