NEW YORK STATE EDUCATION DEPARTMENT

APPR Assessment Guidance:

Guidance for Districts and BOCES on Selecting Third Party Assessments for use with Students in K-2 for Teacher and Principal Evaluations

March 2014

Table of Contents

Section 1. Introduction

Overview of the Guidance

The following Annual Professional Performance Review (APPR) Assessment Guidance document is intended to provide information to districts and BOCES on how to evaluate and identify existing third party assessments that may be used to assess students in grades kindergarten through grade two (K-2) for the purposes of educators' annual professional performance reviews in New York State.

In order to implement the emergency regulations first adopted by the Board of Regents at its February 2014 meeting, the following actions have been taken by the Department:

- *Removal of K-2 Third Party Assessments from the State-approved List:* The Department has removed all third party assessments for use in kindergarten-grade two **only** from the state-approved list on our website (see: http://usny.nysed.gov/rttt/teachers-leaders/assessments/). Effective March 2, 2014, the Department will no longer use the RFQ process to review and publish a list of any state-approved third party assessments for students in K-2 **only**. Instead, districts and BOCES who determine they wish to continue to use, or use a newly selected, third party assessment (developed by any vendor, third party, or other comparable entity) for APPR purposes must now follow the guidance in this document to ensure the third party assessment is consistent with Education Law §3012-c and Subpart 30-2 of the Rules of the Board of Regents, including the prohibition against the administration of traditional standardized assessments to these students.

- *APPR Plan Certification:* Superintendents, district superintendents, or chancellor must now certify in their APPR plan that any third party assessment (developed by any vendor, third party, or other comparable entity) being used for APPR purposes for students in kindergarten – grade two is consistent with this guidance document and is not a traditional standardized assessment.

This guidance is organized into four main sections:

- Section 1 offers an overview and background information about New York State's use of student assessments for the purposes of APPR.

- Section 2 provides a template consistent with regulation 30-2.8 and adapted from the Department's Assessment Request for Qualifications (RFQ) to be used by districts and BOCES when evaluating third party assessments.

- Section 3 provides guidance in identifying K-2 *non-traditional standardized assessments* that may be used for the purposes of APPR based on the New York State Education Department's (NYSED) recommended design principles.

- Section 4 provides explanations of key concepts that are often used in educational assessment.

- The appendix includes a list of resources.

For further guidance on New York State's APPR system, please see:
http://www.engageny.org/sites/default/files/resource/attachments/appr-field-guidance.pdf

If you have further questions regarding New York State's APPR system, generally, please contact educatoreval@mail.nysed.gov. Your questions will be considered for inclusion in future revisions of guidance.

Background

The New York State Board of Regents has committed to the transformation of the preparation, support, and evaluation of all teachers and school leaders in New York State. In 2010, the Legislature enacted Education Law §3012-c to establish a comprehensive evaluation system for classroom teachers and building principals across the State. Chapter 21 of the Laws of 2012 amended Education Law §3012-c to fundamentally change the way teachers and principals are evaluated. The purpose of the evaluation system is to ensure that there is an effective teacher in every classroom and an effective leader in every school. The evaluation system also fosters a culture of continuous professional growth for educators to develop and improve their instructional practices.

Under Education Law §3012-c, New York State will differentiate teacher and principal effectiveness using four rating categories – Highly Effective, Effective, Developing, and Ineffective. The Law requires APPRs to result in a single composite teacher or principal effectiveness score that incorporates multiple measures of effectiveness. The results of the evaluations shall be a significant factor in employment decisions, including but not limited to promotion, retention, tenure determination, termination, and supplemental compensation, as well as teacher and principal professional development (including, but not limited to, coaching, induction support, and differentiated professional development).

The law specifies that student achievement will comprise 40% of teacher and principal evaluations, as follows:

- For the 2011-2012 school year and thereafter, for teachers and principals in subjects and grades where there is no "value-added" model approved by the Board of Regents for such subject and grade: 20% on student growth on State assessments or comparable measures, and 20% on other locally-selected measures that are rigorous and comparable across classrooms in accordance with standards prescribed by the Commissioner.

- For the 2014-2015 school year and thereafter, for teachers and principals in subjects and grades where there is an approved "value-added" model by the Board of Regents for such subject and grade: 25% on student growth on State assessments or comparable measures, and 15% on other locally-selected measures that are rigorous and comparable across classrooms, in accordance with standards prescribed by the Commissioner.

The remaining 60% of teacher and principal evaluations shall be based on multiple measures of teacher/principal effectiveness consistent with standards prescribed by the Commissioner in regulation. This will include the extent to which the educator demonstrates proficiency in meeting New York State's teaching or leadership standards.

On February 11, 2014[1], the Board of Regents adopted as an emergency measure a series of changes to Subpart 30-2 of the Rules of the Board of Regents, effective on that same date, to support the commitment made by the Board of Regents and the Commissioner to ensure that students are not burdened by more testing than necessary or testing that takes away from the core instructional time in our classrooms and schools. Further, these amendments help to ensure that our youngest students in grades kindergarten through second grade are not subject to *traditional standardized testing* [2].

Specifically, the regulatory amendments provide that no APPR plan shall be approved by the Commissioner for use in the 2014-15 school year or thereafter that provides for the administration of *traditional standardized assessments* to students in grades K-2 that are not being used for diagnostic purposes or are required to be administered by federal law, including but not limited to assessments developed by any vendor, third party or other comparable entity; except that nothing in this subdivision shall preclude the use of school- or-BOCES-wide, group or team results using State assessments that are administered to students in higher grades in the school or a district, regional or BOCES developed student assessment that is developed in collaboration with a vendor, if otherwise allowed under this section or guidelines of the Commissioner. However, this prohibition shall not apply to any APPR plan approved or determined by the Commissioner for use in the 2013-2014 school year which remains in effect in the 2014-15 or thereafter in accordance with Education Law §3012-c(2)(l).

The regulatory amendments also require that for any APPR plan submitted to the Commissioner for approval for use in the 2014-15 school year, the plan must include a signed certification by the superintendent, district superintendent, or chancellor that attests that **no more than one percent of total instructional time in each classroom or program of the district or BOCES is spent taking any locally determined *traditional standardized assessments* (developed by a vendor, third party, or other comparable entity or district, regional, or BOCES- developed) for APPR purposes**[3]. The amendments pertain only to the administration of *traditional standardized assessments* to students for APPR purposes– **the amendments do not pertain to those assessments that are used for formative or diagnostic purposes or are required to be administered by federal law.**

At its March 2014 meeting, the Board of Regents made a series of additional revisions to the regulations to clarify what types of third-party assessments may be used for APPR purposes in grades K-2. Based on these amendments, the Department no longer requires a district or BOCES to use the RFQ process when selecting any third party assessment for use in kindergarten, first, or second grade for

[1] Technical amendments to Subpart 30-2 of the Rules of the Board of Regents were adopted at the March 11, 2014 meeting, see: "Proposed Technical Amendments to Subpart 30-2 of the Rules of the Board of Regents to Clarify the Requirements for Districts and BOCES that Opt to Use an Assessment That is Not a Traditional Standardized Assessment for Grades K-2 for Purposes of APPR)"

[2] The Board of Regents items are posted as follows: "Adjustment Options to Common Core Implementation - Full Report of the Work Group"; "Proposed Amendments to Subpart 30-2 of the Rules of the Board of Regents Relating to a Prohibition Against Traditional Standardized Testing for Students in Grades K-2, Removal of K-2 Tests From the List of Approved Student Assessments, Change in the Definition of Core Subjects for the State Growth or Other Comparable Measures Subcomponent and a Limitation on Instructional Time Spent on Taking Local Assessments for Purposes of Annual Professional Performance Reviews (APPR)"; "Proposed Technical Amendments to Subpart 30-2 of the Rules of the Board of Regents to Clarify the Requirements for Districts and BOCES that Opt to Use an Assessment That is Not a Traditional Standardized Assessment for Grades K-2 for Purposes of APPR)"

[3] For further information on the Board of Regents emergency adoption of the amendments to Subpart 30-2 of the Rules of the Board of Regents and the impact of its changes to APPR plans, see: http://www.engageny.org/resource/guidance-on-the-approved-regulatory-amendments-to-appr-to-help-reduce-local-testing

APPR purposes. Rather, pursuant to this regulatory amendment, a district/BOCES may use an assessment for these grades if it is not a traditional standardized assessment and it meets the minimum requirements set forth in this guidance document (see the bolded sections of the Section 2). The Department is issuing this APPR Assessment Guidance, to provide districts/BOCES with the minimum criteria a third party assessment must meet in grades K-2 to ensure the third party assessment is consistent with Education Law §3012-c and Subpart 30-2 of the Rules of the Board of Regents, including the prohibition against the administration of traditional standardized assessments to these students for APPR purposes. The regulatory amendments also require that for any APPR plan submitted to the Commissioner for approval for use in the 2014-15 school year, the plan must include a certification by the superintendent, district superintendent or chancellor that attests that any K-2 third party assessment they are using for APPR purposes is consistent with this APPR Assessment Guidance and is not a *traditional standardized assessment*.

Traditional standardized assessments are defined by NYSED for the purposes of this regulation as a systematic method of gathering information from objectively scored items that allow the test taker to select one or more of the given options or choices as their response. Examples include multiple-choice, true-false, and matching items. NYSED defines this term to focus specifically on those assessments that require the student (and not the examiner/assessor) to directly use a "bubble" answer sheet.

With the rapidly changing assessment landscape as a result of technology (e.g., game-like, adaptive, new tools for performance-based assessment like simulations and on-screen drawing of models), districts and BOCES have an increased selection of assessments that are not traditional standardized assessments; however, districts and BOCES shall evaluate these assessments in the context of this guidance and determine their appropriateness.

The approved regulatory amendments to the APPR are intended to help districts and BOCES reduce local testing and ensure that the amount of testing should be the minimum necessary to inform effective decision-making at the classroom, school, and district/BOCES level. For the purposes of APPR, there have never been K-2 standardized tests administered or required by the State[4]. Decisions about how to measure student progress in these grades are made by local school districts and BOCES. Education Law §3012-c provides districts and BOCES with design flexibility and several assessment options. The Department urges districts and BOCES and their respective collective bargaining units to identify other ways to assess learning progress for these very young students. There are a variety of ways in which a district or BOCES can design a meaningful and authentic assessment program that provides information to drive instructional decisions, ultimately leading to an increase in students' knowledge and skills.

[4] Federal law mandates that the English proficiency of all English language learners enrolled in Grades K-12 be assessed annually. The New York State English as a Second Language Achievement Test (NYSESLAT) is the annual assessment that NYS administers to comply with federal law. The NYSESLAT gives the State and schools important information regarding English language development of English language learners. In the listening, reading, and writing subtests, students in grades K-2 mark their answers in their test booklets by circling or otherwise marking the answer or picture. There is no bubble sheet for these students to use. A teacher or aide must transcribe the student's responses to the answer sheet exactly as the student recorded it in the test booklet. The speaking portion of the NYSESLAT is administered to a student one-on-one. For a parent’s guide, please see: NYSESLAT.

This guidance document next provides information to help districts and BOCES navigate key assessment concepts. Furthermore, this guidance provides information about the design principles that lead to assessments that are reflective of student learning experience.

Districts and BOCES can use this guidance to determine whether the third party assessments they are using for the purposes of APPR are consistent with the changes to Subpart 30-2 of the Rules of the Board of Regents and the Regents action to eliminate *traditional standardized assessments* for use to assess students in K-2.

Section 2: Summary of Required Assessment Information

The Department provides the chart below for districts and BOCES to use when reviewing K-2 third-party assessments for potential inclusion in a district or BOCES' APPR plan. Any third party assessment selected shall not be a traditional standardized assessment and, importantly, to be compliant with the Department's APPR guidelines, assessments must meet the requirements stipulated in the Commissioner's Regulations §30-2.8, listed below in bold. Other assessment characteristics listed below are recommended but not required.

The information in this summary document has been adapted from the Department's Round 4 RFQ Application for Approved Third-Party Assessments.

ALIGNMENT TO STANDARDS: As stated in §30-2.8 of the Regents Rules, the assessment must be aligned with the New York State learning standards or, in instances where there are no such standards that apply to a subject/grade level, evidence of alignment to research-based learning standards.	
MEASURING GROWTH: **As stated in §30-2.8, the provider must have a detailed procedure for measuring growth using the student assessment, that such assessment will result in normative inferences about each individual's student growth.** **The provider must be able to provide information on the one or more norming groups used to calculate normative growth as well as the required test administration procedure, including a recommended testing timeline when using the instrument to measure growth, including the potential use of a pre-test or other tool in the first year of implementation.** If measuring growth, there should be a recommended administration procedure including the points in time at which the assessment should be administered to make valid inferences about student growth. It is recommended that assessments include specific procedures and include whether the same form is administered at two or more points in time, or if multiple equated forms are available to make inferences about growth. Furthermore, if assessment is able to yield inferences about growth using one or more other assessments as the pre-test (i.e., assessments that	

are part of a different product line and/or that are not psychometrically equated), it is recommended that an explanation and references to empirical studies that best support these claims are provided.	
INFERENCES FROM SCORES: Assessments should include an overview of the content and skills purportedly measured by the assessment as well as documentation of scoring processes and inferences made from scores. This may include how the assessment classifies students into performance categories, evidence supporting validity of standard setting process and resulting cut-scores, and/or evidence pertaining to ability to report student results of assessments as growth percentile rank (as well as information on norming groups).	
As stated in §30-2.8, there must be strong evidence that the assessment is aligned with industry standards of reliability and validity as defined in the Testing Standards. **VALIDITY EVIDENCE:** Assessments should include a variety of validity evidence to support inferences about student performance, including but not limited to: evidence of content, construct, concurrent, or predictive validity as appropriate. If available, assessments should also include evidence of validity of using assessment results to support inferences about effectiveness of teacher in producing growth or achievement in student performance.	
RELIABILITY: Assessments should include estimates of reliability / error of measurement (e.g., inter-reliability, estimates of error expressed in confidence intervals for reported scores) and strong evidence that the assessment is aligned with industry standards of reliability and validity as defined in Testing Standards.	
MEASUREMENT ACROSS THE KNOWLEDGE/ SKILL DISTRIBUTION: Assessments should include evidence that the assessment has items of varied difficulty that cover the expected knowledge/ skill distribution for the examinees of interest (i.e., sufficient item coverage at the tails of the distribution).	

SUBGROUPS: Assessments should include documentation supporting use of assessment for disaggregated student sub-groups (e.g., race, sex, poverty level, ethnic groups). Evidence that the assessment does not exhibit bias toward any major subgroups (e.g., through an analysis of differential item functioning).	
TECHNICAL MANUAL AND ASSESSMENT ADMINISTRATION DOCUMENTATION: Assessments should include the most recent technical and administration manuals for the assessment. Additionally, the assessment administration protocol, including when during the year the applicant suggests the assessment should be administered to make inferences about (a) achievement and (b) growth (if not included in manual(s)), should be provided. Assessments should include documentation related to assessment administration that delineates recommended guidelines for assessment security for the assessment being proposed given that the assessment will be used for educator evaluation and should delineate threats to validity for the guidelines provided.	
CROSSWALK: Assessments should provide a suggested crosswalk of the native assessment scores onto two scales used in NYSED's APPR system: 0-15 and 0-20.	

Section 3. Selecting a Third-Party Assessment for APPR Purposes

Identifying Potential Existing Assessments

After thoroughly reviewing APPR requirements under Education Law §3012-c (see Appendices A and B), districts may begin the process of identifying assessments that are not traditional standardized third party assessments that have the potential to be used for the purposes of principal and teacher evaluation at the K-2 level. It is the responsibility of the district to choose any assessments that are used for this purpose, rather than the responsibility of NYSED.

When considering potential existing assessments for K-2, NYSED requires that districts consider only those authentic assessments that reflect the type of student learning that is experienced in the classroom. Appropriate assessments for K-2 students should include questions that require the student to demonstrate understanding of concepts presented to them in their learning environments. For instance, tasks that are required for assessment should resemble the everyday learning experiences of students.

Furthermore, districts shall use assessments that are developmentally appropriate for students at the K-2 level and closely resemble that which occurs in the classroom, not those that are commonly referred to as "traditional standardized tests". While existing, vendor-produced assessments may be available for students within that grade band, it is the responsibility of the district to determine whether those assessments meet the criteria outlined in this guidance document. For example, an assessment that is said to assess students in grades K-2 may not necessarily include tasks that are optimal or even appropriate for that grade level. Therefore, it is the duty of the district to determine the suitability of the assessment and its tasks. With the rapidly changing assessment landscape as a result of technology (e.g., game-like, adaptive, new tools for performance-based assessment like simulations and on-screen drawing of models), districts and BOCES have an increased selection of assessments that are not traditional standardized assessments; however, districts and BOCES shall evaluate these assessments in the context of this guidance and determine their appropriateness.

As an example of a suitable assessment for K-2, consider a literacy assessment that uses a multi-modal approach when assessing students. Rather than simply asking students to select their answer from a pre-determined number of choices that are presented to them, students must sound out letters or words, identify sounds, and demonstrate their understanding of words to an administrator that records student responses. This example demonstrates the use of assessment items that require a student to perform a task in order to demonstrate understanding of a concept.

As another example of a developmentally appropriate assessment for K-2, consider an instrument that incorporates technology and game playing into its assessment of student learning. When assessing vocabulary and reading comprehension, students must use a computer or tablet to engage with questions that include animation similar to a cartoon or video game. Depending on the student's response, the instrument adjusts to the student's level of knowledge (i.e., adaptive testing). This assessment strategy is seamlessly integrated into that which the student is familiar. While assessing the degree to which a student understands a concept that is taught in the classroom, the student's learning is not interrupted by the assessment process. Instead, the student is fully engaged in the process, and the instrument adapts to the student's responses.

These examples demonstrate age-appropriate methods of assessing K-2 student competencies. Other considerations when choosing an assessment for APPR purposes include administration time and the potential to measure growth from two points in time as required for Student Learning Objectives (SLOs) used for the State Growth or Other Comparable Measures subcomponent (see Section D of the APPR Guidance document: http://www.engageny.org/sites/default/files/resource/attachments/appr-field-guidance.pdf)

Evaluating Potential Existing Assessments

Districts should evaluate each potential assessment to determine quality and verify the appropriateness of the assessment for APPR purposes. The basic process involves gathering information about the assessment's purpose and quality and conducting an internal review of the assessment. During the internal review, districts should consider numerous criteria using the assessment concepts that were described in Section 2. Descriptions of the criteria to be evaluated are outlined below to guide districts through the remainder of the process. Districts may also choose to consider the Summary of Assessment Information adapted from the Department's Round 4 of the Assessment RFQ, which is provided in Section 2.

Step 1: Gather Documentation Regarding the Quality of the Assessment (Reviews, Critiques). For each assessment considered for APPR, districts should gather documentation regarding the quality of the instrument. When possible, information should be gathered from sources external to the publisher or developer as well as from the authors/test developers. For assessments historically used in the district, this process may include collecting feedback from teachers who have used the assessment. For assessments developed by third parties, vendors, or other comparable entities, this process may include reviewing documentation from the following sources, where available:

- Published information from sources external to the test developers such as:
  - Formal reviews published by sources external to the author or publisher
  - Informal reviews of the assessment published by research or evaluation groups
- Published or unpublished information from the author or publisher of the assessment such as:
  - Technical manual
  - Administration manual
  - Policies regarding the assessment
  - Any other available information (e.g., recent reliability data, newly created norms)
- Unpublished information from sources external to the assessment authors (e.g., teachers in the district or other districts who are using or have used the assessment) such as:

- Testimonials, ratings, or other user feedback
- Data that demonstrates the quality of the assessment
- Any other available information that can be shared

Vet the material gathered about the assessment's quality with an eye toward how knowledgeable and current the source is. Knowledgeable sources will include information about the assessment concepts outlined in Section 2, including construction, reliability, validity, bias checks, and administration and reporting procedures. Knowledgeable sources will further address the quality of the assessment as it aligns to the purpose or use of the assessment.

Step 2: Use the Documentation to Evaluate the Assessment.
Review the documentation gathered to determine whether the assessment is of sufficient quality for use for APPR. Because the information may exist in one or more locations in the gathered documentation, the criteria below are organized by topic. The tasks detailed below will assist the district in conducting a systematic evaluation.

The following list provides the basic components of and information about an assessment that districts should collect and review. If components or information are missing but the assessment is a promising candidate for use as a local assessment, districts can begin to develop missing components and/or gather evidence of the assessment's quality (e.g., pilot data). Note that the evaluation process can be halted at any time if the review team determines that the assessment is not appropriate for APPR purposes.

The following actions will help districts in determining how to choose an existing assessment for APPR purposes:

1. Identify general information about the assessment.

   a. *Subject/grade*: Identify the subject/grade the assessment will be used for.

   b. *Assessment name*

   c. *Assessment purpose*

   d. *Duration and time of administration:* Indicate how long the assessment takes to administer as well as when the assessment takes place during the year.

   e. *Additional purposes:* If relevant, indicate if there are other uses of the assessment in addition to the primary purpose.

   f. *How the assessment was chosen:* Specify the factors that were taken into consideration when choosing the assessment.

2. Consider the strength of the assessment based on the following factors:

   a. *Rigor:* The assessment should be reviewed based on its rigor in three categories:
      1. The degree to which the assessment measures the intended learning standards
      2. Validity of the assessment
      3. Reliability of the assessment

Please note: 2 and 3 should be based on the *Standards for Educational and Psychological Testing* to the extent practicable.

b. *Comparability:* The assessment should be reviewed based on its comparability, defined as the extent to which the results of the assessment support comparable inferences about student performance and progress when used by different teachers for the same grade and subject.

c. *Informs instruction:* The assessment should be reviewed for the extent to which it supports classroom instruction by providing clear, informative feedback to teachers and learners about both the status and needs of learners.

d. *Supports learning goals:* The assessment should be evaluated on how well it supports the process of student learning and/or provides feedback to students on learning progressions on a continuum from kindergarten to grade 12. The assessment should address the needs of diverse learners including students with disabilities, English Language Learners, accelerated learners, and students achieving and performing below grade level.

e. *Utilizes a diverse set of assessment techniques (i.e. performance-based tasks):* A judgment should be made about diversity of assessment tasks, including whether the assessment promotes the application of knowledge and understanding.

Finalizing a Decision for Selecting an Assessment

After conducting an evaluation of the assessment and recording results, district teams will have enough information about an assessment to begin to make a determination about whether it can be used for APPR. Often, review teams will find that they have some but not all of the information needed to make a decision. If more information is needed, the district can do one of two things:

- Continue to collect information while piloting the assessment
- Collect more information about the potential assessment before making a final decision

Monitoring the Assessment’s Use

The process of maintaining a district’s set of assessment will be ongoing. After a district selects and implements an assessment for APPR, the district should monitor the quality of the assessment to determine if it is living up to its promise of being a high-quality assessment. Districts may wish to continue to monitor the following assessment characteristics:

- Continued alignment to the district’s curriculum and intended degree of rigor
- Instrument security (i.e., procedures intended to ensure that assessment results are not tainted by improper instrument administration procedures or by an overfamiliarity with the exam or the exam contents)
- Reliability

- Validity associated with its use, including useful results and good score reporting
- Feasible administration and scoring procedures

Administering the Assessment

Establish Local Policies

Before administering the assessment, it may be necessary to establish district- or school-level policies to ensure best use and interpretation of assessment scores through the establishment of good score reports, standardization in administration, security, training of administrators, and other policies required to ensure that the assessment and its administration manual are kept up to date. Please also see Section G of the APPR Guidance Document (Scoring and Security of Assessments): http://www.engageny.org/sites/default/files/resource/attachments/appr-field-guidance.pdf

Communicate Results

After administering and scoring the assessments, information about student performance may be communicated to students and their parents as well as to teachers and other educators. These communications can take many forms; what is important is that they are appropriate for the audience.

Monitor Assessment's Use

After your district begins using an assessment operationally, it is important to periodically reevaluate the instrument and the items to ensure they are still performing adequately. Changes in things like the passing rate over time may indicate that instrument items have been compromised through overexposure. Psychometric quality should also be evaluated from time to time, as should alignment of the assessment with the curriculum.

Section 4. Key Assessment Concepts

In this section, key educational assessment concepts are introduced and discussed with respect to identifying high-quality assessments. Unless otherwise stated, this information is relevant for assessment of all grade levels. After reading this section, district and BOCES should have a better understanding of key assessment concepts such as assessment domains, reliability, validity, and bias/fairness in testing.

Assessment Terminology

"Assessments" Versus "Instruments"

Assessment is a general term that, as used in this guide, describes the systematic process of collecting, reviewing, and using information about students. The assessment process includes many steps:

- Identification of the purpose of the assessment and the target audience to be assessed
- Identification of the knowledge, skills, and/or behaviors to be measured
- Identification or development of a data collection tool (e.g., an instrument such as a test) that can be used to gather information about students
- Development of a method to score the instrument
- Administration and scoring of the instrument
- Development and application of policies associated with the reporting, interpretation, and use of the resulting scores
- Development of documentation to detail the assessment process

For the purposes of this guide, the term *assessment* will be used to refer to a complete process of measuring student knowledge and skill primarily in the cognitive (academic) domain; however, it is important to note that some assessments also measure the affective/behavioral, and psychomotor domains. The cognitive domain is concerned with the acquisition of academic knowledge, content, and skills (such as the acquisition of literacy, math, or science knowledge and skill), and is often measured through knowledge and skill tests or through an examination of student work products that address the New York State P-12 Learning Standards.

The term *instrument* will be used to refer specifically to the tool that collects information from the student, regardless of whether the tool being used is a survey, performance task, observation, paper, project, or other type of data collection tool.

Components of Instruments

As described above, the assessment instrument is used to gather information about student learning. All instruments used for the purposes of APPR will include some common components, particularly student instructions, assessment items or some other method of communicating to the student the information on which the student is being assessed, and a method or methods for students to respond:

- *Instructions* provide sufficient directions to enable students to take the instrument (e.g., test). Instructions should be complete, concise and placed in a location that helps ensure students will read them. In the K-2 context, instructions should be developmentally appropriate for the student being assessed, and in many cases are more appropriately read aloud to the student.

- *Items*[5] are statements, questions, or other prompts that elicit the knowledge or skills being assessed. Items are sometimes comprised of an *item stem,* which contains the statement, question, or other prompt to which students must respond, along with other stimulus materials, such as graphs, drawings, or narrative text. In K-2 assessment, items may be seamlessly integrated into the daily classroom life and blur the distinction between assessment and student learning.

- *Response methods* clearly indicate how examinees provide answers to the instrument. Response methods typically include directing examinees to (a) mark answers on the instrument itself (including through the use of technology), (b) provide short or extended responses, or (c) complete a performance task in a public, small-group, or one-on-one situation that may be assessed using a rubric.

Item Types

While there are many different item types, some formats may be considered more appropriate for K-2 students than students in higher grades. An important distinction to consider is whether students select or construct (create) a response. For *selected response* items, students choose the correct response to an item from among two or more options. Selected response items include multiple-choice, true-false, rate/rank, matching, and paired comparison items, among others. These items are typically attractive because they can be scored easily and objectively by using an answer key, which identifies the correct answer for each item. Importantly, traditional selected response items may not be ideal for K-2 students because this item type often allows little room for the student to demonstrate how he or she arrived at an answer and the format may not be as developmentally appropriate as more engaging formats such as performance tasks.

Selected response items may be more suitable for assessing young students if technology is integrated into the assessment process. For example, a student may be presented with a question through a technology platform that interacts with the student, such as a narrator that resembles a cartoon character. A set list of options may then be presented on the screen, and the student is required to physically select their answer by clicking the option with a mouse or touching the screen. This process, which still falls under the category of selected response, may be more engaging for a student and may therefore resemble the learning environment more closely than a paper and pencil test.

[5] *Item* and *question* are used interchangeably in this document.

For *constructed (or open) response* items, students create their own response rather than selecting one from a set of potential responses. Constructed response items are further differentiated by the length and parameters of the student's response. For example, constructed response items include the following:

- Short answer (requests a very short response, usually one word, a short phrase, or a number)

- Restricted constructed response (directs examinees to provide a fairly short response, usually a brief, targeted answer or explanation, or a directed math symbol, equation or scientific drawing is requested)

- Extended constructed response (requires examinees to provide a long, organized response, such as an essay, narrative, criticism, mathematical or scientific reasoning or proof, or other complex response). While appropriate for older students, this response type is typically not appropriate for younger students.

- Performance item (directs examinees to demonstrate capability through a live performance or through the creation of a product, such as a completed lab experiment, research report, art product, or performance piece). For examples of performance tasks for K-2 students, see: http://readingandwritingproject.com/resources/assessments/performance-assessments.html

- Observational/interview item (provides directions for proctors to observe and catalog student behavior through a rating scale, checklist, or anecdotal records form). The observational method can be unobtrusive, in which the examinees are unaware that their behavior (such as their use of speaking or listening) is being examined or when the observer is sufficiently removed from the performance so as not to intrude, or it can require more participation from examinees, in which the proctor requests that examinees demonstrate the assessed behaviors.

Answers to constructed response questions (such as short answer, restricted constructed response, or extended constructed response) may be provided through multiple modalities. For instance, while a written response may be appropriate for some students, an alternative approach may include having a student verbally answer the question. This may be more appropriate for K-2 students in order to keep the students engaged with the assessment administrator, and it may also be more suitable in the context of what is being assessed. For instance, a kindergarten student being assessed on their literacy skills may be asked to pronounce the sound that is associated with a letter or string of letters. In this case, the best way to assess the student's knowledge is to have him or her produce the sound verbally while the administrator records the answer.

Each item type has strengths and potential challenges and their use should be considered based on the testing context, including the population of students to be assessed. For example, selected response items typically take less time to administer, allowing examinees to take more of them during a shorter period of time (increasing content representation and reliability on an assessment). They also typically have more reliable scoring. However, constructed response items can assess student performance of complex behaviors (such as the capacity to research a topic or construct an organized response). Furthermore, constructed response items have the advantage of typically being more engaging than selected response type items. This quality may make constructed response items more effective for assessing younger examinees, as this group of students may generally have a shorter attention span. For the assessment of K-2 students, districts are encouraged to carefully consider whether item types closely resemble classroom activities and can be interwoven into the learning environment of students. A comparison of the advantages of three popular item types (selected response, constructed response, and performance/observational items) is presented in Table 1. Note that the table assumes the item types are well conceived, developed, and implemented.

Table 1. Comparative Advantages of Item Types

	Selected Response (Objectively scored items)	Constructed Response (Subjectively scored items)	Performance / Observational Items
Sampling of Curriculum	Samples a lot of academic standards in a short period of time	Samples less curriculum than selected response items; takes longer examinee administration time	Samples less curriculum than selected response items; takes longer examinee administration time
Item Development	Requires the development of many items	Fewer items are needed	Fewer items are needed, but the items are written to break out the components of the task
Complexity and Rigor	Can sample a range of cognitive complexity (e.g., Bloom's Revised Taxonomy); takes skill to write items at the higher levels of rigor	Should be written for higher levels of rigor and rubric can give credit for partially correct items	Can range the levels of rigor, although some items should represent higher-level demands
Scoring	Objective scoring—efficient with a scoring key	Subjective scoring—requires the use of rubrics/scoring papers and scorer training	Subjective scoring— requires the use of rubrics/scoring papers and scorer training
Influence on Learning	Overuse of the selected response item format can encourage learner passivity; can encourage development of critical thinking skills when items align with higher levels of rigor	Good-quality constructed response items can encourage examinees to demonstrate creativity, organizational skills, topic development, critical thinking skills	Encourages the examinees to demonstrate what they know and can do. Depending on the item content, can encourage the development of critical thinking, organizational skills, and creativity
Reliability	High internal consistency reliability is possible with the inclusion of 20+ high-quality items	Reliability is typically lower than with selected response items due to scorer differences, fewer number of items	Reliability is typically lower than with selected response items due to scorer differences, fewer number of items

Adapted from: Linn & Gronlund (1995).

Other item qualities must also be examined when considering the value of an item. For example, a high-quality assessment should be comprised of items that are not too difficult or too easy for the target audience and should be populated with items that discriminate appropriately between strong and weak performers.

Defining Characteristics of Assessments

Districts are encouraged to employ assessment strategies that enhance instructional practice and promote high-quality learning in classrooms. High-quality assessment programs benefit students and educators alike by providing educators with robust student data they can use to improve their instructional practices, leading to better student outcomes.

The task of selecting assessments for APPR purposes provides districts the opportunity to scrutinize their assessment practices in order to answer important questions:

- How well is the district measuring student learning?
- What types of assessments are used in the district and why?
- Are there other types of assessments the district could/should be using for certain educator roles?
- Does the district have systems in place to ensure that all students have a fair and equal opportunity to demonstrate what they know?
- Does the assessment measure specific Common Core Learning Standards (CCLS)?
- Does the assessment blend into the learning environment of students, minimally interrupting time that is devoted to instruction?

This section of the guidance describes examples of defining characteristics of assessments that districts should consider as they select assessments for APPR purposes.

Direct and Indirect Measures of Student Learning

Direct measures of student learning assess student learning, growth, or achievement with respect to specific content represented by key standards and learning objectives identified by the district/BOCES. Direct measures are strongly preferred for evaluation because they measure the most immediately relevant outcomes from the education process. Examples include assessments of student achievement in a subject, culminating student projects, essays or performance assessments.

Indirect measures of student learning, growth, or achievement provide information about students from means other than student work. These measures may include student record information (e.g., grades or other data related to student growth or achievement such as high school graduation or college enrollment rates). These measures are not appropriate for APPR purposes, but may have other value within a district or BOCES.

Assessment Types

Assessments administered at different times capture learning at different periods and intervals. End-of-year (EOY) and end-of-course (EOC) assessments may initially seem to be the most logical choice for APPR, but they are not the only options.

Traditional standardized assessments are defined by NYSED for the purposes of this regulation as a systematic method of gathering information from objectively scored items that allow the test taker to select one or more of the given options or choices as their response. Examples include multiple-choice, true-false, and matching items. NYSED defines this term to focus specifically on those assessments that require the student (and not the examiner/assessor) to directly use a "bubble" answer sheet.

Performance assessment requires examinees to perform a task, often an authentic or “real” task. The purpose of performance assessment is to allow a student to display understanding of a concept through a performance. Well-constructed performance assessments are often engaging and meaningful for students, making this type of assessment particularly beneficial to students in earlier grades. Performances may include demonstrations, explanations, conducting work, problem solving, etc. Examinees are then scored on their performances, which may include products that may be components of the performance.

Groups of educators often create high-quality performance assessments over time by trying the performance tasks out first with students for instructional purposes, then using an iterative process to gradually hone the tasks to improve their instructional and measurement qualities. After the tasks are administered, teachers or test developers review the responses using a rubric to determine how well the task items and prompts elicited the targeted student behaviors. The review team then adjusts the tasks, items, prompts, and rubrics in an effort to improve the students’ demonstration of the learning objectives.

Portfolio assessments involve the use of purposeful and systematic collection of student work over time. For APPR purposes, the culminating work(s) of the student portfolio could be used to measure student growth in the course. Portfolios are assembled in accordance with a protocol and scored using a well-defined rubric or scoring papers. When appropriately designed and implemented, portfolios provide an opportunity to conduct an “authentic assessment” (one that is intimately embedded in instruction and limits time spent away from instruction) and allow for the examination of each student’s work product. To that end, they can demonstrate complex thinking, organizational and problem-solving abilities. For the purposes of APPR, educators and/or peers are not permitted to actively participate in the creation or revisions process of products that are included in a student’s portfolio when the summative product is used for the evaluation of educators in the State Growth or the Other Comparable Measures subcomponent or the Locally selected Measures subcomponent; however, student work for the portfolio that is revised over time, or that is used throughout the year for formative or other instructional purposes may be used as an artifact for the Other Measures subcomponent (e.g., portfolio of student work). Portfolio assessments **must** reflect independent student work only and may not reflect educator or peer supported efforts. As is the case with constructed response items, considerable time will be spent developing scoring rubric guidelines and materials, providing scorers with training, and establishing that the scoring of the student work is done reliably.

Hybrid assessments combine an assessment with a portfolio or performance assessment to achieve a more balanced type of assessment, one that provides a broad representation of content (in the on-demand assessment) and includes complex tasks (presented in the portfolio or performance assessment).

Interim assessments are those that are aligned with key educational standards and learning objectives (identified through the district’s curriculum scope and sequence or curriculum map) and administered in a single grade level across all applicable schools, typically in a single subject. This assessment type is better suited for formative or diagnostic purposes and is not admissible for APPR purposes (see F11: http://www.engageny.org/sites/default/files/resource/attachments/appr-field-guidance.pdf); however, in some instances an interim assessment may be used summatively and thus may be used for APPR purposes (districts and BOCES are encouraged to check with their vendors for more details on the prescribed uses of their assessments).

Commercial Assessments

Commercial assessments are assessments developed by a vendor or third party that are purchased by districts and BOCES. There are two basic types of commercial assessments used for APPR purposes: criterion-referenced and norm-referenced.

Criterion-referenced assessments measure how well a student has learned a specific body of knowledge and skills (i.e., the "criterion" or "domain" of interest). Most tests and quizzes written by teachers are criterion-referenced assessments. The commercial interim assessment programs used in some schools and districts are examples of district-based criterion-referenced tests. These tests are designed to represent the district curriculum. Criterion-referenced assessments have the advantage of connecting examinee performance to a set of standards, grounding the score reporting in a description of how well a student has met the criterion (learning standards).

Norm-referenced assessments, on the other hand, provide an estimate of how an individual student performed on the assessment compared to a predefined group. The "norm group" is intended to be a representative group of students based on the country's (state's, district's, etc.) demographic characteristics at the time the test is developed. The commercial test developer first administers the assessment to the norm group and generates scaled scores and percentiles based on that group's performance. Students taking the published test then receive scores that are compared to the norm group. Percentiles are one type of comparative score that report the percentage of students (in the norm group) that the examinee performed better than (For example, a percentile of 75 indicates that a student performed higher than 75% of other students, but does not provide information about what the student knows and can do with respect to learning standards).

Alignment of Content to the Curriculum

An important characteristic of high-quality assessments is that items on the instrument are representative of the intended curriculum. The intended curriculum is usually described in the district's curriculum scope and sequence or curriculum map. A curriculum map typically will present the following information:

- The time frame for each section (unit) of curriculum content
- The associated standards
- The curriculum unit connections
- Assessments associated with the content

Assessments should include a Table of Test Specifications that demonstrates how the assessment relates to the content identified in the curriculum. When evaluating an assessment for APPR purposes, district and BOCES teams should review the Table of Test Specifications to ensure the assessment is well aligned to the local curriculum scope and sequence or curriculum map. In some cases, the test items can be linked directly to the academic standards. In other cases, the standards will need to be broken out into observable (e.g., measurable) learning skills or objectives.

Introduction to Reliability and Validity

Issues of the reliability and validity of educational assessments are integral to the use of instruments for all educational purposes. Reliability and validity issues affect everyday use of assessment results in classrooms, but because the effects are usually confined to the classroom and because the decisions are so numerous and rapid, the effects are not often noticed. Issues related to reliability and validity become more noticeable when test scores are used to make decisions about students and teachers, as is the case with APPR.

This document next explores issues of reliability and validity, as these are concepts that are often described in technical manuals of educational assessments. These terms are discussed in the context of using educational assessment for principal and teacher evaluation. For more information on the terms "rigorous" and "comparable" please see Section F of the APPR Guidance: http://www.engageny.org/sites/default/files/resource/attachments/appr-field-guidance.pdf.

Reliability

Reliability, in its broadest sense, refers to the consistency or stability of scores from an assessment. It is an indication of the confidence one can have that differences in test scores reflect actual differences in the characteristic being measured (e.g., what a student knows and can do in math), as opposed to "error" (sometimes referred to as "noise"). Understanding the reliability of an assessment is essential to understanding and interpreting the results of that assessment. For example, consider a math test comprised of a single problem. Students' responses are probably not a good reflection of what they actually know or do not know. The test did not provide enough coverage of the content to provide reliable information. This section describes some common factors that influence reliability, and therefore interpretability of scores but are not related to what students actually know or do not know in a subject. It is imperative that assessments show sufficient reliability so that they can be used to inform an educator's student impact rating.

Many factors may introduce error and hence reduce reliability in test scores. For example, students may be "lucky" and be presented with a testing scenario they experienced in class or "unlucky" and presented with two vocabulary words from texts they had not yet read. In fact, no test is perfectly reliable. Generally, longer tests with more items—assuming the items are of similar quality and are focused on the same content domain—tend to be more reliable than shorter tests. Standardization of test administration procedures also increases reliability.

Reliability is usually measured on a 0 to 1 scale, with "1" representing "perfect reliability" and no measurement error (that is, the different between a student's test score and that student's actual knowledge of what is being tested) and "0" representing "no reliability" and all measurement error. However, in practice, no test has a reliability of 1 or 0. High reliability, then, is *one* characteristic of a high-quality instrument. Measures of reliability are provided in the documentation that accompanies high-quality commercial assessments.

Because reliability is very hard to "see" or detect in classroom testing, there are four basic methods for estimating it. The factors affecting stability and consistency play out in different ways in different test situations:

- *Internal Consistency Reliability* refers to consistency among the items in an assessment. If the items represent the same tested content (e.g., content representing 1st grade curricula in math or reading), they are internally consistent. Consider internal consistency when you have a traditional assessment with many items (such as an end-of-the-year test). Because internal consistency is concerned with the degree to which items represent the same content, it is important to consider the internal consistency reliability for each test section separately when tests represent two or more content areas. Internal consistency reliability can be estimated using just one administration of a single form of the assessment.

- *Test-Retest Reliability* refers to consistency (i.e., stability) in test scores over time. To estimate this type of reliability, the same exam needs to be administered to the same group of students after a period of time (say, 2–3 weeks) without teaching students the tested content between administrations. This is a form of estimating reliability for assessments designed to measure student growth using a pretest-posttest design.

- *Alternate (Parallel) Form Reliability* refers to consistency in test scores across different forms of the same test. This type of reliability is important to consider when educators create different forms of the same test (such as parallel forms for a pre- and a posttest) or when educators compare tests that have been altered in some way (e.g., items or directions have been changed). The process for estimating alternate form reliability is similar to test-retest reliability. Administer the two forms to the same group of students, but this time keep the time interval between testing periods short (e.g., within a week or so to measure the relationship between the parallel tests without respect to the time interval). The correlation between the two sets of scores is the

Illustrating the Importance of Reliability and Validity

Three tailor's assistants measure the waist sizes of clients for fitting pants and skirts. The assistants have slightly different ways of measuring clients' waists. This results in three slightly different measurements going to the tailor and three slightly different fits for the clients. One assistant may pull the tape tighter than the others; one may put his or her fingers inside the tape measure while the others are careful to keep their fingers outside of it. Even for the same assistant, unintentional differences in procedures from one fitting to the next may result in slightly different measurements and fits (e.g., measuring at the "true" waist versus measuring above or below the true waist). These differences in measurement are examples of reliability, specifically:

- Inter-rater: differences in measurement within each assistant
- Internal Consistency: differences in measurement for a single individual assistant

The tailor's business also offer dresses and suit jackets. For fitting jackets, the assistants take shoulder and arm measurements. Client complaints illustrate different types of validity.

- Content Validity: The jackets did not fit well because no chest measurements were taken.
- Relationships with Other Variables and Outcomes: The jackets did not fit as well as the jackets made by the tailor down the street.
- Consequential Validity: Some clients alleged they were harmed by the poor fits. A bride and groom asked for their money back after their groomsmen complained the jackets didn't fit and they were forced to wear their own clothes to the wedding instead.

alternate form reliability estimate. This form of reliability will only be important for APPR assessments if districts and BOCES are using alternate forms of the same assessment for the pre- and posttest administration or if districts are changing a test form from one year to the next and want to ensure that the revised form is consistent with the earlier form.

- *Inter-Rater Reliability* refers to the degree of consistency among raters who are rating the same performance. This type of reliability is especially important for the purposes of APPR assessments as they will likely contain performance items or portfolio responses that must be scored using a rubric or other scoring mechanism that relies on rater judgment. The process for estimating inter-rater reliability is to have multiple raters rate a student on some measure of performance and to evaluate the agreement between the raters. The goal in assessing this type of reliability is to ensure that regardless of the rater, students who perform the same will receive the same or a very similar score. Inter-rater reliability is enhanced when there are clear scoring materials, such as rubrics, and training for scorers. After these procedures are in place, inter-rater reliability must be monitored to ensure that scoring is conducted as expected. The simplest method for estimating inter-rater reliability is to use a concordance table that records the scores for two raters rating students on the same item. The agreements are recorded along the diagonal of the table. The simple percentage agreement is computed by dividing the number of times the two raters agree by the total number of items scored. As shown in Table 4, the two raters agree 47% of the time (14 ÷ 30), which is not a terribly high level of agreement. The reliability coefficient of .47 indicates that (a) more can be done to improve scoring consistency for this item, (b) the scorers need additional training, or (c) both.

Table 2. Concordance: Two Raters

	Rater 1					
		1	**2**	**3**	**4**	**Total**
Rater 2	**1**	4	2	1	0	7
	2	3	5	3	1	12
	3	0	2	2	0	4
	4	1	1	2	3	7
	Total	8	10	8	4	30

These four types of reliability are summarized in Table 3.

Table 3. Four Types of Reliability

Type of Reliability	Description	Type of Reliability Coefficient	When to Establish This Type of Reliability
Internal Consistency	The degree to which the items are measuring a similar set of content in the same way	Coefficient Alpha or similar	When the assessment is comprised of many items representing the same content area
Test-Retest	The degree of stability of scores over time, estimated in the absence of instruction in the content area	Correlation between the scores at Time A and Time B	When an assessment is used to establish pre- and posttest scores
Parallel Forms	The degree of similarity between two different but parallel forms of an instrument	Correlation between the test scores from Form A and Form B	When two parallel forms of an assessment are used or when the assessment is changed from one year to the next
Inter-Rater	The degree of consistency between two or more raters (scorers)	Simplest: Percentage of joint agreement using a concordance table; More Sophisticated: Cohen's Kappa, Correlational Methods	When the assessment contains items that require rater judgment

When selecting an assessment, it is recommended that districts and BOCES examine reliability evidence to determine the degree to which the assessment is providing stable and consistent measures.

Validity

Reliability is a required feature of a high-quality instrument. However, reliability alone is not enough; an instrument with high reliability may not be valid for a particular use. Validity is concerned with whether the inferences of an instrument are appropriate for the intended use. Below are three types of validity evidence. Validity is concerned with both the development and use of the assessment; as a result, districts and BOCES are responsible for ensuring the validity of the inferences of test results, just as test developers must present validity evidence for the assessment that they produce. Validity is an ongoing process, meaning that if an assessment is continuously used from year to year, investigations of validity must also take place over time.

- *Content Validity*. Content validity evidence demonstrates that the instrument content aligns with and samples appropriately from the intended content. The intended content for many educational instruments is specified in the district's curriculum scope and sequence or curriculum map. Determining content validity is an important up-front consideration in the selection of an assessment for APPR purposes The district should examine the alignment of test specifications to the curriculum of the district, and then determine how well the instrument represents both. Furthermore, alignment to rigor is necessary to determine if the cognitive (or other) demands represented in the curriculum map are apparent in the instrument itself and in the test specifications. Note that other characteristics of the instrument development process are also important in establishing content validity and development evidence, including the degree to which the instrument is administered in an appropriate and standardized way and that the time allotted to take the test and other administration conditions are also appropriate.

- *Relationships to Other Instruments, Outcomes, and Variables*. This type of validity evidence is collected when assessment results are related to the results of similar measures or other intended outcome measures. It can also be collected when results are unrelated to dissimilar measures and outcomes. For example, one would expect for scores on two general 2nd grade reading tests to be related (high scores on one test being associated with high scores on the other; low scores on one test being associated with low scores on another). The inverse can also be true. Validity evidence is accumulated if the instrument is less related or even unrelated to dissimilar instruments, measures, and outcomes. For example, one would not expect scores on a 2nd grade reading test to be strongly related to scores on a physical fitness test, and a weak relationship between the two also provides validity evidence. Similarly, validity evidence may be collected if scores are related to expected outcomes. For example, validity evidence would be accumulated if scores on math achievement at the end of 1st grade were related to scores on an end-of-year math test in 2nd grade. Predictive validity evidence would allow a district to use an assessment to predict the results on a future outcome, such as using an assessment to predict students' results on a state or national exam.

- *Validity evidence based on consequences*. This type of validity evidence is accumulated if the use of the scores is generally experienced as fair and beneficial for the students and other persons affected by the test results. To establish validity evidence based on consequences for an assessment, the instruments should be shown to contribute to student learning and to provide benefits to teachers. For example, results can be used to ensure the following:

    - Improvement of instruction to students

    - Realignment of the curriculum to provide all students with more opportunity to learn the key material

    - Provision of high-quality professional development opportunities for teachers

Table 4. Types of Validity Evidence

Type of Validity	Description	Question to Ask
Content Validity	The degree to which the content of the assessment aligns with the district curriculum at the expected level of rigor	Does the assessment represent the content and rigor of the instructional/curricular content?
Relationships to Other Measures, Outcomes, and Variables	The degree to which the scores are in agreement with or predict other tests and/or criterion	Is the assessment measuring what it is purporting to measure?
Consequential Validity	A comparison of the intended use(s) of the assessment to the intended and unintended outcomes of that use(s)	Does the assessment confer the intended benefits and reduce unintended harms for students and teachers?

Fairness

To be acceptable to teachers, students, the public, and other interested stakeholders, assessments must provide examinees from diverse backgrounds an equal chance to show what they know and can do. Numerous practices indicate that an assessment is fair, including whether all students have an adequate chance to demonstrate their knowledge during the assessment process, whether students had an ample opportunity to learn the content, and if the instrument is not biased (see definition of bias below). It is important for selected assessments to be perceived as fair by stakeholders, including educators, district administrators, and others.

Bias

The primary source of bias is item bias. Individual instrument items may perform in a biased way against specific groups by referencing persons, groups, experiences, or cultures that the examinees may be more or less familiar with. To avoid bias in item writing, test developers are advised to review the instrument content to ensure that the following guidelines are adhered to:

1. Selected language on the instrument holds the same semantic meaning for all examinees.

2. Selected language on the instrument communicates the intended affective (emotional) effects for all examinees

3. Stereotypical language is avoided, especially language that characterizes groups as more or less powerful, advantaged, smart, attractive, etc., than other groups.

4. To the extent that cultural or demographic groups are represented on the instrument, the representation attempts to acknowledge all groups.

5. Main characters in texts show good representation across cultural and demographic groups.

Specifically, instruments should be balanced based on gender, cultural, and other demographic factors, and should be inclusive of the groups of students taking them. While it is important to note that instrument items will draw from examinees' backgrounds in often unintended ways, it is the work of the instrument review team to ensure that the instrument consists of a reasonable range of experiences, group representations, and backgrounds. With respect to experiences, the instrument should include items that tap into common experiences for the group and not include experiences that favor one group over another (e.g., using more than one sports example when the group of examinees are not all athletes).

National Evaluation Systems (NES; 1991, pp. 4 and 15) provides examples of biased and nonbiased language in assessments to assist review teams in reducing instrument bias. Table 5 provides a few examples:

Table 5. Examples of Bias in Item Writing

	Poor	Better
Gender Bias	Identify the major stages in the evolution of man.	Identify major stages in human evolution.
Power and Status	Congress finally granted African Americans broad enforcement and protection of their right to vote in 1964.	After a long struggle, African Americans won legal enforcement and protection of their right to vote in 1964.
Power and Status	Many universities are now permitting retirees to enroll in degree programs.	Older persons are now enrolling in university courses and degree programs in ever increasing numbers.

A related problem is known as the *stereotype threat*[6] (Steele & Aronson, 1995). A stereotype threat is one in which examinee performances on an assessment is changed when examinees belonging to a particular subgroup are reminded that that subgroup performs better or worse on that particular type of task. For example, if a narrative prompt on an assessment describes negative impacts of poverty on student performance, the prompt can activate that stereotype in examinees who are receiving free or reduced-priced lunch, perhaps negatively affecting their performance on the exam.

Documentation

To ensure that assessments are used in an appropriate and standardized way, they are typically accompanied by documentation. This documentation ensures transparency in the development and administration of assessments and can include the following documents:

- *Technical manual*. An assessment's technical manual is a comprehensive technical document. It should identify the purpose of the assessment as well as when, how, and to whom it can be appropriately administered. The document should explain how the instrument content was identified and developed, specific requirements regarding the administration of the instrument, the process for scoring the instrument, the types of scores reported by the instrument, and information regarding the proper interpretation of scores. It should include information a potential user may need for determining the assessment's psychometric quality, such as reliability, validity, and bias analyses. The technical manual may also include other policies describing the appropriate use of the instrument, such as the training requirements for instrument administration and the interval of time before which the instrument must be reevaluated. Technical manuals are typically developed for commercial assessments; districts wishing to use commercial assessments for APPR purposes should consult the technical manual to review the instrument quality information reported there, including information about reliability and validity.

- *Administration manual*. This document details the instrument administration procedures. When followed closely, it standardizes the administration procedures, enhances instrument security, supports the equitable treatment of examinees, and minimizes errors in instrument administration and scoring. The instrument administration manual typically includes a list of the examinee resources (e.g., calculators or dictionaries) that are required and prohibited during administration, a description of the appropriate conditions for administering the instrument, a script that the

[6] For more information about stereotype threat, see http://www.reducingstereotypethreat.org/definition.html.

administrator reads to students, details regarding what can and cannot be said or done by students and by those administering or proctoring the assessment, instructions for timed tests, insights into how to deal with emergencies that may arise during a test, and a list of the examinee accommodations that are permitted (e.g., offering extra time or administering the exam in a quiet setting outside the classroom). It should explain clearly the procedures for scoring the instrument and procedures for training scorers to score the instrument items reliably. Finally, the instrument administration manual should include instructions to ensure instrument security, including procedures for accounting for instrument materials (e.g., how to check out and return instrument materials) and other administrative details (e.g., how to process or score answer sheets).

The content described above should be provided by the test developer; and if not, it can be requested from the test developer, as the information contained in these documents will help determine whether the assessment is of high quality and whether it is appropriate for APPR purposes.

Appendix: Resources

Assessment-Related Resources

American Educational Research Association, Psychological Association, & National Council on Measurement in Education. (1999). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

Educational Measurement: Issues and Practices – ITEMS: The Instructional Topics in Educational Measurement Series:
http://ncme.org/publications/items/

Darling-Hammond, L. & Adamson, F. (2010). Beyond basic skills: The role of performance assessment in achieving 21st century standards of learning. Stanford, CA: Stanford University, Stanford Center for Opportunity Policy in Education: https://scale.stanford.edu/system/files/beyond-basic-skills-role-performance-assessment-achieving-21st-century-standards-learning.pdf

Linn, R. L., & Gronlund, N. E. (1995). *Measurement and assessment in teaching* (8th Edition). Upper Saddle River, New Jersey: Prentice-Hall Inc.

National Evaluation Systems, Inc. (1991). Bias issues in test development. Amherst, MA: National Evaluation Systems, Inc.

Shute, V. J. (2011). Stealth assessment in computer-based games to support learning. *Computer Games and Instruction*, 55(2), 503-524. See: http://myweb.fsu.edu/vshute/pdf/shute%20pres_h.pdf

Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of personality and social psychology*, *69*(5), 797. See
http://mrnas.pbworks.com/f/claude%20steele%20stereotype%20threat%201995.pdf

Teacher's College Early Elementary Performance Assessment Resources:
http://readingandwritingproject.com/resources/assessments/performance-assessments.html

Resources Related to New York State APPR

Section 3012-c of the Education Law can be found at:
http://public.leginfo.state.ny.us/LAWSSEAF.cgi?QUERYTYPE=LAWS+&QUERYDATA=$$EDN3012-C$$@TXEDN03012-C+&LIST=LAW+&BROWSER=EXPLORER+&TOKEN=38733959+&TARGET=VIEW

The regulations that implement Education Law §3012-c can be found at:
http://www.regents.nysed.gov/meetings/2012Meetings/March2012/312bra6.pdf.

Revisions were made to the regulations that implement Education Law §3012-c at the June 2013 and the February 2014 meetings of the Board of Regents. The revised regulations can be found at:

http://www.regents.nysed.gov/meetings/2013Meetings/June2013/613p12hea1.pdf and
http://www.regents.nysed.gov/meetings/2014/February2014/214p12hea1.pdf and
http://www.regents.nysed.gov/meetings/2014/March2014/314brca11.pdf

The New York State Education Department will provide additional or updated guidance as necessary on its website, www.nysed.gov. See: http://www.engageny.org/sites/default/files/resource/attachments/appr-field-guidance.pdf for updated APPR guidance.

Other NYSED Early Elementary APPR Assessment Resources

NYSED information on early elementary assessments and the K-2 assessment pathways document can be found at: http://www.engageny.org/resource/early-elementary-assessments

Acknowledgement

This document was partially based on *Technical Guide A: Considerations Regarding District-Determined Measures*, Massachusetts Department of Elementary and Secondary Education (http://www.doe.mass.edu/edeval/ddm/)