Assessment Workgroup Toolkit

Document	Tool Description	Possible Uses
Resource: Glossary of Assessment Terms 	A shared understanding of key terms is an essential component of ensuring success when working on assessment projects. This chart fleshes out fifteen key assessment terms by providing a list of characteristics, a definition, an example, and a non-example for each term.	As a stand-alone document, the glossary can serve as a reference point or resource for members of a community. This glossary can also be used in conjunction with the activity described below to help educators establish a common understanding of terms.
Activity: Developing a Deeper Understanding of Assessment Terms 	This activity, which uses the Frayer Model, is designed to help participants to deepen their shared understanding of key terms by exploring characteristics, definitions, examples and non-examples.	Education leaders and PD facilitators can use the first part of this activity to assess participants' understanding of terms and to determine who in the group can be asked to provide specific insights and examples to explain terms. They can then use the second part of the activity to help participants to explore ways to incorporate examples into their curriculum design process and classroom practice.
Resource: Taxonomy of Outcomes 	These seven outcomes were developed by Learner-Centered Initiatives, Ltd. as a framing device to help educators to establish a vision for their graduates and to link that vision to various school/district or community programs.	Groups can work collectively on this chart to ascertain patterns in valued outcomes and to explore additional ways to meet the desired outcomes. Leaders can look for the presence of the outcomes in their school or district mission and vision. Curriculum designers can use it to deepen the alignment between these outcomes, standards, and assessments.
Activity: Developing Criteria for Assessing Recommendations 	This activity provides eleven criteria to consider when adopting or rejecting a recommendation that will impact a system or the different constituencies in a system.	School or district leadership teams that are working through recommendations may find the activity useful for assessing and narrowing down proposals or otherwise making their thinking transparent. The tool is flexible enough for participants to revise or add additional criteria.
Activity: Exploring Intended Results, Unintended Consequences, and Influence on Stakeholders 	This activity is designed to scaffold participants' systems thinking skills around what could or might happen as a result of a proposed action or recommendation. The activity contains several guiding questions as well as sample responses.	Individuals or groups considering an action or change can find it helpful to identify what it is they want to happen and what might happen as a result of adopting the recommendation. Participants can review the examples before considering the implications of their recommendations.

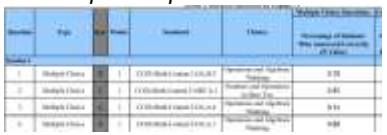
Glossary of Key Assessment Terms

A shared understanding of key terms is an essential component of ensuring success when working on assessment projects. As a stand-alone document, this glossary can stand a as a reference point or resource for members of a community or can be used in conjunction with the human Frayer model activity to help educators think through a variety of terms related to assessment.

Term	Definition	Characteristics	Example	Non-Example
1. Anchor	The representative products or performances used to illustrate each level on a scoring scale. The product or performance aligned to the top or highest level is called the exemplar. (Arter & McTighe, 2003)	They're work products, portions of products, or recordings of demonstrations. Students can use them to self-assess their work against qualities associated with each level of a scoring scale. They can also be used to clarify criteria for scoring for students to determine how to improve their work.	Student writing sample that illustrates common errors reflected in the lowest level of a source citation rubric.	Student work hanging on a bulletin board marked with a gold star but with no indication as to what makes it quality or exemplary. *Explanation*: *While the work may be a model, it needs a clear connection to the scoring scale.*
2. Assessment	The purposeful and strategic collection of direct, observable, tangible evidence of student learning throughout the learning process, used to inform teaching and learning (Martin-Kniep, 2013). Any systematic basis for making inferences about characteristics of people, usually based on various sources of evidence (Arter & McTighe, 2003).	This umbrella term includes tests, exams, products, performances, demonstrations, and processes. They're usually accompanied by a set of directions or guidelines, some sort of scoring criteria, and a designated time period for completion. They are used for different purposes at different moments in time during the assessment process; diagnostic (before), formative (during), and summative (after).	Essay, multiple choice test, portfolio, dance recital, Regents exam.	Unrecorded, undocumented class discussion. *Explanation*: *While useful for informing teaching and student interactions, absent of documentation, it's difficult to make purposeful decisions.*
3. Assessment Bias	The presence of one or more items in an assessment that differentially affects the performance of different groups of test takers and consequently the reliability and validity of interpretations. (AERA, APA, & NCTM, 2014)	It contains content that insults, irritates, or causes pain to students (Popham, 2006) because of the student's personal characteristics, resulting in a lack of fairness to students.	A math teacher presents a series of tasks around football games that require knowing *YAC* but doesn't explain what the abbreviation means.	A student accidentally throws away the directions to a task, so misses several steps in a sequence, and therefore, does not do well on the final product. *This occurred due to an individual student's action, not because of the design of the assessment.*

Term	Definition	Characteristics	Example	Non-Example
4. Authentic Assessment	Assessments that engage students in real-life problems or tasks for an audience who cares about or has a stake in what students create or do (Martin-Kniep, 2000)	Students are given, or identify themselves, a real purpose and audience for their work. They search for in-depth understanding and engage in high levels of thinking, working alone and with others often across content areas.	Students collect data around the flow of traffic in the high school cafeteria as well as perceptual data related to bullying during the lunch periods. They develop a proposal for the Board of Education to address the schedule and structure of lunch periods to increase efficiency and inclusivity.	Students write a multi-page paper explaining how bills are introduced to Congress. *Asking students to describe something that happens in the world, without engaging in it, is a sign that the task lacks a real purpose or audience.*
5. Constructed-Response Items or Tasks	Prompts in which the student must create their own answer or products rather than choosing from a specific set of possible answers. (AERA, APA, & NCTM, 2014).	The student does most of the work generating a response or answer. The teacher provides the framework or schema and time for student to work on the task.	Students write responses to the essential question, "When does a sound become music?" at the beginning of a unit.	Students are given 10 questions, each followed by four choices, and must pick the word or phrase that best answers the question. *The teacher has provided the prompt and response.*
6. Dispositions	Abiding tendencies or habits of mind that reflect the values, commitments, practices, and ethics that influence behaviors and actions (Martin-Kniep, 2008).	These are the affective (soft-skills) dimensions of individuals (e.g., flexibility, persistence, open-mindedness, commitment to understanding).	Students demonstrate *Courage and Initiative* during a civics conversation about a controversial topic. When self-assessing, they consider the class norms related to courage: • We explore assumptions and discuss issues to support a productive learning experience. • We take initiative to speak up about what is necessary to achieve the workshops' goals.	Students articulate the differences among religions. *Knowing factual knowledge, while important, is a different set of skills and abilities.*
7. Formative Assessment	An assessment process used by teachers and students during instruction that provides feedback to adjust ongoing teaching and learning (AERA, APA, & NCTM, 2014).	A common analogy compares them to the chef tasting the soup while it's still in the kitchen; the assessments at this point are low-stakes since the student is still learning the content. They're un-graded.	Students respond to a quick series of short response questions using their phones at the end of a lesson. The teacher uses student responses to form small groups the next day and gives results to students, so they can identify their own strengths and weaknesses.	Students present a speech in front of audience. The speech is recorded and judged by a panel from the community. Their performance is evaluated and scored, and one speaker is declared the winner. *Since it comes at the end, and is judged, the student is no longer practicing.*

Term	Definition	Characteristics	Example	Non-Example
8. Performance Assessment	An assessment that requires students to do something with their knowledge (Brookhart and Nitko, 2011). They are tasks that require students to apply their knowledge, skills, and strategies by creating a response or a product (Rudner & Boston, 1994; Wiggins, 1989).	The assessment is typically more complicated than writing a short response or answering basic questions. Students are usually required to think deeply, reflect and revise, and work towards submitting their best work.	Students create a newspaper around the content they are studying or learning about that parallels content learned by younger students. They generate an op-ed, several news articles, an editorial cartoon, and classified ads. Younger students read the paper and provide feedback on readability.	Students answer 50 multiple choice questions. *The task doesn't require students to think deeply or revise their work.*
9. Reliability	The degree to which an assessment is consistent or trustworthy. Generally, documented or captured via three types of evidence such as internal consistency, stability, and scorer consistency (Popham, 2006).	When a speaker refers to this trait for an assessment, they are typically speaking about how well it can be trusted. Evidence related to the trait varies based on the nature and scope of an assessment; that is, a test given to 25,000 students requires different evidence than one given to 25.	From a test analysis: "Cronbach's alpha and Feldt-Raju ranged from 0.89 to 0.93. All were at least .89 across all grades and both subjects, which is a good indication that the tests are acceptable."	A student notices that several items on the test come from content that's unfamiliar. A review of the test map shows the content wasn't explicitly taught to students. *The student has noticed a lack of alignment, which does contribute to trustworthiness, but is about something else.*
10. Rubric	A tool that defines and differentiates levels related to attributes of quality or performance (LCI, 2000).	All versions of this tool have levels, dimensions, and descriptors. Levels indicate the range of performance; dimensions are the criteria used to judge a demonstration, process, or product; and descriptors define the dimensions at different levels.		*This is a Likert Test. It helps the student understand levels of improvement but does not describe the quality at different levels.* 
11. Standardized Assessment	When an assessment's directions, administration conditions, and scoring follow the same procedures and structures for all students (AERA, APA, & NCTM, 2014).	Although generally used to refer to large-scale, federally-mandated tests or large-scale college admission tests, it technically refers to any assessment that looks similar for all students.	A teacher writes and gives the same final exam to all 110 of her students and scores all the exams using the same answer key. She provides versions in Spanish for her ELLs and accommodations for students with IEPs.	Students design their own portfolio project and set their own submission deadlines. A teacher works with each student to set expectations of quality. Students submit work on a rolling basis, depending on their goals. *The conditions and scoring structure are different for each student.*

Term	Definition	Characteristics	Example	Non-Example
12. Selected-Response Prompts	The student must select from a list of possible answers (Popham, 2006).	The teacher does most of the work generating a response or answer. The student is responsible for identifying the correct answer.	Students are given 10 questions, each followed by four choices, and must pick the word or phrase that best answers the question.	Students write a brief response to the essential question, "When does a sound become music?" at the beginning of a unit. *This task requires students to construct their own response.*
13. Standard Error of Measurement	A statistic that indicators (non) consistency in the scores generated by an assessment. It provides a value that reflects how much of a spread there would be in a student's scores if the student were to take the test multiple times (AERA, APA, & NCTM, 2014).	The statistic is directly related to reliability; the higher the reliability, the lower this statistic.	From a test analysis: "They ranged from 2.75 to 3.91 across subjects, grades, and the two methods of estimation, which is reasonable."	Two students in the same study group use the same handouts and structures to study. One student gets all but one question correct, and one student fails the test. *Different test results for different students caused by differences in their depth of understanding are not a function of the test design itself.*
14. Test Map/Blueprint	A document that articulates the alignment between each item or task, the design process, scoring criteria or expectations, and/or p-values, if known.	At a minimum, the document should detail alignment between assessment prompts, items, or questions and standards. This alignment is essential to making claims related to validity.	*Test map example* 	An assessment has English 9 Shakespeare Project written in the footer. Students know they're working on the Shakespeare Project but neither they or their teachers can articulate the learning outcomes or standards for this assignment. *Although the general topic and content have been identified, it's difficult to link the task to individual standards.*
15. Validity	The degree to which an inference about an assessment is accurate (Popham, 2006).	When a speaker refers to this trait, they are typically speaking about the alignment between what the assessment claims to measure and what it actually measures.	A student notices that several items on the test come from content that's unfamiliar. A review of the test map shows that the content wasn't taught to students. It's difficult to be confident in the inferences drawn from student's performance on the test.	From a test analysis: "Cronbach's alpha and Feldt-Raju ranged from 0.89 to 0.93. All were at least .89 across all grades and both subjects, which is a good indication that the tests are acceptable." *These statistics speak to the trustworthiness of the assessment.*

Activity for Developing a Deeper Understanding of Assessment Terms

This activity is best positioned as an interactive experience that enables pairs or small groups of individuals to share and discuss their responses. It can be used to gain a deeper understanding of any term or phrase and does not need to be limited to assessment terms. The activity facilitator(s) should provide the terms on this handout as well as create a poster for each term. In addition, they'll need to create separate cards with the definition, characteristic, example, and non-example for each term and distribute them randomly to participants. Participants should then work together to complete a poster with the correct components.

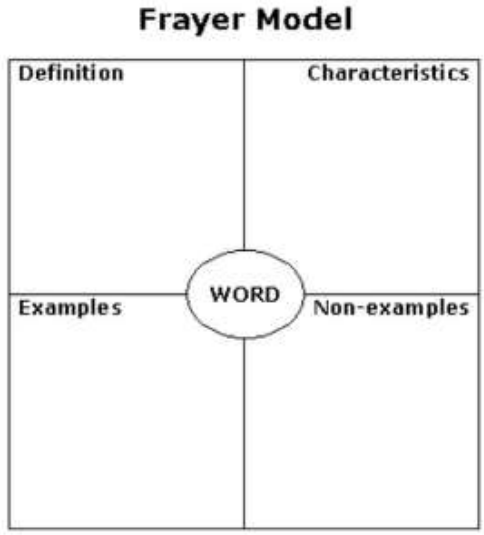
Part One: Review the list of terms below. Before we begin, take a moment to reflect on your level of understanding related to the term. Put a checkmark in the column that best reflects your understanding of the term.

Term	This term is unfamiliar to me.	I've heard of this term before.	I know this term and can define it.	I know this term and can teach it to others.

Take a moment to consider the card you were given when you arrived. Your card is related to one of the terms in the list above and somewhere in the room are people who hold cards related to the same term as yours. As a group, you complete a Frayer Model[1] for the term. Posters containing the term can be found around the room.

Part Two

Large group: When the activity begins, work with others in the room to locate the other three components of your term. When you've located the other members of your group, make your way to your poster and attach your parts.



[1] Frayer, D., Frederick, W. C., and Klausmeier, H. J. (1969). *A Schema for Testing the Level of Cognitive Mastery*. Madison, WI: Wisconsin Center for Education Research.

Taxonomy of Learning Outcomes

This resource can be used by a teachers and leaders from schools or districts to ascertain patterns in the outcomes that are valued. Leaders could also look for the presence of those outcomes in their school or district mission and vision. Additionally, curriculum designers can use it to deepen the alignment between these outcomes, standards, and assessments.

Outcomes for learning can be described in a variety of ways. As you review the examples below, consider how your community expresses that you value this type of knowledge.

1. Factual (content) Knowledge (example: ***geographic features*** *– 5th grade SS project on maps;* ***structure of a cell*** *– 7th grade Science activity)* Example from my school, district, community:
2. Conceptual Knowledge (example: ***beauty*** *– elementary unit on fairy tales; change – HS science department unit, "Can/should* ***change*** *be stopped?" looking at impact on local water table)* Example from my school, district, community:
3. Procedural Knowledge (example: ***how to balance a checkbook*** *– Economics class;* ***how to write code*** *–Code.org lesson during Genius Hour)* Example from my school, district, community:
4. Meta-Cognitive Knowledge (example: *student* ***self-assessment on statements about health*** *before starting a new unit – middle school PE;* ***student reflection on study habits*** *after mid-term – HS math)* Example from my school, district, community:
5. Thinking Processes Skills and Abilities (example: *"It Says… and so… I know…" graphic organizer to* ***draw inferences*** *– 8th grade ELA;* ***comparing*** *claims and counterclaims – elementary Social Studies)* Example from my school, district, community:
6. Subject-specific Skills, Abilities, and Practices (Example*:* ***playing a musical instrument*** *– orchestra;* ***using the scientific method*** *– middle school science;* ***reading a map –*** *Social studies)* Example from my school, district, community:
7. Dispositions or habits of mind (Example: ***perspective-taking*** *– students read first-person reflections about an event from multiple perspectives in unit on The Great Depression; open-mindedness – students document how a text aligns to Teaching Tolerance's Standards for Social Justice related to* ***courage****).* Example from my school, district, community:

Preliminary Checklist for Quality Recommendations

Using established criteria to make decisions ensures greater credibility, viability and transparency. School or district leadership teams can use this tool to establish the criteria for a quality recommendation before they assess and narrow down proposals. that are working through recommendations may find the activity useful for assessing and narrowing down recommendations or otherwise making their thinking transparent. The tool is flexible enough for participants to revise or add additional criteria.

Background: One of the workgroup goals was to: *make specific, credible, and viable recommendations to their school community.* Part of meeting that goal means reconciling, as a group, the criteria of a quality recommendation. This activity focuses on a generic recommendation (Recommendation X).

Directions:

1. Review the various criteria, including the first three examples. Then, select 3-4 of the other criteria and provide your own rationale for those criteria.
2. As a group, decide on the criteria for accepting recommendations.
3. Use the established criteria to review recommendations and to explain why they are or are not being accepted.

Criteria *Recommendation X should be accepted because it …*	**Why is this criterion for a quality recommendation important?**
1. … will lead to improved student learning.	*It's important that we ensure all recommendations connect back to student learning. This criteria will help us ensure we go back to student learning.*
2. … will do no harm.	*This criteria is an important reminder to attend to unintended consequences and the ethics of education.*
3. … will lead to improved teacher practice.	*Similar to the first criteria, this will help remind us to ensure the recommendation has an impact on an essential aspect of education.*
4. … can be implemented within a reasonable period (2-3 years).	
5. … will lead to improvement in other areas (curriculum design, etc.).	
6. … validates quality teachers' practice.	
7. … promotes standard attainment.	
8. … will increase parental goodwill/confidence.	
9. … will improve communication between stakeholders.	
10. … will increase the public's confidence in NYS schools.	
11. … is cost effective.	
Other?	

Exploring Intended Results, Unintended Consequences, and Influence on Stakeholders

Individuals or groups considering an action or change can find it helpful to identify what it is they want to happen as well as what might happen as a result of adopting the recommendation. Participants can review the examples before considering the implications of their recommendations and then use that work to revise their plans and recommendations..

Question that informed this activity: *Are we thinking ahead about how this would look in order to avoid unintended consequences?*

Part 1: Exploring Intended Results and Unintended Consequences

Recommendation: *Teachers should evaluate their current classroom assessment system by completing an assessment inventory. Administrators should support this work by facilitating data collection, organization and analysis.*	
Intended result: What we expect will happen as a result of taking action based on this recommendation.	*We expect teachers will be able to determine the quality, focus, and uses of the assessments in their classrooms. We expect school leaders to learn more about teacher assessment practices in each classroom, department, or grade level.*
Unintended positive consequences: Positive outcomes which may occur and which we did not anticipate.	• *Teachers will revisit the outcomes they value in student learning.* • *Teachers may be able to eliminate assessments that are not serving them well or may modify assessments to improve them.* • *Teachers may uncover opportunities for the use of assessments they had not previously considered.* • *Teachers will identify changes that are within their control.* • *Administrators will get a better sense of patterns across their school.* • *Teachers and administrators will learn more about student assessment experiences across the school.*
Unintended negative consequences: Negative outcomes which may occur and which we did not anticipate.	• *Time devoted to evaluating assessments may take away time from instruction.* • *Teachers may not know what to do with the results of their audit.* • *Teachers may compare their results.* • *Principals may compare the teachers' results.* • *Without quality criteria for the audit, teachers may end up comparing apples to oranges, and compromising the analysis.* • *Teachers may feel it's extra work or that there's a "right" answer or ideal distribution of assessments.*

Part 2: Influence on Stakeholders

Recommendation: *Teachers should evaluate their current classroom assessment system by completing an assessment inventory. Administrators should support this work by facilitating data collection, organization and analysis.*	
How might students be impacted by this?	*Students may end up with greater access to a more balanced assessment system.*
	Students may take more or fewer assessments during the audit process.
How might parents be impacted by this?	*Parents may end up with greater clarity about the different assessments in the school and their purposes.*
	Parents may feel discomfort as the assessment system begins to look less familiar.
How might support staff be impacted by this?	*Support staff may develop a better understanding of why teachers use the assessments they use.*
	Support staff may be burdened with some of the audit tasks.
How might test publishers be impacted by this?	*Test publishers may have a greater sense of what is needed by schools.*
	Schools might purchase fewer commercially-produced assessments.