PENNSYLVANIA DEPARTMENT OF EDUCATION

Local Assessment Validity Evaluation Handbook

For the Keystone Examination System

Developed by the National Center for the Improvement of Educational Assessment in Collaboration with the Pennsylvania Department of Education

11/17/2011

Pennsylvania Local Assessment Validity Evaluation Handbook

Local Assessment Validity Evaluation Handbook

The purpose of this handbook is to help districts in Pennsylvania understand the requirement for local end-of-course assessments for the purpose of meeting state graduation requirements set forth by the Pennsylvania Department of Education (PDE). This handbook includes how-to discussions of various important components, lists of evaluation criteria, and sample submissions and evidence needed to meet the independent evaluation standards. It pulls information from many resources, including manuals from the states of Maine, Nebraska, Rhode Island and Wyoming. It includes excerpts of papers written by staff from the National Center for the Improvement of Educational Assessment (www.nciea.org). This is meant to provide technical assistance, not a research paper, therefore each reference is not cited individually, however all must be acknowledged.

This handbook is divided into five parts. The first three parts corresponds to the three main parts of the test development cycle. Part four provides information on the submission and evaluation process, including examples.

Note: The term "district" used in this document applies to all LEAs.

Part I: Test Development

The criteria used to evaluate the test development process will focus on two aspects: **alignment** and **fairness**.

Alignment is generally defined as a measure of the extent to which the content standards and assessments agree, and the degree to which they work in conjunction to guide and support student learning. This criterion is not a simple determination but is a considered judgment based on a number of factors that collectively determine the degree of match between content standards and the assessment which will gauge how well students are achieving those standards. In other words, does the local assessment do an effective job of measuring the knowledge and skills set forth in the eligible content PDE developed for each Keystone subject?

Fairness involves the interaction of the assessment with the individual student. To be considered fair, an assessment must provide each student with relatively equal opportunities to appropriately demonstrate what he or she knows and are able to do. The evaluation requires results from alignment studies and evidence of test design as well as review procedures to minimize, detect, and eliminate bias from the assessment.

Part II: Achievement Standards

Determining appropriate rigorous achievement standards will be a focus of the evaluation of the validity of local assessment systems. The achievement standards for the Keystones include the performance

level descriptors (PLD) and cut scores for each subject. There are four achievement levels for the Keystones: **Below Basic, Basic, Proficient** and **Advanced**.

Districts choosing to develop local assessments do not have to use the same levels, but must distinguish proficient performance from performance that does not meet that mark. However, the proficient PLDs developed for the Keystones must be used to define proficiency on the local assessment. The process of developing other PLDs, if the full set of Keystone PLDs is not adopted, will be scrutinized. In addition, the process used to set the cut scores on the local assessment will be evaluated to verify that it followed a documented, validated procedure and produced appropriately rigorous achievement standards.

Part III: Technical Quality

The technical quality of the assessments will also be examined to ensure the test produces consistent results across students and from one assessment administration to the next. The reliability of the results is very important to ensure that the state and districts are consistent in the requirements they are asking students to meet. Any tests that include tasks that are graded by teachers must include evidence that a second teacher would provide the same or similar rating. Tests that include items that are updated each year will require proof that the new items are of the same difficulty from one year to the next.

Part IV: Evaluation Process

This section will provide the evaluation criteria, the submission template and sample submissions. Examples of evidence that will need to accompany the submission are referenced in this section and expanded upon in the next section.

Part V: Evidence

This final section will provide examples of the kinds of evidence needed to show the validity of the assessment. Districts will be able to use these examples as models in drafting their own evidence for local assessment validation.

Part I: Test Development

Alignment

Alignment is generally defined as a measure of the extent to which a state's standards and assessments agree and the degree to which they work in conjunction to guide and support student learning. This criterion is not a simple determination but is a considered judgment based on a number of factors that collectively determine the degree of match between a state's standards and the assessment which will gauge if students are achieving those standards. *In other words, does the assessment do an effective job of measuring the knowledge and skills set forth in the content standards for each course/examination?*

A typical alignment study strives to answer several key questions:

- Is there a strong **content match** between the test items (and the test as a whole) and the state's content standards (as described in the assessment anchors and eligible content)?
- Are the test items (and the test as a whole) more rigorous, less rigorous or of comparable **rigor** to the state's content standards?
- Is the **source of challenge** for test items appropriate? That is, is the content the item is assessing the hardest thing about the item? There should not be, for example, an underlying factor such as an difficult algebra demand embedded in a measurement item; or the need for extensive background knowledge of a topic in order to answer reading comprehension questions.
- Are the **text passages** for the reading assessments of appropriate length and complexity for this course?
- To what degree does this set of items (or test) reflect the **balance** of content and performance delineated in the corresponding content standards, assessment anchors or eligible content for the course/examination?

Norman Webb's Dimensions of Alignment

1. Categorical concurrence: The same or consistent categories of content appear in both the content standards and the test items.

> Do the same "categories" of knowledge and skills appear in both the standards and the assessment? For example, does the mathematics test include math skills, math concepts and math problem solving if all three "categories" are included in the state's mathematics standards? For Pennsylvania: to what degree are the Keystone assessment anchors represented by test items?

2. Depth-of-Knowledge (DOK): What is elicited from students on the assessment is as demanding cognitively as what students are expected to know and do as stated in the standards.

> Do the assessment items/tasks reflect the cognitive complexity of the concepts and processes described in the standards? In other words, are the assessments as

cognitively challenging as the standards? This is generally the most overlooked alignment component. It requires careful analysis to determine if the assessment is accurately targeting the depth of knowledge levels called for in the standards. For the Keystones, the minimum DOK allowed for any item is a DOK 2. Information will be available about the DOK of each assessment anchor and eligible content. The local assessment should include items that either match or exceed the DOK levels included in the Keystones.

3. Balance of Representation: The extent to which items are appropriately distributed across standards.

Do the assessments (or item bank) reflect the same degree and pattern of instructional emphasis (in terms of content and skills tested) as found in the state's academic content standards?

4. Range-of-Knowledge: The criterion for correspondence between span of knowledge for a standard and the assessment (or item bank) considers the number of assessment anchors within the standard measured, with at least one related assessment item/activity.

Will the assessments yield scores that reflect the full range of achievement implied by the Keystone assessment anchors? Alignment should occur at the anchor level or finer grain at the eligible content level to ensure that the strands and standards are appropriately sampled. This does NOT mean that each eligible content should be assessed separately; rather it simply means that the assessments should be built from a blueprint that reflects the appropriate weight of each assessment anchor.

Pre- and Post-alignment

Generally, those developing an assessment should think about alignment before and after development. When designing the assessment, consider creating a test blueprint that shows how each of the assessment anchors and eligible content will be measured—by multiple-choice items, short-constructed response items, an essay or a performance task. Using the blueprint to connect the eligible content to the items will help ensure strong categorical concurrence. Determining the number of items to write to each anchor will lay the ground rules for the balance of representation. Finally, consider the depth-of-knowledge each item should represent. Spelling all of these points out in a document will satisfy the requirement for a test specification or blueprint document. (See example A1, for instance.)

Next, focus on the individual items or tasks. Clearly specify which assessment anchor and eligible content that item is to address, the type of item, the number of points it will be worth and the depth of knowledge it is meant to measure. These details are considered item specifications. Use an item specification document as the basis for providing instructions to assessment writers. "Another consideration when designing and writing a local assessment is the use of universal design. Universal design principles address policies and practices that are intended to improve access to assessments for all students. Universal design principles are important to the development and review of assessments because some assessment designs hinder a student from illustrating their skills and knowledge. Universal design techniques can result in a more accurate understanding of what students, especially students with disabilities, know. To ensure that an assessment complies with the principles of universal

design, each item needs to be written to respect the diversity of the population to be assessed and be sensitive to test taker characteristics and experiences such as gender, age, ethnicity, socioeconomic status, region, disability and language. Directions to item writers should encourage them to avoid content that might give an unfair advantage or disadvantage to any student subgroup. This would include minimizing the effects of extraneous factors like avoiding unnecessary use of graphics that cannot be presented in Braille, using font size and white space appropriate for clarity and focus and avoiding unnecessary linguistic complexity when it is not being assessed. Using universal design considerations in the directions to item writers will also cover one of the fairness criteria described more fully in the next section.

Next, each item should be reviewed for content, depth-of-knowledge and fairness. Having a separate group of teachers review each item will provide additional evidence of the alignment. Creating written direction for the item reviewers, even a checklist of features to look for (positive or negative) will satisfy the criteria under alignment. Even beyond this step, a formal alignment study will be required by PDE. The process for an alignment study is described below.

General Methodology for Conducting an Alignment Study

Phase I Review: Item-by-Item Analysis

To determine how closely each item (of a given test form) is aligned to the related grade-level content standard, teams of educators who are knowledgeable of the content and skills expected for that grade will first review (1) the content match and then (2) the DOK match of each test item. Working in pairs, educators review and code each item using a simple coding system.

- **Reviewer Question 1 - Content Alignment: Is there a strong content match between the test items and the state's content standard?** Reviewers should focus on the "nouns" in each assessment anchor/eligible content to get to the concepts or skills being assessed.

  Materials: state curriculum documents, coding templates and test items for each course reviewed

F = test item **fully** addresses or exceeds the content (e.g., mean, mode, and median addressed in a single item; all literary elements addressed in a single item) described in the corresponding assessment anchor.

P = test item **partially** addresses the content described in the corresponding assessment anchor. Often, there will be content/skills assessed in the test item that address only a narrow aspect described in the assessment anchor or eligible content for the course. For each item coded as "partial" content match, coders make notes as to which part is assessed in the item. (Use far right column on coding sheet for comments about partial content matches.)

No match = test item **does not** address the content described in any corresponding assessment anchor for this course. These will be items that might assess (a) related content, but not content described in any eligible content; (b) content included for instruction and assessment at a higher or lower level course, but not this course; (c) content that is related, but exceeds the assessment limits for this course;

or (d) unrelated content. **"No match" items** are listed at the end of each section of the coding sheets with notes as to why the items did not align to course assessment anchors

During phase II of the review, each item coded as "partial" content match must go through a second, holistic review to determine whether the "set" of items corresponding to the assessment anchor now fully assess all aspects. Notes about these items should identify aspects that will facilitate the phase II review.

- **Reviewer Question 2 - DOK Alignment: Are the test items more rigorous, less rigorous, or of comparable rigor to the states' standards/assessment anchors?** Reviewers should focus on the "verbs" in each objective to understand how students will demonstrate understanding of the concepts or skills being assessed with each test item.

  Materials: DOK content-specific charts with descriptors, coding templates and test items for each course reviewed

To determine how closely each item (on a given test form) is aligned to the intended DOK of the assessment anchors and eligible content, pairs of educators review the same items again. They use a content-specific Depth of Knowledge Levels Table and the intended DOK levels of the assessment anchors and eligible content to determine whether the cognitive demand of the test item is more rigorous, less rigorous, or of comparable rigor to the corresponding description in the assessment anchors/eligible content (from the content match).

"Intended" DOK levels of the state's assessment anchors and eligible content should be listed next to corresponding objectives in the coding forms prior to completing the study. Sometimes, more than one DOK might apply. These are listed to assist reviewers, but may not include all DOK possibilities. Example:

Content Area: Literature		**Alignment** List items by # notes go in far right column	
Strand/Standard 1.0 Reading Comprehension	***State's Intended DOK***	**Content "F or P or No" coding**	**DOK "F-P-0" coding**
3.a. Use context to determine the meanings of words	**2a -** Use context cues or resources to identify the meaning of unfamiliar words	# #	# #
4.b. Identify and explain what is directly stated in the text	**1d-** Locate or recall facts or details explicitly presented in text **2d-** Recognizing appropriate generalizations about text (e.g., possible titles, main ideas)	# #	# #

To complete the DOK review for each test item, raters determine the DOK of the item and compare it to the "intended" DOK of the content objective to which it is aligned. One of three possible ratings is recorded on the template for each item:

F= test item **fully** addresses or exceeds the DOK intended in the corresponding assessment anchor and eligible content

P= test item **partially** addresses the DOK intended in the corresponding assessment anchor and eligible content. For each item coded as a "partial" DOK match, coders make notes as to which part orDOK is assessed in the item (use far right column on coding sheet for notes about DOK). For example, if the intended DOK could be a level 1 (e.g., identify a pattern) and or a level 2 (e.g., extend a pattern), one test item might address only one DOK level of the two or more possible DOK levels.

0= test item **does not** address the DOK intended in any aspect of the corresponding assessment anchor and eligible content. Generally this is when there is a content match with the item, but the item has a lower DOK level (e.g., recall) or has a higher DOK level (e.g., justify an answer) for the content described in the anchor.

Make notes for partial and no DOK matches at the far right, by listing the actual DOK level of the test question.

No DOK notations are made for items that were already coded as "no match" for content.

- **Reviewer Question 3 - Analyze Source of Challenge: Is the source of challenge appropriate?** In other words, is the hardest thing about the test item that which is targeted for assessment in the item? There should be VERY few notations here. This is NOT about what makes the question hard to answer; but whether something else besides the content of the question is making the question difficult to answer. For example, the source of challenge in social studies or reading might include the need to understand cultural references or background information not assessed in the test item, but important in understanding the test item itself. **Source of challenge analysis can be done while reviewing alignment questions about content or DOK**.

Important Notes for the Phase I Review:

- Each coding recorded represents agreement of the two raters reviewing the test questions. Ask the facilitator for assistance if you cannot reach agreement.
- Each test question is only aligned to one assessment anchor and eligible content to avoid double counting items. If a test question "might" align to more than objective, select the best objective to use for alignment purposes.
- Coding must include the test item number, since it will be important to know if the information about all items was reviewed and recorded; this information will be used in the phase II analyses (e.g., does more than one item for this objective assess a DOK level 2 when the assessment anchor and eligible content ceiling is a DOK 3?).
- Items that do not match any assessment anchor or eligible content are listed at the end of each section (by strand or standard) with a notation – which might include something such as "content not assessed in this course," etc.
- Once phase I review has been completed, go back and check to be sure all items have been recorded. For example, if 70 items were reviewed, check to see that each test question, by numbers one through 70, have been noted on the coding forms.

- Complete the last page of the coding forms with totals for each standard for content and DOK alignment findings. This is the item-level summary for alignment.

Phase II Review: Holistic Analysis (analysis of the test form or item bank)

Phase II of the alignment study uses results from the item-by-item analysis in phase I to review holistically items as a set or as a complete test. A review of an item bank is not the same as actually reviewing a test form because different teachers or a computer will: "select" different individual items; use a different overall total number of items; or have more or less emphasis on certain items depending on the purpose(s) for creating an assessment for classroom use.

When there is no actual test form to review, it is important to know what percentage of items is actually being used to make the determination of the degree of alignment for a given exam. If two simulated test forms are reviewed, each having 40 items and there are 200 items in the item bank, then interpretations will not be as solid as if all test forms, or most of the available test items, were analyzed.

Three different levels of grain size can be analyzed holistically during Phase II.

Three holistic analyses can be used in this phase of the study to determine the alignment of items in the item bank to all or some of the state's standards, assessment anchors and eligible content.

A. **To what degree are all assessment anchors in the state's content standards, for this exam, represented in the item bank or test form?**

B. **To what degree are the assessment anchors and eligible content assessed on the end-of-course assessment represented in the item bank or test form?**

C. **To what degree do individual teachers identify items (of those items reviewed in phase I) as having the potential to be selected for an end-of-year review for that exam?**

The process for both holistic analyses is to review each "set" of items aligned to each assessment anchor and eligible content, and then consider the alignment to each content strand or standard as a whole.

- **Reviewer Question 4A - Content Alignment: Is there a strong content match between the set of items reviewed and all assessment anchors included in the state's content standards for this course?**
- **Reviewer Question 4B - Content Alignment: Is there a strong content match between the set of items reviewed and the state's assessment anchors assessed on the state assessment for this course?**

A "strong content match" means that OVERALL, the set of sample test questions fully assesses all, or almost all, aspects of the state standards for the specified course. A "moderate content match" means that while there was strong alignment to some aspects of the standards or assessment anchors, other aspects were either not assessed at all or not fully assessed. "Weak/no content match" means that test items assess few or no aspects of the assessment anchors, or do not assess the aspects as described in assessment anchors and eligible content.

General Overall (holistic) Content Decision Guidelines

Strong Alignment – if 70-100 percent alignment with standards and assessment anchors and eligible content

Moderate Alignment – if 40 percent – 69 percent alignment with standards and assessment anchors and eligible content

Weak/No Alignment – if zero – 39 percent alignment with standards and assessment anchors and eligible content

- **Reviewer Question 5A – DOK Alignment: Overall, is the set of test items more rigorous, less rigorous, or of comparable rigor to <u>all assessment anchors</u> included in the state's content standards for this exam?**
- **Reviewer Question 5B – DOK Alignment: Overall, is the set of test items more rigorous, less rigorous, or of comparable rigor to the state's <u>assessment anchors and eligible content assessed on the state assessment</u> for this exam?**

Reviewers determine to what degree the DOK range for all aligned items – from DOK "ceiling" to any lower DOK levels for assessment anchors - is assessed by the set of items aligned for content.

General Overall (holistic) <u>DOK</u> Decision Guidelines

More Rigorous – if most items reviewed are at a higher DOK level than standards and assessment anchors and eligible content

Similar Rigor – if most items reviewed are similar to the DOK range of standards and assessment anchors and eligible content

Less Rigor – if most items reviewed are at a lower DOK level than standards and assessment anchors and eligible content

- **Reviewer Question 6 - Text Complexity Review (for ELA committees only):** Support materials: articles on increasing text complexity and text structure (http://www.nciea.org/publications/TextComplexity_KH05.pdf)

Using descriptors listed for each grade span, identify those descriptors that apply to *most or all* of the text passages in the test form.

Important Reminder: Use the item review coding from phase I to focus the phase II review on the alignment of the set of items (or the entire test) as a whole, not individual items.

Claims about Content

Because the local assessment must provide data on a student's readiness for college and careers that is equally good or better than the Keystone exams, it is important to focus on the claims related to the content. When the assessment is adequately aligned, we can expect the following claims to be true:

- The items on the local assessment represent the content standards to the same breadth and depth as the Keystone items; and
- The content coverage of the local assessment is aligned with the Keystone assessment.

Fairness

Fairness is a major design consideration in the development of local assessments. The validity of the inferences from an assessment system are threatened when students do not have relatively equal opportunities to appropriately demonstrate what they know and are able to do, or when the design or interpretation of the test is otherwise unfair to certain students or groups of students. Fairness concerns judgments about an individual in isolation, judgments about an individual in relation to others, and judgments about groups of students. The central question regarding fairness in assessment is:

> *Does the assessment system provide all examinees an equitable opportunity to demonstrate their acquired knowledge in the content area of the assessment?*

To the extent this question may be answered affirmatively, one may claim fairness in the assessment system. Fairness encompasses technical matters but also matters of policy. A district may not be able to achieve perfect fairness for every individual, but that should be the goal. Evidence will need to be provided that every effort was made to ensure an equitable opportunity to all students in the district.

The issue of fairness is interwoven throughout the assessment process. The design considerations discussed previously all have a hand in ensuring fairness. An assessment that is well-aligned, has undergone a defensible standard-setting process, and is consistent in its scoring and categorization as an assessment that should be most fair to all stakeholders. Indeed such conditions should be considered prerequisites to fairness, as the absence of any one would create an assessment situation that disadvantages at least a portion of examinees in a tangible way.

Fairness in assessment is about both process and product. Steps must be taken to ensure fairness in development and administration, as well as in the inferences drawn from test scores. Both the Code of Fair Testing Practices in Education (2004) and the American Education Research Association/American Physiological Association/National Council on Measurement in Education Standards include obligations for ensuring fairness to test takers. The" Standards" also address obligations to ensure fairness through all stages of test development, test administration and test use. Taking steps to assure fairness begins with defining the purpose and aims of the assessment. The purpose of the assessment should align with the policy context in which it resides. The assessment should be expected to do no more than achieving the informational needs specified by the assessment system policy. With purpose and aims in place, addressing fairness then centers on communication with stakeholders. All stakeholders should be made aware of the purpose of the assessment and how scores from the assessment will be used. Rules and expectations (e.g., articulation of proficiency standards) must be communicated to all students, teachers, school administrators, parents and anyone else who stands to be impacted by the assessment. There are three main areas of focus for the fairness criterion: item development, administration and reporting.

Item Development

When people talk about fairness in an assessment, they often discuss test content. Items make up the "meat" of the test and are the primary source of interaction between examinees and the assessment system. Surely the goal is to avoid situations where items are designed specifically to favor one group of

students over another or where such favoritism is known but disregarded. Unfairness in test items can occur in more subtle ways. Examinees come from varied backgrounds and therefore may react to or interpret test items in ways unintended by the test developers. Consider the example question, below:

> *Joe challenged his older brother, Dave, to a race of 100 meters. Joe had a head start of 3 seconds and ran at an average speed of 5 meters per second. If the brothers finished in a dead heat, what was Dave's average speed?*

This question contains an idiom, "in a dead heat," that could confuse some students, particularly English language learners, and limit their ability to demonstrate knowledge of the algebraic skill that is of interest here. To minimize these unintended situations, each item developed for the assessment, before it is placed on a live test, should be reviewed by an independent panel representing diverse backgrounds and perspectives. Suspect items should be revised or eliminated all together.

Applications should show evidence that item writers were trained on writing items free of bias. Ideally, tests will be developed using universal design principles to make them fair to all. Universal design is an approach to assessment development based on principles of accessibility for a wide variety of end users. Thompson, Johnstone, and Thurlow (2002) describe seven elements of universally designed assessments: inclusive test population; precisely defined constructs; accessible, non-biased items; tests that are amenable to accommodations; simple, clear and intuitive procedures; maximum readability and comprehensibility; and maximum legibility.

Another element that would support the claim that the items are fair to all is evidence of a review for bias and sensitivity. This review, conducted by a diverse group of educators and stakeholders, helps ensure that items are evaluated from diverse viewpoints, including different income levels, ethnicity, races, and genders. Likewise, at least some of the reviewers should be familiar with issues of English language learners and students with disabilities. Documentation should include the characteristics of the review panel, the training given to the item reviewers and the results of their analysis.

Test Administration

Test administration protocols should establish an environment free of distractions, so that examinees have a relatively equal opportunity to demonstrate their abilities. Further, security measures that reduce opportunities for cheating and ensure examinee confidentiality provide the greatest likelihood that test scores can be interpreted as representative of the test takers or examinee's ability. Adherence to test administration protocols also facilitates comparisons among test scores of different students or groups. An examinees' scores can be interpreted against a common standard with greater confidence when the tests are completed under comparable testing conditions. Another way to ensure fairness in test administration is through the appropriate use of accommodations and alternate assessments. For students with disabilities or students for whom English is not the primary language, the standard administration protocol may not provide sufficient opportunity to demonstrate their abilities. On the other hand, using inappropriate accommodations or modifications may create unfair testing situations. Additionally, opening access to students with severe disabilities, via alternate assessments, can provide

these students the opportunity to demonstrate knowledge that is generally thought to be beyond their reach.

For the submission, evidence could include test security protocols, monitoring procedures, documentation of any accommodations used that are not part of the state accommodations manual, and the administrators guide to the assessment.

Reporting

Finally, local districts should be mindful of issues of fairness in their plan for disseminating assessment results. As with communicating the purposes of the assessment, communication of results should reach all stakeholders. That means that individualized reports should be available to the student, parent, teacher, and principal; and aggregated results should be available to the district superintendent as well as the general public. Focusing on the individual student, the reporting process should allow for equitable opportunity for remediation.

Statistical analysis aimed at evaluating fairness begins with the disaggregation of results across various student groups. We discussed above that it is difficult to assess fairness at the individual level and much of that difficulty is due to the requirements of common statistics. Statistical analyses generally depend on groups containing enough individual units to derive stable estimates of the qualities of interest. Even non-statistical evaluations cannot account for the myriad of unique qualities held by all of those involved in an assessment system. Therefore, to evaluate fairness we must aggregate individuals into groups of relative homogeneity. Such commonalities should represent characteristics that plausibly could impact test performance. When disaggregating results by group, it is important to think beyond the common social categories; of race, ethnicity, and gender, as they do not represent the only classifications across which unfairness may arise. Other classifications, such as primary language spoken in the home, documented learning disabilities, and free or reduced lunch eligibility reflect a student's social context and could impact test performance.

More must be done, however, than simply examining differences across groups in terms of average performance. A difference in performance between two groups could indicate unfairness in the content of the test, but alone it is not enough evidence to support such a claim. Instead, performance differences are more likely to result from deeper problems residing in the educational system that provide certain groups greater opportunities to learn tested content. In these cases, differential test performance is an accurate reflection of difference in opportunity to learn and not an unfair assessment. A better statistical analysis would consider the difficulty of each test item for each group in relation to that group's average performance on the overall assessment. Wide disparities, such as an item-correct proportion 20 percent below other groups for a group that scored around the overall average, indicate items that may possess some form of unintended bias. Likewise, correlating the score on an individual item to the total score for each student group will highlight items that seem overly difficult or easy for specific groups of students.

Exhaustive analyses can detect a phenomenon termed differential item functioning (DIF), which indicates an item that "functions" differently for different student groups. An old example was on the

SAT anagrams section where the anagram related to military hardware. Boys significantly outperformed girls on that question even though girls outperformed boys on the overall anagram section. That item was determined to show DIF favoring boys. Such analyses require specialized software packages, statistical acumen and labor resources that would often be beyond what is available to a local district. However, evidence of some statistical analyses and their results should be included. Information about which incorrect option a student chose could provide helpful information to students and teachers, as well as identification of any distractor that is showing an undue influence in a specific student group. In addition, the reporting template and any accompanying student or parent guides should be part of the submission. Finally, any policies regarding retakes, remedial instruction, and auditing or verifying scores challenged by students should be documented.

When fairness is adequately provided in an assessment, one can expect the following claims to be true:

- Test scores across all identifiable and relevant student groups will have comparable interpretations with respect to the course content area; and
- All identifiable and relevant student groups receive equitable treatment within the assessment system.

References and Suggested Readings

Alignment

Bhola, D. S., Impara, J. C., & Buckendahl, C. W. (2003). Aligning tests with states' content standards: Methods and issues. *Educational Measurement: Issues and Practice (fall)*, 21-29.Collins, A. (1998).

Bloom B. S. (1956). Taxonomy of Educational Objectives, Handbook I: The Cognitive Domain. New York: David McKay Co Inc.

Hess, K., (2008). Exploring cognitive complexity and depth of knowledge. Dover, NH: Center for Assessment. Available at www.nciea.org

Porter, A. C. (2002). Measuring the content of instruction: Uses in research and practice. *Educational Researcher, 31*(7), 3-14.

Webb, N.L. (1997). *Criteria for Alignment of Expectations and Assessments in Mathematics and Science Education*. Council of Chief State School Officers and National Institute for Science Education Research Monograph 6. Madison, WI: University of Wisconsin–Madison, Wisconsin Center for Education Research.

Webb, N.L. (1999). *Alignment of Science and Mathematics Standards and Assessments in Four States* (Research Monograph No. 18). Madison, WI: National Institute for Science Education.

Webb, N.L. (2002). Alignment Study in Language Arts, Mathematics, Science, and Social Studies of State Standards and Assessments for Four States. Washington, DC: Council of Chief State School Officers.

Webb, N.L. (2002). Technical issues in large-scale assessment. Washington, DC: CCSSO

Fairness

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

Camilli, G. (2006). Test fairness. In R. L. Brennan (Ed.), *Educational measurement (4th ed.;* pp. 221-256). Westport, CT: American Council on Education/Praeger.

Code of Fair Testing Practices in Education (2004). Washington, DC: Joint Committee on Testing Practices.

Council of Chief State School Officers. (2003). Quality Control Checklist for Item Development and Test Form Construction. Washington, DC: CCSSO. Available at http://www.ccsso.org/Documents/2003/Quality_Control_Checklist_2003.pdf.

Johnstone, C. J., Altman, J., & Thurlow, M. (2006). A state guide to the development of universally designed assessments. Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Thompson, S. J., Johnstone, C. J., & Thurlow, M. (2002). *Universal design applied to large-scale assessments* (NCEO Synthesis Report 44). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Zieky, M. (2006). Fairness review in assessment. In S. M. Downing & T. M. Haladyna (Eds.) *Handbook of Test Development*. Mahwah, NJ: Lawrence Erlbaum Associates.

Part II: Achievement Standards

An important step in assessment development is the crafting of achievement standards. The achievement standards are written descriptions of "how good is good enough." Achievement standards are also referred to as proficiency levels. As described in the Pennsylvania regulation, local assessments are required to include "performance level expectations and descriptors that describe the level of performance required to achieve proficiency comparable to that used for the Keystone exams." This section will provide best practices and examples on how to meet that goal.

Establishment of Proficiency Levels

Establishing proficiency levels is an integral part of creating a criterion-referenced test; that is, one where the goal is to demonstrate a certain level of knowledge and skill regardless of the performance of others. The process results in levels used to report student scores and communicate to stakeholders whether the performance meets certain criteria.

First, start by developing a common vocabulary. To be clear, **content standards** are statements of the knowledge and skills that students are expected to learn; **proficiency levels** indicate the degree of mastery required. In other words, content standards are the what; proficiency levels are the how much or how well. Proficiency levels are also referred to as **performance standards** or **achievement standards**. However, the term "standards" has many meanings in educational discussions. We have the standards for educational and psychological testing, often called simply "the standards" (AERA, APA, & NCME, 1999). Every Keystone exam has content standards that include assessment anchors and eligible content. Due to concerns that performance or achievement standards will become confused with other types of standards, we have chosen to call them proficiency levels here.

A proficiency level typically is defined as the minimally adequate level of performance for some purpose (Kane, 1994) or the level of performance that is expected of examinees (Hambleton, 2001). A proficiency level consists of three components: the name of the level, a written description of the level, and a minimum cutoff score; a fourth component, exemplar items or sample student work at each level, is optional but very helpful.

Proficiency Level Descriptors

Readers can skip this section if they are adopting the Keystone proficiency level descriptors (PLD) as they are, or read further to learn more about intended purposes or uses.

Policymakers developing local assessment systems must first decide whether or not to adopt all of the Keystone performance levels as their own proficiency levels. Adopting the performance level for proficiency is required, but districts have the option of developing additional levels specific to their own assessment. If they choose to adopt all of the Keystone proficiency levels as is, then the process of naming the levels and writing descriptions is already complete. Local jurisdictions that adopt the Keystone performance levels would meet satisfactory requirements for the claim regarding the level of rigor of the proficiency levels. If, however, the local policymakers wish to craft their own proficiency

levels, they must then provide evidence that the process used to draft the descriptions was sound and that the resulting descriptors are equally or more rigorous than the Keystone descriptors.

While these PLDs are most commonly associated with setting cut scores, they are also a useful development and reporting tool. The descriptors state, in words, what the cut scores mean and can help students, teachers, and parents interpret what their students know and can do and, in turn, what they do not know and cannot do. Ideally, if an assessment program has clearly defined levels and purposes for using those levels, the levels should be designated early in the test development process. It should be clear to those designing and using the assessment what levels of performance are expected from the students and how those levels are distinguished from one another in terms of knowledge and skills. That way, items can be written to clearly distinguish among the levels and ensure a more reliable categorization of student abilities.

First, policymakers will need to determine the number and name of the levels. Currently, the Keystone Exams use four levels. It can become difficult to describe meaningful differences across more than four levels. In addition, any particular test has a fixed amount of measurement power that depends primarily on the number and quality of the questions in the test. The more cut scores there are in any given test, the less measurement power the test developer can devote to each cut score and the less information there is around each cut score. In addition, the greater the number of performance levels, the greater the work required to produce PLDs and cut scores. The level of effort required can soon become unmanageable.

Once the number and names of the levels have been determined, the descriptors need to be written. Ideally, policymakers will convene a small group of content experts for a PLD-writing workshop. The policymakers need to explain their expectations for the rigor of each level, particularly as related to the rigor of the Keystone levels. To develop PLDs, content experts start with those expectations and add specific knowledge, skills, and abilities required at each level for each subject. PLDs should be built from test content, either in the form of content standards, eligible content or blueprints, depending on when in the process they are being written and what is available at that point. The test items can also be used as supplemental information to help develop the descriptors. Care should be taken, however, to ensure that the descriptors are not written to address a specific item. Rather, they should list the knowledge and skills required to answer that item correctly and others like it. It is important to keep in mind that test items are periodically replaced for security purposes. Therefore, we do not want descriptors that are specific only to the test form that was operational when the descriptors were written. In other words, the PLDs should describe the more general knowledge and skills that the test items are designed to measure rather than the knowledge and skills of the specific items.

Often, the language in a descriptor relies on models of cognitive processing, such as those defined in Bloom's taxonomy. That is, a lower level of performance may include words such as "identify" or "describe" while a higher level of performance may include words such as "analyze" or "evaluate." There is much research on the progression of learning and instruction that focuses on the type and quality of knowledge (c.f., Jong & Ferguson-Hessler, 1996). For example, consider the differences between concrete and abstract knowledge or among declarative, conceptual and procedural knowledge.

Facilitating a discussion on the hierarchy of cognitive learning with those on the PLD writing committee may help them distinguish among levels while still addressing similar content. Following this approach would most likely result in descriptors that reflect a similar breadth of content but different depths of knowledge and understanding.

The process of developing and reviewing the descriptors should be documented and presented as evidence of the validity of the process. The table below provides a checklist for local assessment developers to consider when developing and reviewing the descriptors.

Table 1. Checklist for Reviewing Proficiency Level Descriptors (PLDs)

	PLDs are generalizable across test forms (i.e., no part of the PLD is written to a specific item)
	Terms are clear and understandable (particularly to teachers)
	Format of PLDs is parallel across levels and courses (e.g., paragraph or bullet format; main ideas in same order across levels)
	Additional PLDs are clearly aligned to the Keystone course standards
	Each major component (e.g., assessment anchor) is addressed in every PLD
	Every knowledge, skill or behavior listed in the PLD is measured on the assessment
	Verbs are concrete (e.g., use CAN rather than MAY or SHOULD)
	PLDs clearly reflect greater difficulty and complexity from one performance level to the next (i.e., Advanced is more rigorous than Proficient which is more rigorous than Basic)
	PLDs clearly reflect greater complexity from one sequential subject to the next (e.g., Proficient in Algebra II is clearly more rigorous than Proficient in Algebra I)
	Most distinctions among levels are made using concrete differences in knowledge and ability rather than just in frequency or consistency of application (e.g., avoid use of rarely, sometimes or usually distinctions)

Cut Scores

Regardless of whether the Keystone PLDs are adopted as is or new ones are written, the cut scores associated with those PLDs on the local assessment need to be determined. Thus the next step of establishing proficiency levels is to run a cut score study to obtain recommendations for the location of the cut scores. A cut score is "a point on a test's score scale used to determine whether or not a particular test score is sufficient for some purpose" (Zieky, Perie, & Livingston, 2008). In this case, that "purpose" is to determine mastery of coursework important to college and career success. All methods of setting cut scores are based on human judgments. Zieky et al. categorizes cut score methodologies into four broad classes depending on the kinds of judgments that the participants make: judgments about test items; judgments about patterns of subscores (i.e., profiles); judgments about individual people or the products made by those people; and judgments about groups of people. There are also some compromise methods that combine absolute and normative judgments.

Most state assessments (including those in Pennsylvania) rely on judgments about test items to set their cut scores. Methods based on judgments about test items focus panelists' attention on the content of the test item and incorporate judgments from a large number of panelists. Regardless of the method chosen, it is important to elicit judgments from a representative pool of panelists, including content experts, teachers who work with special populations (e.g., students with disabilities and English language learners), and teachers who represent fully the demographics of the district.

The two most popular types of methods based on judgments of test items are the modified Angoff and Bookmark. Detailed procedures for conducting a cut score study based on either of these methods can be found in Zieky, Perie, & Livingston, 2008, and Cizek & Bunch, 2007.

Modified Angoff

Modified Angoff is arguably the most widely used and well-researched method for setting cut scores. Typically, panelists are asked to state the probability that a student who is just barely proficient (or just barely basic, advanced, etc.) would answer each test item correctly. Each probability is treated as an expected score on a 0-1 scale. So, if a barely proficient student is expected to have a 50-50 chance of answering an item correctly, a 0.50 is recorded for that item. The probabilities of all the items are summed to calculate an expected score for the just barely proficient student. This expected score then becomes the initial cut score for that panelist. The expected scores are then averaged across panelists to determine the recommended cut score.

The Modified Angoff method can be extended to work with open-ended items that are scored polytomously (e.g., with possible scores of 0–3, 1–6, etc). Instead of stating the probability that a just barely proficient student would answer the item correctly, panelists estimate the average score that a large group of just barely proficient students would obtain on the item. This average score does not have to be an integer. For example, if an essay is scored on a scale from 1–6, one participant might estimate that a group of just barely proficient students would obtain an average score of 3.5, while another might estimate the average score to be 4.2. Again, these numbers are summed along with the probabilities for the multiple-choice items to get the total expected score for the just barely proficient student, which is the cut score for the proficient level.

Bookmark

The Bookmark method was developed to be used with tests that are scored using Item Response Theory (IRT). It is now one of the most widely used methods for setting cut scores on state K–12 assessments. To use this method as it was designed, one must have a test that was calibrated using IRT. In addition, one must also have a statistician available who knows how to use IRT and who has access to the software required for the necessary calculations.

The panelist is given a special test booklet called an *Ordered Item Booklet* that displays the questions in order of difficulty from easy to hard. The participant's task is to place a bookmark at the spot that separates the questions into two groups—a group of easier questions that the just barely proficient student would probably answer correctly (with probably meaning a chance of at least 2 out of 3 or .67), and a group of harder questions that the borderline test taker would probably not answer correctly (i.e.,

the test taker would have a probability of less than .67 of answering correctly). After the panelist has made this placement of the bookmark, the statistician can calculate the expected test score for the barely proficient student, which again becomes the cut score for proficient.

Methods for setting cut scores on multiple pieces of evidence

One option for districts creating local assessments is to bring in multiple pieces of information to gauge student's mastery of the course materials. For instance, students might take a short multiple-choice (MC) test on the required factual knowledge, complete a performance task to show their understanding of procedures and write an essay analyzing a more complex component of the subject. The district would then need to determine how to set proficiency levels on these multiple pieces.

One option would be to set cut scores on each piece separately and then determine what patterns of performance are required on each piece. For example, let's assume that the MC test is worth 20 points (20 items worth 1 point each), the performance task is scored with a rubric worth a total of 10 possible points, and the essay is scored with a 6-point rubric. We could decide that a "proficient" student must receive at least 15 out of 20 points on the MC test, at least a 7 out of10 points on the performance task and at least 4 out of 6 points on the essay.

Another option would be to weight each component. For example, maybe we decide that the performance tasks is the most important component, so we multiply that by three to make it worth a total of 30 points. Then, maybe the essay score is also multiplied by three to make it worth 18 points. Now we have a complete test worth a total of 68 possible points (20 MC + 30 performance + 18 essay). Then, we could simply determine the number of points out of 68 a student must receive in order to be categorized in each of the proficiency levels.

The first example is what we call a conjunctive method while the second is called a compensatory method. That is, in the first example, a high MC score cannot overcome a low essay score, but in the second example, because the cut score is based on a total score a high MC score could compensate for a lower essay score.

Once policymakers determine how they want the pieces to be combined and weighted, there are several methods that can be used. We encourage district assessment staff to find one of the two books mentioned (Cizek & Bunch, 2007, or Zieky, Perie, & Livingston, 2008) and read chapters on the following methods: Body of Work, Analytic Judgment and Dominant Profile.

Other methods

There are many other methods besides the two described in the earlier sections or the three mentioned in the previous section. A handful of states use a contrasting group approach to set or evaluate cut scores. This approach is based on judgments of students and on the idea that students can be divided into two contrasting groups on the basis of judgments of their knowledge and skills: a group that is proficient and a group that is not. For example, teachers who do not yet know their students' test scores could categorize their students as either Proficient or not based on the PLD and their observations of the student in their classroom. After large numbers of teachers have provided these judgments, a score can be calculated that best separates the two groups.

All methods should be researched carefully and one selected purposefully. Matching the method to the test is an important step in developing strong proficiency levels. First examine the composition of the test. Is it comprised primarily of multiple-choice or open-ended items? Next, consider the type of scaling that will be done. Will IRT be used to calculate scale scores or will results be reported out as raw scores? Also consider the pool of panelists. Are there enough to consider a method based on judgments of students? Chapter 5 of Zieky, Perie, and Livingston (2008) provides a complete list of questions to consider when selecting a method.

Other considerations

Regardless of the method chosen, the rationale for choosing a method needs to be documented. The approach to running the cut scores study also needs to be documented with sections detailing:

- Selection of panelists
- Training given to panelists
- Number of rounds of judgments collected
- Feedback given between rounds
- Data provided to assist with judgments (e.g., p-values or impact data)
- Evaluation of the cut score study

Each of these steps is important and needs to be done well to establish strong proficiency levels. More details can be found in either of the two books mentioned previously. In addition, Hambleton (2001) provides a good discussion of evaluating a cut score study and the establishment of proficiency levels. In addition, the Pennsylvania Technical Advisory Committee prepared a document summarizing the information that should be presented in a standard setting plan and technical documentation, which is appended to this handbook. Any Pennsylvania district developing a local assessment should follow those recommendations.

Exemplar Items

Once the PLDs have been written and used to establish cut scores, they can be fleshed out even further through the use of exemplar items; that is, when the cut scores are known, psychometricians can identify items that students at one proficiency level are likely to answer correctly and that students in the lower proficiency level are unlikely to answer correctly. Different criteria have been used to identify exemplar items, such as the p-value of an item for students scoring proficient must be at least 0.65 and must be at least 0.35 higher than the p-value of the same item for students in the lower level. An item's location on a difficulty scale based on IRT also can be used to identify items that fall in the middle of a performance level. Describing common characteristics of these items or including 2 to 3 of these items with each PLD can add a richness and depth to the final PLD and may provide valuable interpretive information for teachers, students and parents. These descriptions can be updated periodically throughout the local assessment program as new items are released.

Conclusion

If proper procedures are followed to develop PLDs and set cut scores, and if those procedures are well documented, the proficiency levels should meet the satisfactory criteria given in the evaluation matrix. Specifically, the evidence should support the following claims:

1. The local assessment system maintains an adequate level of rigor in the proficiency levels; and
2. Judgments of student proficiency are set using a researched and established methodology.

References and Suggested Reading

AERA, APA, NCME. (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests.* Thousand Oaks, CA: Sage.

Hambleton, R. H. (2001). Setting performance standards on educational assessments and criteria for evaluating the process. In G. J. Cizek (Ed.) *Setting performance standards: Concepts, methods, and perspectives* (pp. 89–116). Mahwah, NJ: Lawrence Erlbaum.

Jong, T. & Ferguson-Hessler, M. (1996). Types and qualities of knowledge. *Educational Psychologist, 31*, 105–113.

Perie, Marianne. (2008). A guide to understanding and developing performance level descriptors. *Educational Measurement: Issues and Practice, 27*(4) pp. 15-29.

Zieky, M., Perie, M., & Livingston, S. (2008). *Cutscores: A manual for setting performance standards on educational and occupational tests*. Princeton, NJ: Educational Testing Service.

Part III: Technical Quality

The technical quality of an assessment is vital to ensuring that scores are accurate and reliable, that is that they could be reproduced under another testing condition. For this manual, we have focused on "consistency" as the primary component of technical quality. Consistency, also known as reliability, refers to consistency in scores across items or tasks, scorers, forms and years. For example, if a person was weighed twice, one minute apart on the same scale and got different weights, the accuracy of the scale would probably be questioned. LLikewise, if a student shows very different performance on two forms of the same test, that are meant to measure the same knowledge, might call the consistency in the forms into question. In the same sense, if the same student essay received two different scores from two scorers, it might raise questions about the clarity of the rubric, or the qualification or training of the scorers. All of these factors must be analyzed to show that the technical quality of the local assessment is high.

Consistency across Items or Tasks

"Internal-consistency reliability" is a measure of the consistency of a student's performance across the items or tasks in an assessment. A formula known as "KR- 21" is a convenient short-cut method for estimating the internal consistency of tests made up of multiple-choice items (or any items scored correct or incorrect).

KR-21 Calculation

KR-21 = [N/(N-1)] [1 - {M(N-M)}]/(N*V)

where:

N = number of items in the test
M = arithmetic mean of the test scores
V = variance of the raw scores

A statistical cousin of KR-21, "Cronbach's alpha," can be used for assessments that are made up of items or tasks scored using multiple values, such as a short-answer question worth four points.

Cronbach alpha Calculation

$$\alpha = \frac{K}{K-1}\left(1 - \frac{\sum_{i=1}^{K}\sigma_{Y_i}^2}{\sigma_X^2}\right)$$

K = number of items
σ_X^2 = variance of the observed total test scores
$\sigma_{Y_i}^2$ = variance of component *i* for the current sample of persons.

KR-21 and Cronbach's alpha are appropriate for assessments where all items or tasks tap a similar skill, competency, or ability, such as algebra. If a test measures more than one skill, such as an informative essay compared to a persuasive essay, other calculations are needed. One option is to calculate either of those two specifics for the items that measure the same skill (e.g., items measuring only vocabulary separate from items measuring comprehension). Another option is to employ the "split-half" calculation. This method requires dividing the test in half (e.g., odd items vs. even items), scoring the two halves separately for each student, and correlating the two sets of scores. This correlation, after a minor adjustment, is the "split-half reliability" of the complete test. All of these reliability statistics can be calculated through a simple computer program, such as Statistical Package for the Social Sciences or even using a spreadsheet, such as Excel.

Split –half Reliability Calculation

$$\text{Reliability of scores on total test} = \frac{2 \times \text{reliability for } ½ \text{ test}}{1 + \text{reliability for } ½ \text{ test}}$$

Once the calculations have been made, the next question is how high should the reliability be? This is a difficult question to answer and depends largely on how high the stakes are for test takers. For classroom tests that are only one score in a series, acceptable reliability levels are often at 0.60 or above. For very high stakes tests, like a college admissions test, a reliability of 0.90 or higher is required. Due to these local assessments being used in part for graduation decisions, test designer should aim for a reliability of 0.90 and not accept anything below 0.80.

Consistency across Scorers

For assessments that call for human ratings, such as essays, portfolios or performances, reliability can be established by determining the amount of agreement among the scorers, also known as raters. This form of reliability is known as "inter-rater reliability." As an example, suppose two raters independently evaluated 50 essays using a six-point scoring rubric. One inter-rater reliability index, the "percentage of exact agreement," reports the percentage of students who receive the same score from both raters. A less stringent criterion is the percentage of agreement within one point (e.g., a student receives a 5 from one rater and a 4 from the other).

High agreement between raters is the goal; exact agreement should be above 0.70 and adjacent agreement levels should be close to or above 0.90. Where agreement is low, further study is required to identify and reconcile the source(s) of non-agreement. Generally, low agreement between scorers indicates a problem with the criteria (ambiguous, unclear), the scorers' application of the criteria (introducing a bias or not understanding the process), or both. The former problem would require revisiting the rubric or scoring criteria while the latter problem would require improving the training and monitoring of the scorers.

Consistency across Forms

"Equivalent-forms reliability" is determined by administering two equivalent, or parallel, forms of an assessment to the same students and then establishing the similarity between the two sets of scores. For instance, if a district wants to create a second form of a test to give to anyone who was absent on testing day to reduce the possibility of cheating, the district then needs to ensure that the regular form and make-up form are equivalent. Like inter-rater reliability, this form of reliability can surface either as a percentage or as a coefficient. Let's consider the percentage first. In the example of a regular form and a makeup form, a student judged to be proficient (or not) on the basis of one form of the assessment should be similarly judged on the basis of the other. To examine reliability in this regard, the superintendent could ask a sample of 40 students to take both forms of the test and then calculate the percentage of students who receive the same proficiency judgment on the two occasions. A high percentage indicates that the test yields similar judgments regarding proficiency, irrespective of which form of the test is taken. That is, the test is reliable. The example above can be modified slightly to illustrate the use of Pearson for establishing equivalent-forms reliability. Imagine that this test was worth a total of 80 points. Each student in the superintendent's sample thus has two scores: one from each form. The correlation between the two sets of scores is an expression of "equivalent-forms reliability": the higher the correlation, the greater the similarity in a student's relative performance on the two forms, and hence the greater the reliability of the proficiency test.

Consistency across Years

Consistency across years typically involves two variables: ensuring that a new form given the next year is equivalent to the form given the previous year; and ensuring that scorers are applying the criteria with the same level of rigor. For the first condition, a district could use the equivalent forms approaches listed in the previous section. Alternatively, they could only replace a few items each year and use the items that are the same across the two forms to calculate whether the new items introduced any changes in difficulty. As an example, let's say in the first year students scored an average of 40 points on 60 items. In the second year, they scored an average of 42 points on 60 items. Did the students really perform better in year 2 or was the test easier? If 50 items were the same in the two years and only 10 items were different, we could compare the performance on the 50 items in year 1 to year 2 to see if the students' performance changed. If not, then we would assume that the 10 new items were easier than the 10 old items, and we would need to adjust the scores to make them equivalent to the previous year.

For tests with constructed-response items that are scored by human raters, we need to be sure that the scorers are applying the same judgments from one year to the next. One method for ensuring equal rigor is to give them papers from the previous year and ask them to score them. The goal is for the ratings to be the same from one year to the next. If they are not, some retraining of scorers will be needed.

References and Suggested Reading

Airasian, P. W. (1994). *Classroom assessment (2nd ed.)*. New York: McGraw Hill.

Allen, M.J., & Yen, W. M. (2002). *Introduction to measurement theory*. Long Grove, IL: Waveland Press.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297-334.

Cronbach, Lee J., and Richard J. Shavelson. (2004). My current thoughts on Coefficient alpha and successor procedures. *Educational and psychological measurement 64*, no. 3 (June 1): 391-418.

Devellis, R.F. (2011). *Scale development (3rd ed.).* Thousand Oaks: Sage Publications.

Linn, R. L. & Gronlund, N. E. (2000). *Measurement and evaluation in teaching (8th ed.)*. New York: Macmillan.

Popham, W. J. (1999). *Classroom assessment: What teachers need to know (2nd ed.)*. Boston: Allyn & Bacon.

Salvia, J. & Ysseldyke, J. E. (1998). *Assessment (7th ed.)*. Boston: Houghton Mifflin.

Part IV: Evaluation Process

Each LEA is required to assemble a submission packet that will be evaluated by an independent evaluator. The directions for assembling the packet and sample submissions are provided in this section. The evaluation criteria are provided in Exhibit 1. First, however, it is important to understand what a validity evaluation is and why it is important for ensuring that the scores truly represent students' actual knowledge and skills. First, we start with a validity primer and then provide submission templates, instructions and examples. Sample evidence to accompany the submission is discussed in the next chapter.

Each submission will be evaluated on four criteria: alignment, proficiency levels, consistency and fairness. These four categories have been thoroughly discussed in the previous sections of this handbook. Exhibit 1 provides the rubric on which the evidence will be judged. Each submission must score at least satisfactory on all four dimensions in order for the local assessment to be approved as an alternative to the Keystone Exams.

Exhibit 1. Evaluation Criteria

Evaluation Criteria	Superior	Satisfactory	Insufficient
Alignment	In addition to the evidence characterizing the satisfactory level: • Evidence of depth of knowledge alignment from results of "think-aloud" protocols or other similar analyses • Evidence from an external alignment study • No gaps in coverage of the standards, all items/tasks are aligned to specific standards, and depth of knowledge represented by the items/tasks matches the expectations for depth of knowledge in the standards	• Documentation of adequate sampling of all content standards • Evidence from an internal alignment study that used a two-way alignment process • Few gaps in the coverage of the standards, all of the items/tasks are aligned to specific standards, and there is a range of depth of knowledge (including DOK 4) represented by the items/tasks • Plans for periodic review of alignment	• Items represent content standards, but many standards are unaddressed • The content standards are represented well, but the depth of knowledge required to correctly answer items is not in alignment with the standards

Evaluation Criteria	Superior	Satisfactory	Insufficient
Fairness	In addition to the evidence characterizing the satisfactory level: • Universal design principles were adhered to in developing the assessment • Assessment results are communicated in a manner that allows for equitable remediation opportunities • Analysis of distractor choices across student groups (for multiple-choice items) • Disaggregated results show no large discrepancies between total scores and item difficulties	• Procedures are in place to ensure that the items allow individuals from all subgroups to demonstrate their knowledge • Documentation from bias and sensitivity reviews show the items are free of noticeable bias • Accommodations and alternate assessments are provided as needed/appropriate • Performance expectations are communicated clearly to all stakeholders • The district produces and examines results disaggregated by student groups to search for differences in opportunity to learn • Test administration and security protocols ensure that all students experienced an equitable test environment	• Review procedures are in place, but lack the sophistication to dependably detect potential bias • Results are not disaggregated by important student groups (e.g., ones identified by the state on state-level report cards)

Evaluation Criteria	Superior	Satisfactory	Insufficient
Establishment of proficiency levels	In addition to the evidence characterizing the satisfactory level: • Evidence that items represent a wide enough range of difficulties so that the assessment may provide adequate information across the range of cut scores • A plan for evaluating the appropriateness of cut scores once data is available from the assessment (predictive association) • The process for establishing proficiency levels involved individuals from a diverse representation of roles within the school community • Sample items are included in the descriptive information regarding each proficiency level	• The process for establishing proficiency levels followed a researched and validated methodology and documentation of the process is provided • A convincing rationale for the chosen method used to recommend cut scores is provided • Panelists had knowledge of the content and were demographically representative of all potential panelists in the district • Performance level descriptors are written to a level equally or more rigorous than Keystone's (adoption of Keystone descriptors is adequate)	• The performance level descriptors are not as rigorous as to the Keystone descriptors • Percent correct or course grade measures define the cut scores • The cut scores are either too idealistic or too lenient (i.e., they do not conform to the performance level descriptors) • Reasonable cut scores have been advanced, but documentation of the process for establishing proficiency levels is lacking
Evaluation Criteria	**Superior**	**Satisfactory**	**Insufficient**
Consistency	In addition to the evidence characterizing the satisfactory level: • A plan for ongoing calibration of raters' scores to ensure that raters don't become more rigorous or more lenient from one year to the next • Test equating procedures ensure comparable test difficulty across forms and/or years • Inter-rater agreement and internal consistency (whichever is applicable) far exceeds minimum requirements	• Evidence is presented for measuring inter-rater agreement on open-ended items and internal consistency (i.e., reliability) on closed-ended items • Numbers meet minimum requirements for inter-rater agreement and/or internal consistency • Evidence of training for consistency within and across years for scorers of open-ended items (if applicable) is presented • A plan for periodic review of the equivalence of test difficulty across forms and/or years exists	• Inter-rater agreement and/or internal consistency is too low to support the uses of the assessment results • Inter-rater agreement was not calculated or numbers were not provided • Only one rater was used for every open-ended item (i.e., zero percent read behind)

Validity Primer and Introduction to Producing Evidence for a Validity Evaluation

Validity is defined as the "degree to which evidence and theory support the interpretations of the test scores entailed by proposed uses of the test" (AERA, NCME, & APA, 1999). In 2002, Kane described an approach that works with those proposed interpretations to develop a validity argument. Working backwards from the proposed interpretations, we develop a series of claims and assumptions that must be true for the interpretation to be valid. We then provide evidence and data to support each assumption and claim to show that our proposed interpretations are valid. This approach asks test developers to think of reasons why the intended inferences might not be supported. Basically, we state what we think the test does and then try to disprove it. That is, if student scores are higher in the second year than in the first, does that mean that they know more or that they and their teachers are more familiar with the tests? We want to prove that students actually know more but we prove it by trying to disprove it or the alternatives. If we disprove the alternatives and cannot disprove our desired assumption, then we have evidence that the test does what we say it does. In practice, it is not possible to search for all the reasons, but the framework provides us guidance for developing studies that refute other possible explanations for a finding. Shown below are the proposed interpretations and claims for the Pennsylvania Local Assessment System—first in a list, then as a diagram in figure 1.

Proposed interpretations:

- The local assessment provides data on a student's readiness for college or careers that is equally good or better than the Keystone Exams.
- Proficiency scores on the local assessments are equally or more rigorous than proficiency scores on the Keystone exams and cover equivalent material.

Alignment claims

- The items on the local assessment represent the content standards to the same breadth and depth as the Keystone items.
- The content coverage of the local assessment is aligned with the Keystone assessment.

Fairness claims

- Test scores across all identifiable and relevant student groups have comparable interpretations with respect to the course content area.
- All identifiable and relevant student groups receive equitable treatment within the assessment system.

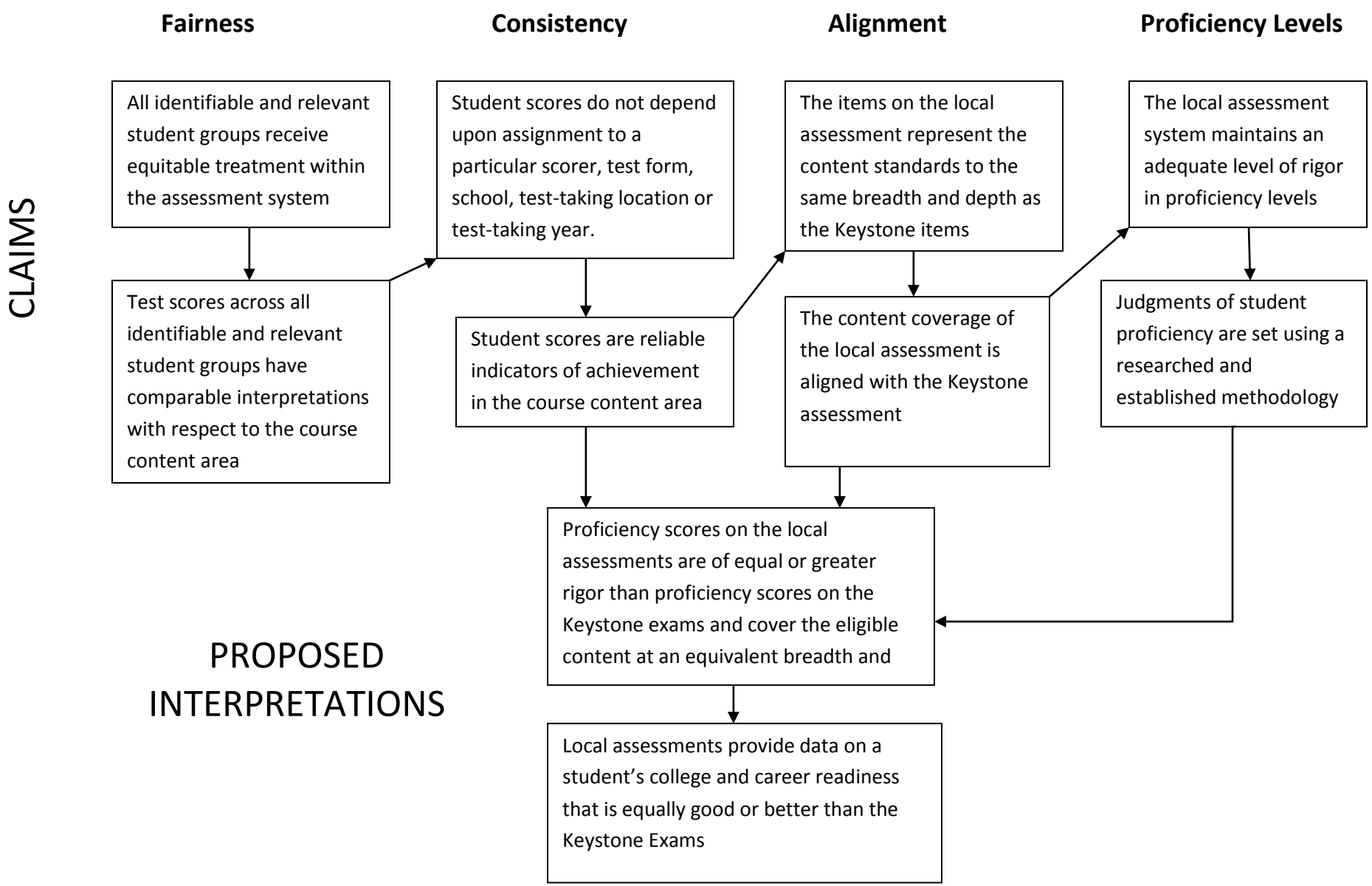
Establishing proficiency levels claims

- The local assessment system maintains an adequate level of rigor in the proficiency levels.
- Judgments of student proficiency are set using a researched and established methodology.

Consistency claims

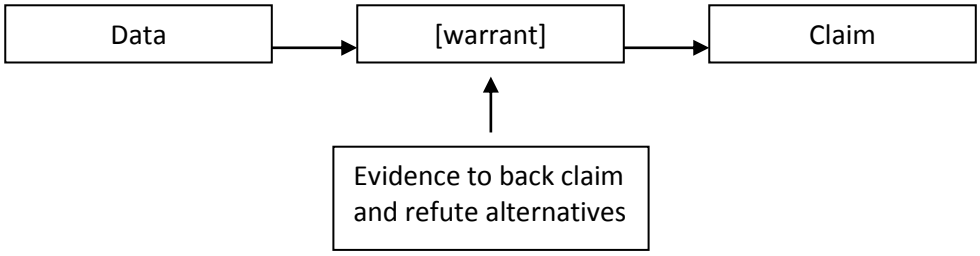
- Student scores do not depend upon assignment to a particular scorer, test form, school, test-taking location or test-taking year.
- Student scores are reliable indicators of achievement in the course content area.

Figure 1. Relationship between Validity Claims and Proposed Interpretation



Each of the claims is further explicated by the statements in the evaluation criteria matrix. The evaluation criteria are meant to help the districts determine what type of evidence is needed. Figure 2 derives from the general principle underlying Kane's (2006) work, which asks us to provide the data, warrant and claim. That is, we start with data and determine how it provides evidence to support a claim.

Figure 2. Toulmin's model adapted from Kane (2006)



Remember the evidence that proves an alternative claim is untrue, also supports the claim. Consider for example the statement "an increase of student scores reflects a greater understanding of the content." An alternative hypothesis could be that "an increase in student scores reflects greater teaching of test-taking strategies." Collecting evidence both to refute the second claim as well as to support the first would strengthen the validity evaluation.

The goal of each validity evaluation submission is to provide evidence for each of these claims. Thus for each claim, the submitter should provide evidence and an explanation of how that evidence supports the claim. Again, to strengthen the evaluation, evidence refuting alternative hypotheses will strengthen the application. An application package from PDE will include a template with three columns—one for the data, explanation of how it supports the claim (or refutes an alternative hypothesis), and the claim the evidence supports—with only the third column completed, the LEA will be responsible for completing the first two columns.

For example, let's start with the alignment claim: The items on the local assessment represent the content standards to the same breadth and depth as the Keystone items. Evidence could include training materials for item writers and item reviewers, as well as an external alignment study. We are looking for evidence that that the items are fully aligned with the assessment anchors and that all of the assessment anchors are covered by the assessment anchors. Likewise, we are disproving the claim that certain anchors were omitted or reduced in importance compared to what was intended with the Keystones. The sample templates on the following pages show how it could be completed for evaluation. In addition, the actual instructions for item writers and reviewers would be included, as well as the full report from the external alignment study.

Template A: One Test Supplants the Keystone Exam

Alignment

Data or Evidence	Explanation of how it supports the claim	Claim
		The items on the local assessment represent the content standards to the same breadth and depth as the Keystone items.

Establishment of Proficiency Levels

Data or Evidence	Explanation of how it supports the claim	Claim
		The description for Proficient was adopted directly from the Keystones; other levels and descriptions were added appropriately.
		The process for establishing the Proficient cut score was appropriate for the test and implemented effectively.

Fairness

Data or Evidence	Explanation of how it supports the claim	Claim
		Items were developed to be free of bias against any student group.
		Test administration procedures ensure all students experience an equitable testing environment.
		Reports provide applicable data and information for all student groups.

Consistency

Data or Evidence	Explanation of how it supports the claim	Claim
		Scoring procedures are implemented and monitored to ensure reliability of scores.
		Procedures are in place to ensure that the tests are equivalent in difficulty from one year to the next (or one form to the next).

Suggested Evidence and Instructions for the Completion of Template A

The following tables are populated with ideas for evidence that can be attached, as well as factors to consider when writing the narrative for the middle column. Evidence is ordered appropriately from basic evidence that is required to extra documentation that will move a district into the "superior" column. Note: The items in **BOLD** are required.

Alignment

Data or Evidence	**Explanation of how it supports the claim**	**Claim**
Include specific evidence that shows that the items are aligned to the course content standards. This evidence could include: • ***Test blueprint or specifications*** • *Item specifications* • ***Written instructions for item writers*** • ***Written instructions for item reviewers*** • *Sample tasks of high DOK* • ***Alignment study done by district or school staff*** ○ ***Technical report explaining process and results*** ○ ***Matrix of items to course content standards*** • *External alignment study* ○ *Technical report explaining process and results* • *Research studies examining the constructs tested by each item, such as a cognitive lab or think-aloud studies*	*Describe how the evidence submitted shows that the local assessment matches the course content standards (assessment anchors and eligible content) and/or the test blueprints for the Keystones. Be sure to discuss matching both the breadth and depth of knowledge of the target course content standards. Demonstrate that you've matched EVERY content standard with sufficient balance of representation and are testing a range of depth through DOK Level 4.* *If using a unique testing approach, showing the content measured might require additional evidence, such as an external alignment study or research evidence of student knowledge tapped by the assessment.*	*The items on the local assessment represent the content standards to the same breadth and depth as the Keystone items.*

Fairness

Data or Evidence	Explanation of how it supports the Claim	Claim
Include as many as possible: • ***Description of policies and procedures to ensure test items are not biased against any student group*** • *Evidence of a bias review committee meeting* ○ *Notes from meeting* ○ *Steps taken as a result of the meeting* • *Evidence of universal design for learning (UDL) procedures in item development* ○ *Instructions for writers* • *Statistical analyses of item difficulty across various student groups* • *Statistical analysis of distractor choice across student groups*	*Demonstrate that your item development process included multiple steps to ensure that they were free of bias, e.g., train item writers on universal design technique; include a bias and sensitivity review committee comprised of members of multiple racial/ethnic groups and representatives of students with disabilities and English language learners in the review process; and/or conduct statistical analyses of items by student group. At least one procedure should be undertaken and documented. The more that is done, the higher the score will be.*	*Items were developed to be free of bias against any student group.*
• ***Written criteria for taking local assessments versus alternative pathways to graduation*** • ***Evidence that local accommodations policy follows PDE policy*** • ***Test administration and security protocols*** • *Test administration and security monitoring plans*	*First, demonstrate that students are treated equitably during test administration. Note that equitable is not necessarily equal. For instance, some accommodations may help some students access the test but not others. Any deviations from standard PDE accommodations policy should be approved by PDE prior to submitting the validity evidence, and the approval from PDE should be included in the packet.* *Provide evidence that IEP teams are given similar guidelines for graduation pathways as in the other districts of Pennsylvania or provide a rationale for the difference.* *Demonstrate that students across the district follow the same administration and security protocols to avoid giving certain students an undue opportunity to cheat. Describe procedures for monitoring test administration and*	*Test administration procedures ensure all students experience an equitable testing environment.*

	security protocols – to ensure that all students have equal access to the test and none are unduly influenced.	
• ***Sample score reports*** • *User guides to score reports relating scores to graduation requirements* • *Reports communicating expectations to students, teachers and parents* • *Policy regarding appeals process for disputed student scores*	*Show how the reports provide ample information to all stakeholders. If the district has a large population of Spanish-speaking students, provide an explanation of what will be translated. Describe how expectations and results to students, parents, and teachers will be communicated.*	*Reports provide applicable data and information for all student groups.*

Establishment of Proficiency Levels

Data or Evidence	Explanation of how it supports the claim	Claim
PLDs and evidence that they were adopted by your local district. *Description of how they were adopted (who decided whether the Keystone descriptors should be adopted as is; if new descriptors were developed, how were they developed and by whom?)*	*Show that the description for Proficient was adopted directly from the Keystones and that other levels and descriptions were added appropriately. The description of Advanced or Basic could be more rigorous than the Keystones. If the definitions are less rigorous, explain how they will be used.*	*The local assessment system maintains an adequate level of rigor.*
Standard-setting technical report that includes: • ***Name of method used to set cut scores*** • ***Rationale for selecting that method*** • ***Composition of standard-setting panel*** • ***Evidence that the process followed a research-based and documented application of that method*** • ***Evidence of internal consistency of panelists (e.g., convergence in cut scores over rounds)*** • *Results of an evaluation of panelists about their comfort level with the process and results*	*A complete standard-setting technical report should minimally net a satisfactory score on this dimension. Follow the model provided in the PDE Technical Advisory Committee (TAC) document appended to the validity handbook. Show that a documented, validated method was used and provide evidence that it was implemented effectively. Document the range of panelists (representing different student groups and geographical areas in the district) and the rationale for any decisions made about the procedures.*	*The process for establishing the Proficient cut score was appropriate for the test and implemented effectively.*

Consistency

Data or Evidence	Explanation of how it supports the claim	Claim
Description of scoring procedures, including: • ***Rubric*** • ***Scoring process*** • ***Number and qualification of scorers*** • ***Training of scorers*** • ***Monitoring of scorers*** *Data showing inter-rater agreement on scoring of open-ended items*	*If the test is all selected response (multiple choice), document the scanning procedures and the quality control procedures to ensure the key is correct.* *If the test includes open-ended items, it is important to show that they are scored reliably. Document the review of the rubric, selection of scorers, training of scorers, tests of the scorers' accuracy and reliability, and continual monitoring of the scores. The goal is to demonstrate that the scores themselves are calculated accurately.*	*Scoring procedures are implemented and monitored to ensure reliability of scores.*
Calculations of internal consistency of multiple-choice items. *Other reliability statistics*	*There are several statistical methods to show internal consistency of the assessment. Select one and provide the calculation. Also, consider showing the item/test correlation for all items to show their contribution to the total score.*	*Student scores are reliable indicators of achievement.*
Description of procedures used to ensure comparable difficulty of items/forms over time • ***Judgmental*** • ***Statistical*** • ***Formal equating***	*Explain the process to refresh the item pool while maintaining comparability over time.* • *Will human judgment determine a replacement item is of the same difficulty as the old item?* • *Is there a statistical procedure?* • *What procedure will be in place to ensure scorers score open-ended items with the same rigor from one year to the next?*	*Procedures are in place to ensure that the tests are equivalent in difficulty from one year to the next (or one form to the next).*

Example of a Submission under Template A

The following provides a simple example of a submission that would be satisfactory, assuming the attachments provided information that fully met the evaluation criteria.

Introduction:

We created a local assessment for our students in chemistry for the following reasons:

1. We wanted to maintain control over the integration of the assessment score into the grade.
2. We felt that a strong science assessment should include a performance component whereby students integrate their understanding of scientific procedures with their knowledge of a specific content area; therefore, our assessment contains 25 multiple-choice items focused on the content knowledge, five open-ended items that focus more on the process, and one performance task that requires students to implement scientific procedure to investigate an issue in the chemistry content area. Students who choose this subject as one of their graduation requirements must score Proficient or above to graduate.

Alignment

Data or Evidence	Explanation of how it supports the claim	Claim
Test blueprint (Exhibit A1)	Blueprint shows the distribution of items across assessment anchors with target depth noted.	The items on the local assessment represent the content standards to the same breadth and depth as the Keystone items.
Instructions for item writers (Exhibit A3)	The instructions clearly show that item writers were to develop items of a similar or greater breadth and depth as the Keystones while maintaining a similar balance of representation.	
Instructions for item reviewers (Exhibit A4)	The item reviewers were asked to independently rate the assessment anchor and depth of knowledge (DOK) assessed by each item. These ratings were then compared to the original targets to be sure the final item pool contained items of similar or greater depth and breadth and at a similar ratio as the Keystones.	
External alignment study – see the technical report with final results (Exhibit A5)	The two-way alignment study completed by an independent contractor confirms that the local assessment is fully aligned with the eligible content and that the balance of representation is similar to that of the Keystones.	

Fairness

Data or Evidence	Explanation of how it supports the claim	Claim
Policy of bias-free testing (Exhibit F1) Report from bias & sensitivity review committee (Exhibit F6) Statistical analyses (Exhibit F8;	Our policy is to develop and administer tests free from bias. We have attached our formal policy. Our item writers are instructed on ways in which items could be slanted toward (or against) a particular student group and	Items were developed to be free of bias against any student group.

F9)	given tips on writing items that are fair to all. During item review, a bias and sensitivity committee convened; it was composed of diverse educators (i.e., black, Hispanic, and white; those who work with low-income students, students with disabilities, and English language learners). We then ran statistical analyses to see if the difficulty of items varied by more than 0.20 between any two student groups. Those that did were further analyzed by the bias and sensitivity committee. We also examined distractor choices of those items to see if one distractor was chosen more often by a particular group of students.	
Test administration and security protocols (Exhibit F4) Test administration and security monitoring plans (Exhibit F10)	We follow PDE's guidance on when alternative pathways to graduation can be used. Only 0.8 percent of our students have such significant cognitive disabilities that they cannot take our end-of-course assessments. We use PDE's accommodations manual and follow PDE policy for assessing students with disabilities. We monitor accommodations and test administration and security protocols by visiting about 30 percent of our schools unannounced on testing days.	Test administration procedures ensure all students experience an equitable testing environment.
Score reports (Exhibit F5) User Guides (Exhibit F11) Letter to parents (Exhibit F12)	Our student-level score report provides information about performance level, areas in need of improvement, and the score that will be averaged into the final grade. A separate score is given for the performance task and linked back to the rubric. Our school and district reports clearly provide information on average scores and percent in each performance level by student groups (white, black, Hispanic, Asian and other, Students with disabilities, English language learners) Our user's guide explains how to interpret each portion of the student report and which score will be incorporated into the student's grade. We send home a letter to the parents each year with the score report explaining the grading policy, graduation policy, and appeals policy.	Reports provide applicable data and information for all student groups.

Establishment of Proficiency Levels

Data or Evidence	Explanation of how it supports the claim	Claim
Performance Level Descriptors (PLD) and summary record from district board meeting where Keystone PLDs were adopted verbatim. (Exhibit P1)	We adopted the Keystone levels and PLDs as they were with no changes.	The description for Proficient was adopted directly from Keystones; other levels and descriptions were added appropriately.
Standard setting technical report (Exhibit P2)	The standard-setting technical report shows that we followed a standard implementation of the Modified Angoff combined with an extended Angoff method that best combined information from our performance tasks with the multiple-choice/short-answer portion of the assessment. We recruited 12 panelists from across the district who represented both the rural and small town areas in our district, as well as representing minority students. We also included nine teachers—two of whom had experience working with students with disabilities and three from a neighboring district—one district curriculum supervisor, one parent who works in the science field, and one high school principal. The standard setting employed several rounds with data collected and analyzed after each round. Data show strong convergence at the end and the variance across panelists is shown. Panelist evaluation forms show a strong understanding of the process and an endorsement of the final cut scores.	The process for establishing the Proficient cut score was appropriate for the test and implemented effectively.

Consistency

Data or Evidence	Explanation of how it supports the claim	Claim
Description of scoring procedures (Exhibit C1)	Scoring was done over a one-week period by a committee of teachers meeting in a central location. We have attached a full report of scoring procedures. It includes an explanation of the rubric, recruitment process for scorers, demographics of scorers, training, qualification, and monitoring procedures. In addition, we included the data showing the qualification results of the scorers, the inter-rater reliability on each performance task (both percent of exact agreement and percent within one score point), and the re-calibration exercise done at the start of each day.	Scoring procedures are implemented and monitored to ensure reliability of scores.
Document with reliability calculations (Exhibit C4)	We used a split-half method of calculating the internal reliability of the first test form of the traditional section only (not the performance task) and came up with a reliability calculation of 0.77. We also performed a KR-21 calculation on the MC section of the test and calculated a reliability of 0.82. We used the inter-rater reliability to evaluate the performance task.	Student scores are reliable indicators of achievement.
Equating manual (Exhibit A2; C2)	We use a judgmental process during item development to match item difficulty. Then we field test the new items to see how close in difficulty they are to the old items. We also keep half the test the same each year so we can determine if any score changes are due to changes in student knowledge or changes in the items. If the latter, we adjust the scores statistically.	Procedures are in place to ensure that the tests are equivalent in difficulty from one year to the next (or one form to the next).

Template B: Multiple Components Supplant the Keystone Exam

Alignment

Data or Evidence	Explanation of how it supports the claim	Claim
		The items on the local assessment represent the content standards to the same breadth and depth as the Keystone items.

Fairness

Data or Evidence	Explanation of how it supports the claim	Claim
		Items were developed to be free of bias against any student group.
		Test administration procedures ensure all students experience an equitable testing environment.
		Reports provide applicable data and information for all student groups.

Establishment of Proficiency Levels

Data or Evidence	Explanation of how it supports the claim	Claim
		The description for Proficient was adopted directly from Keystones, and other levels and descriptions were added appropriately.
		The process for establishing the Proficient cut score was appropriate for the test and implemented effectively.

Consistency

Data or Evidence	Explanation of how it supports the claim	Claim
		Scoring procedures are implemented and monitored to ensure reliability of scores.
		Procedures are in place to ensure that the tests are equivalent in difficulty from one year to the next (or one form to the next).

Suggested Evidence and Instructions for the Completion of Template B

The following tables are populated with ideas for evidence that can be attached, as well as factors to consider when writing the narrative for the middle column. Evidence is ordered appropriately from basic evidence that is required to extra documentation that will move a district into the "superior" column. Note: The items in **BOLD** are required.

Alignment

Data or Evidence	Explanation of how it supports the Claim	Claim
Include specific evidence that shows that, when combined, the sum of the parts of the assessment components is fully aligned to the course content standards. This evidence could include the following:: • ***Test blueprint or specifications for each component*** • ***Explanation of how components are combined*** • *Item specifications* • ***Written instructions for item writers*** • ***Written instructions for item reviewers*** • *Sample tasks of high DOK* • ***Alignment study done by district or school staff*** ○ ***Technical report explaining process and results*** ○ ***Matrix of items to course content standards*** • *External alignment study* ○ *Technical report explaining process and results* • *Research studies examining the constructs tested by each item, such as a cognitive lab or think-aloud studies*	*Describe in words how the evidence submitted shows that the local assessment matches the course content standards (assessment anchors and eligible content) and/or the test blueprints for the Keystones. Discuss matching both the breadth and depth of knowledge of the target course content standards. Provide evidence to confirm that through the combination of the components, EVERY content standard has a sufficient balance of representation and tests a range of depth through DOK, Level 4.* *If using a unique testing approach, showing the content measured might require additional evidence, such as an external alignment study or research evidence of student knowledge tapped by the assessment.*	*The items on the local assessment represent the content standards to the same breadth and depth as the Keystone items.*

Fairness

Data or Evidence	Explanation of how it supports the claim	Claim
Include as many as possible: • ***Description of policies and***	*Describe how the item development process included multiple steps to ensure that they were free of bias (e.g., train*	*Items were developed to be free of bias against any student group.*

procedures used to ensure test items are not biased against any student group • *Evidence of a bias review committee meeting* ○ *Notes from meeting* ○ *Steps taken as a result of the meeting* • *Evidence of universal design for learning (UDL) procedures in item development* ○ *Instructions for writers* • *Statistical analyses of item difficulty across various student groups* • *Statistical analysis of distractor choice across student groups*	*item writers on universal design techniques; include a bias and sensitivity review committee comprised of members of multiple racial/ethnic groups and representatives of students with disabilities and English language learners in the review process; and/or conduct statistical analyses of items by student group). At least one procedure should be undertaken and documented. The more that is done, the higher the score will be.*	
• ***Written criteria for taking local assessments versus alternative pathways to graduation*** • ***Evidence that local accommodations policy follows PDE policy*** • ***Test administration and security protocols*** • *Test administration and security monitoring plans*	*First, show that students are treated equitably during test administration. Note that equitable is not necessarily equal. For instance, some accommodations may help some students access the test but not others. Any deviations from standard PDE accommodations policy should be approved by PDE prior to submitting the validity evidence, and the approval from PDE should be included in the packet.* *Also, provide evidence that IEP teams are given similar guidelines for graduation pathways in your district as in the rest of Pennsylvania or provide a rationale for the difference.* *Demonstrate that students across the district follow the same administration and security protocols to avoid giving certain students an undue opportunity to cheat. Describe procedures for monitoring test administration and security protocols to ensure that all students have equal access to the test and none are unduly influenced.*	*Test administration procedures ensure all students experience an equitable testing environment.*
• ***Sample score reports*** • *User guides to score reports relating scores to graduation requirements* • *Reports communicating*	*Show how the reports provide ample information to all stakeholders. If there is a large population of Spanish speaking students, document what information will be translated. Describe how expectations*	*Reports provide applicable data and information for all student groups.*

expectations to students, teacher,s and parents • *Policy regarding appeals process for disputed student scores*	*and results will be communicated to students, parents and teachers.*	

Establishment of Proficiency Levels

Data or Evidence	Explanation of how it supports the claim	Claim
PLDs and evidence that they were adopted by your local district. *Description of how they were adopted (who decided whether the Keystone descriptors should be adopted as is; if new descriptors were developed, how were they developed and by whom?)*	*Show that the description for Proficient was adopted directly from Keystones and that other levels and descriptions were added appropriately. Descriptions of Advanced or Basic could be more rigorous than the Keystones. If they are less rigorous, explain how they will be used.*	*The local assessment system maintains an adequate level of rigor.*
Standard-setting technical report that includes: • ***Name of method used to set cut scores*** • ***Rationale for selecting that method*** • ***Composition of standard-setting panel*** • ***Evidence that process followed a research-based and documented application of that method*** • ***Evidence of internal consistency of panelists (e.g., convergence in cut scores over rounds)*** • ***Discussion of how scores from each component are combined to determine an overall proficiency level.*** • *Results of an evaluation of panelists about their comfort level with the process and results*	*A complete standard-setting technical report should net you at least a satisfactory score on this dimension. Follow the model provided in the PDE Technical Advisory Committee document appended to the validity handbook. Show that a documented, validated method was used and provide evidence that it was implemented effectively. Document the range of panelists (representing different student groups and geographical areas in the district) and the rationale for any decisions made about the procedures.* *The one piece that is unique for the option of using multiple components is that you must show how the scores across the different components are combined into a single determination of a proficiency level.*	*The process for establishing the Proficient cut score was appropriate for the test and implemented effectively.*

Consistency

Data or Evidence	Explanation of how it supports the claim	Claim
Description of scoring procedures, including: • ***Rubric*** • ***Scoring process*** • ***Number and qualification of scorers*** • ***Training of scorers*** • ***Monitoring of scorers*** *Data showing inter-rater agreement on scoring of open-ended items*	*If the test is all selected response (multiple choice), document the scanning procedures and the quality control procedures to ensure the key is correct.* *If the test includes open-ended items, it is important to show that they are scored reliably. Document the review of the rubric, selection of scorers, training of scorers, tests of the scorers' accuracy and reliability, and continual monitoring of the scores. The goal is to show that the scores themselves are calculated accurately.*	*Scoring procedures are implemented and monitored to ensure reliability of scores.*
Calculations of internal consistency of multiple-choice items. *Other reliability statistics*	*There are several statistical methods that can be used to support internal consistency of the assessment (e.g., Pearson r, KR-21, split-half, Chronbach alpha). Select one and provide the calculation. Also, consider calculating the item-to-test correlation to show each item's contribution to the total score.*	*Student scores are reliable indicators of achievement.*
Description of procedures used to ensure comparable difficulty of items/forms over time • ***Judgmental*** • ***Statistical*** • ***Formal equating***	*Explain how you will refresh your item pool while maintaining comparability over time.* • *Will you use human judgment to determine a replacement item is of the same difficulty as the old item?* • *A statistical procedure?* • *How will you ensure the scorers score open-ended items with the same rigor from one year to the next?*	*Procedures are in place to ensure that the tests are equivalent in difficulty from one year to the next (or one form to the next).*

One Example of a Submission under Template B

The following provides a simple example of a submission that would be satisfactory, assuming the attachments provided information that fully met the evaluation criteria.

Introduction:

We created a local assessment for our students in English composition because we feel that composition needs to be measured over time and not during a single on-demand assessment. We are a small district with only one high school; thus, the teachers work together to develop, administer, and score the assessment. Our assessment contains three main components:

1. The teachers develop a debate topic each year. The teachers then work with students in our oral communication class to develop the arguments and set up a debate. Students from our composition class are required to attend the day of the debate. They must listen to both sides and write a three to five page persuasive essay briefly summarizing the two sides of the debate in which they explain which side they felt gave the better argument and why. They are given one hour to write this paper. The paper is scored by both the English composition and the oral communication teachers and returned to the students the next day to revise. The revision is also scored by both teachers.
2. The second component is a research assignment. Students must choose one of five topics to research. English teachers work with the social studies department to select topics related to history or economics. Students must research the topic and write a 10 to 12 page expository essay about the topic. They are given two weeks to complete the assignment. The first draft is scored by both the English composition and the social studies teachers and given back to the students to revise. The revision is also scored by two teachers.
3. The third component is an on-demand assessment, testing their knowledge of grammar and writing conventions. Students who choose this subject as one of their graduation requirements must score Proficient or above to graduate.

Alignment

Data or Evidence	Explanation of how it supports the Claim	Claim
Test Blueprint (Exhibit A1)	The blueprint shows that each component focuses on a specific assessment anchor. The first component focuses on "writing to persuade--persuasion" and "revisions--persuasion" and covers all eligible content under those anchors. The second component covers "writing to inform--exposition" and "revisions--exposition" and covers all the eligible content. The third component covers the eligible content under both "editing for conventions--exposition" and "editing for conventions--persuasion."	The items on the local assessment represent the content standards to the same breadth and depth as the Keystone items.
Instructions for item writers (Exhibit A3)	The instructions developed at the district level convey to teachers the need to develop topics aligned with the course content standards. Likewise, the scoring rubrics had to align with the eligible content. The instructions for writing items for the on-demand assessment component clearly show that teachers had to develop items similar to the breadth and depth of the Keystone assessment component that measured Editing for Conventions.	
Internal alignment study (Exhibit A5)	The two-way alignment study was completed by all teachers of English language arts in our high school as well as our district assessment coordinator. We showed that all eligible content is assessed by at least	

	one of our three components at an equal or greater depth than that of the Keystones.	

Fairness

Data or Evidence	Explanation of how it supports the Claim	Claim
Policy of bias-free testing (Exhibit F1) Statistical analyses (Exhibit F8)	Our policy (attached) is to develop and administer tests free from bias. Our teachers have reviewed research papers that describe ways in which items could be slanted toward (or against) a particular student group and have been given tips on providing prompts and writing items that are fair to all. Our district is 20 percent black and 80 percent white. Approximately 7 percent of our students have an IEP. We focused our analysis on these two characteristics. For the first two components, we checked total scores to see if there was any evidence of bias in scoring but found none. For the third component, we ran statistical analyses to see if the difficulty of items varied by more than 0.20 between any two student groups. Those that did were further analyzed for bias.	Items were developed to be free of bias against any student group.
Test administration and security protocols (Exhibit F4)	We have policies to dissuade students from cheating and to monitor the essays for evidence of cheating. We follow PDE's guidance as to when alternative pathways to graduation can be used. None of our students have such significant cognitive disabilities that they cannot take our end-of-course assessments. We use PDE's accommodations manual and follow PDE policy for assessing students with disabilities.	Test administration procedures ensure all students experience an equitable testing environment.
Score reports (Exhibit F5) Letter to parents (Exhibit F12)	Our student-level score report shows all five scores that were calculated, the proficiency level for each component, and the overall proficiency level. Our school and district report provide aggregate information on average scores and percentages in each performance level by our primary student groups (white/black; students with disabilities/ students without disabilities) We send home a letter to the parents each year with the score report explaining the grading policy, testing policy, graduation policy, and appeals policy.	Reports provide applicable data and information for all student groups.

Establishment of Proficiency Levels

Data or Evidence	Explanation of how it supports the claim	Claim
PLDs and summary record from district board meeting where	We adopted the Keystone levels and PLDs as they were with no changes.	The description for Proficient was adopted directly from the Keystones and other levels

Keystone PLDs were adopted verbatim. (Exhibit P1)		and descriptions were added appropriately.
Standard setting technical report (Exhibit P2)	Our assessment results in five scores: • first draft of the persuasive essay • revised draft of the persuasive essay • first draft of the expository essay • revised draft of the expository essay • score on the on-demand edit for conventions test. We used a Body of Work process to determine a total proficiency level for the persuasive essay and again for the expository essay. Then, we used a Modified Angoff approach to determine the proficiency level for the on-demand section. Next, we developed a table that combined the proficiency levels for each component into one overall proficiency level. For instance, a score of Advanced on the persuasive essay, Proficient on the expository essay, and Basic on the editing for conventions portion resulted in a total level of Proficient for that student. Thus, a student is assigned three component proficiency levels and one overall level. The standard setting technical report shows how we arrived at each cut score for Basic, Proficient and Advanced. We had nine panelists from across the district, including three English teachers, one history teacher, one community college professor, one district curriculum supervisor, one parent who works as a writer, one high school principal, and a local journalist. The standard setting for each component was done in two rounds. We reached consensus on all portions of the standard setting. Panelist evaluation forms show a strong understanding of the process and an endorsement of the final decisions.	The process for establishing the Proficient cut score was appropriate for the test and implemented effectively.

Consistency

Data or Evidence	Explanation of how it supports the claim	Claim
Description of scoring procedures (Exhibit C1)	Two teachers scored every paper. One teacher was the student's classroom teacher while the other was either the oral communication teacher (for the persuasive piece) or a social studies teacher (either a history or economics teacher depending on the topic chosen by the student). The papers were scored separately and an average of the two scores was used for the final score. Each paper was scored out of 100 points. If the teachers differed by more than 10 points, they (along with the principal) met to go over the scores and determine a final score. We have attached a full report of scoring procedures. It includes an explanation of the rubric, the distribution of scores for	Scoring procedures are implemented and monitored to ensure reliability of scores.

	each teacher, and a report noting the number of discrepant scores and how the conflicts were resolved. We also calculated inter-rater reliability on each essay, both the original and revised. The third component was all multiple-choice and was only scored once.	
Document with reliability calculations (Exhibit C2)	We used a Pearson r correlation to determine the reliability of the first two components. The first component had a reliability of r=0.76 and the second component was 0.67. We also calculated the Chronbach alpha of the third component, which was 0.83.	Student scores are reliable indicators of achievement
Description of process for developing new items and prompts (Exhibit C2)	We use a judgmental process for the first two components to maintain a comparable difficulty level across years. When teachers develop the debate topic and choices for the expository writing topic, they discuss the difficulty level of the topic and attempt to maintain an equivalent level of difficulty. For the third component, we embed several new items each year to see how close in difficulty they are to the old items. We also keep 67 percent of the test the same each year so we can determine if any score changes are due to changes in student knowledge or changes in the items. If the latter, we adjust the scores statistically.	Procedures are in place to ensure that the tests are equivalent in difficulty from one year to the next (or one form to the next).

Template C: One Component to Supplement the Keystone Exam

Alignment

Data or Evidence	Explanation of how it supports the claim	Claim
		The supplemental component is aligned to at least one assessment anchor.

Fairness

Data or Evidence	Explanation of how it supports the claim	Claim
		Items were developed to be free of bias against any student group.
		Test administration procedures ensure all students experience an equitable testing environment.
		Reports provide applicable data and information for all student groups.

Establishment of Proficiency Levels

Data or Evidence	Explanation of how it supports the claim	Claim
		The description for Proficient was adopted directly from Keystones and other levels and descriptions were added appropriately.
		The process for establishing the Proficient cut score was appropriate for the test and implemented effectively.

Consistency

Data or Evidence	Explanation of how it supports the claim	Claim
		Scoring procedures are implemented and monitored to ensure reliability of scores.
		Procedures are in place to ensure that the tests are equivalent in difficulty from one year to the next (or one form to the next).

Suggested Evidence and Instructions for the Completion of Template C

The following tables are populated with ideas for evidence that can be attached, as well as factors to consider when writing the narrative for the middle column. Evidence is ordered appropriately from basic evidence that is required to extra documentation that will move a district into the "superior" column. Note: The items in **BOLD** are required.

Alignment

Data or Evidence	Explanation of how it supports the claim	Claim
Include specific evidence that shows that, when combined, the sum of the parts of the assessment components is fully aligned to the course content standards. This evidence could include: • ***Test blueprint or specifications for each component*** • ***Explanation of how components are combined*** • *Item specifications* • ***Written instructions for item writers*** • *Written instructions for item reviewers* • *Sample tasks of high DOK* • ***Alignment study done by district or school staff*** ○ ***Technical report explaining process and results*** ○ ***Matrix of items to course content standards*** • *(or) External alignment study* ○ *Technical report explaining process and results* • *Research studies examining the constructs tested by each item, such as a cognitive lab or think-aloud studies*	*Describe in words how the evidence submitted shows that the local assessment matches the course content standards (assessment anchors and eligible content). The supplemental component does not need to align to every standard, but it does need to align to some and should measure at least one standard at a greater depth than the Keystones. Be sure to discuss matching both the breadth and depth of knowledge of the target course content standards.* *If using a unique testing approach, showing the content measured might require additional evidence such as an external alignment study or research evidence of student knowledge tapped by the assessment.*	*The supplemental component is aligned to at least one assessment anchor.*

Fairness

Data or Evidence	Explanation of how it supports the claim	Claim
Include as many as possible: • ***Description of policies and procedures used to ensure test items***	*Describe how the item development process included multiple steps to ensure that they were free of bias (e.g., train item writers on universal*	*Items were developed to be free of bias against any student group.*

are not biased against any student group • *Evidence of a bias review committee meeting* ○ *Notes from meeting* ○ *Steps taken as a result of the meeting* • *Evidence of universal design for learning (UDL) procedures in item development* ○ *Instructions for writers* • *Statistical analyses of item difficulty across various student groups* • *Statistical analysis of distractor choice across student groups*	*design techniques; include a bias and sensitivity review committee comprised of members of multiple racial/ethnic groups and representatives of students with disabilities and English language learners in the review process; and/or conduct statistical analyses of items by student group). At least one procedure should be undertaken and documented. The more that is done, the higher the score will be.*	
• ***Written criteria for taking local assessments versus alternative pathways to graduation*** • ***Evidence that local accommodations policy follows PDE policy*** • ***Test administration and security protocols*** • *Test administration and security monitoring plans*	*First, show that students are treated equitably during test administration. Note that equitable is not necessarily equal. For instance, some accommodations may help some students access the test but not others. Any deviations from standard PDE accommodations policy should be approved by PDE prior to submitting the validity evidence, and the approval from PDE should be included in the packet.* *Also, provide evidence that IEP teams are given similar guidelines for graduation pathways in your district as in the rest of Pennsylvania or provide a rationale for the difference.* *Demonstrate that students across the district follow the same administration and security protocols to avoid giving certain students an undue opportunity to cheat. Describe procedures for monitoring test administration and security protocols to ensure that all students have equal access to the test and none are unduly influenced.*	*Test administration procedures ensure all students experience an equitable testing environment.*
• ***Sample score reports*** • *User guides to score reports relating scores to graduation requirements*	*Show how the reports provide ample information to all stakeholders. If there is a large*	*Reports provide applicable data and information for all student groups.*

• *Reports communicating expectations to students, teachers and parents* • *Policy regarding appeals process for disputed student scores*	*population of Spanish speaking students, document what information will be translated. Describe how expectations and results will be communicated to students, parents and teachers.*	

Establishment of Proficiency Levels

Data or Evidence	Explanation of how it supports the claim	Claim
PLDs and evidence that they were adopted by your local district. *Description of how they were adopted (who decided whether the Keystone descriptors should be adopted as is; if new descriptors were developed, how were they developed and by whom?)*	*Show that the description for Proficient was adopted directly from the Keystones and that other levels and descriptions were added appropriately. The description of Advanced or Basic could be more rigorous than the Keystones. If the definitions are less rigorous, explain how they will be used.*	*The local assessment system maintains an adequate level of rigor.*
Standard-setting technical report that includes: • ***Name of method used to set cut scores*** • ***Rationale for selecting that method*** • ***Composition of standard-setting panel*** • ***Evidence that the process followed a research-based and documented application of that method*** • ***Evidence of internal consistency of panelists (e.g., convergence in cut scores over rounds)*** • ***Discussion of how scores from each component are combined to determine an overall proficiency level.*** • *Results of an evaluation of panelists about their comfort level with the process and results*	*A complete standard-setting technical report should minimally net a satisfactory score on this dimension. Follow the model provided in the PDE Technical Advisory Committee (TAC) document appended to the validity handbook. Show that a documented, validated method was used and provide evidence that it was implemented effectively. Document the range of panelists (representing different student groups and geographical areas in the district) and the rationale for any decisions made about the procedures.* *The one unique piece for the option of using multiple components is the need to explain how the scores across the different components are combined into a single determination of a proficiency level.*	*The process for establishing the Proficient cut score was appropriate for the test and implemented effectively.*

Consistency

Data or Evidence	Explanation of how it supports the claim	Claim
Description of scoring procedures, including: • ***Rubric*** • ***Scoring process*** • ***Number and qualification of scorers*** • ***Training of scorers*** • ***Monitoring of scorers*** *Data showing inter-rater agreement on scoring of open-ended items*	*If the supplemental component is all selected response (multiple choice), document the scanning procedures and the quality control procedures used to ensure the key is correct.* *If the component includes open-ended items, it is important to show that they are scored reliably. Document the review of the rubric, selection of scorers, training of scorers, tests of the scorers' accuracy and reliability, and continual monitoring of the scores. The goal is to show that the scores themselves are calculated accurately.*	*Scoring procedures are implemented and monitored to ensure reliability of scores.*
Calculations of internal consistency of multiple-choice items. *Other reliability statistics*	*There are several statistical methods to demonstrate internal consistency of the assessment (e.g., Pearson r, KR-21, split-half, Chronbach alpha). Select one and provide the calculation. Also, consider correlating the supplemental component to Keystone score to show how consistent or different the supplemental component is.*	*Student scores are reliable indicators of achievement.*
Description of procedures used to ensure comparable difficulty of items/forms over time • ***Judgmental*** • ***Statistical*** • ***Formal equating***	*Describe how the supplemental component will be kept fresh while maintaining comparability over time. Describe the process to ensure the scorers score open-ended items with the same rigor from one year to the next.*	*Procedures are in place to ensure that the tests are equivalent in difficulty from one year to the next (or one form to the next).*

One Example of a Submission under Template C

The following provides a simple example of a submission that would be satisfactory, assuming the attachments provided information that fully met the evaluation criteria.

Introduction:

We wanted to supplement the Keystone assessment in English composition with an in-depth writing component because we feel that composition needs to be measured over time and not during a single on-demand assessment. Therefore, our supplemental component is a research assignment. Students must choose one of five topics to research. English teachers work with the social studies department to select topics related to history or economics. Students must research the topic and write a 10 to 12 page expository essay about the topic. They are given two weeks to complete the assignment. The first draft is scored by both the English composition and the social studies teachers and given back to the students to revise. The revision is also scored by two teachers. A standard setting study determined how the initial and final scores were to be combined into a proficiency level. Then we developed a chart to show how this proficiency level would be combined with the Keystone level to determine whether or not the student met the proficiency standard.

Alignment

Data or Evidence	Explanation of how it supports the claim	Claim
Test Blueprint (Exhibit A1)	The blueprint shows that the supplemental component focuses on all three assessment anchors about "exposition."	The supplemental component is aligned to at least one assessment anchor.
Instructions for item writers (Exhibit A3)	The instructions developed at the district level convey to teachers the need to develop topics aligned with the course content standards, although that left a lot of latitude regarding the particular topics. Likewise, the scoring rubrics had to align with the eligible content.	
Internal alignment study (Exhibit A5)	A committee of English language arts teachers reviewed the prompts, scoring rubrics, and sample essays from our tryout session and compared them to the eligible content for English composition assessment anchors C.E.1.1, C.E.2.1, and C.E.3.1. The rubrics were fully aligned with the eligible content. Not every piece of eligible content was captured in the rubric as we assumed the Keystone exam will cover that missing eligible content.	

Fairness

Data or Evidence	Explanation of how it supports the claim	Claim
Policy of bias-free testing (Exhibit F1) Statistical analyses (Exhibit F8)	Our policy (attached) is to develop and administer tests free from bias. Prior to serving on the prompt development committee, teachers were asked to review research papers that describe ways in which items could be slanted toward (or against) a particular student group and given tips on providing prompts and writing items that are fair to all (See Exhibit A). We examined total score by student group to see if there	Items were developed to be free of bias against any student group.

	seemed to be consistent bias in performance. We also examined student scores by teacher to see if any bias by a teacher could be detected.	
Test administration and security protocols (Exhibit F4)	We have policies to dissuade students from cheating and to monitor the essays for evidence of cheating. We follow the PDE guidance regarding alternative pathways to graduation and when they can be used. Less than 0.5 percent of our students have such significant cognitive disabilities precluding them from taking end-of-course assessments. In these cases, IEP teams work to develop appropriate graduation requirements. We use the PDE accommodations manual and follow PDE policy for assessing students with disabilities.	Test administration procedures ensure all students experience an equitable testing environment.
Score reports (Exhibit F5) Letter to parents (Exhibit F12)	In addition to the Keystone report, we develop a supplemental page that shows the original and revised essay score, as well as the overall proficiency level. We then provide the final end-of-course score and proficiency level. Our school and district report provide aggregate information on average scores and percent in each performance level by our primary student groups. We send home a letter to the parents each year with the score report explaining the grading policy, testing policy, graduation policy, and appeals policy.	Reports provide applicable data and information for all student groups.

Establishment of Proficiency Levels

Data or Evidence	Explanation of how it supports the claim	Claim
PLDs and summary record from district board meeting where Keystone PLDs were adopted verbatim. (Exhibit P1)	We adopted the Keystone levels and PLDs as they were with no changes.	The description for Proficient was adopted directly from Keystones and other levels and descriptions were added appropriately.
Standard setting technical report (Exhibit P2) Proficiency Table (Exhibit P3)	The supplemental component resulted in two scores: initial draft and revision. We first used a dominant profile approach to determine the range of scores that we thought would be acceptable to show basic, proficient, and advanced performance. We then examined actual papers that fell in those score ranges to finalize the cut scores for basic, proficient, and advanced. We then created a table to combine the proficiency level from the Keystone with this supplemental essay to determine the final proficiency level. We recruited 12 English teachers from our four high schools, plus asked our district's English curriculum supervisor; additionally, a high school principal (former English teacher) was recruited to participate. This group was split into two committees who first	The process for establishing the Proficient cut score was appropriate for the test and implemented effectively.

	worked separately to determine a proficient and advanced cut score. The two results were compared and the two committees worked together to come to consensus on a final cut score. Full details are contained in our technical report.	

Consistency

Data or Evidence	Explanation of how it supports the claim	Claim
Description of scoring procedures (Exhibit C1)	Two teachers score every paper. One teacher is the student's classroom teacher while the other is either another English teacher or a social studies teacher familiar with the topic selected by the student. The papers are scored separately, and then the teachers meet to compare scores and agree upon a final score. If agreement cannot be reached, the principal is called in to arbitrate. We have attached a full report of scoring procedures. It includes an explanation of the rubric, the training provided to teachers, the distribution of scores for each teacher before meeting together and after, and the number of arbitration meetings required. We also calculated inter-rater reliability on each essay based on each teacher's original score.	Scoring procedures are implemented and monitored to ensure reliability of scores.
Document with reliability calculations (Exhibit C3)	We correlated the percent of exact agreement on the initial rating as 0.72 and adjacent agreement as 0.98.	Student scores are reliable indicators of achievement.
Description of process for developing new writing prompts (Exhibit A2; C2)	We use a judgmental process to maintain a comparable difficulty level in prompt options across years. When teachers develop the choices for the expository writing topic, they discuss the difficulty level of the topic and attempt to maintain an equivalent level of difficulty.	Procedures are in place to ensure that the tests are equivalent in difficulty from one year to the next (or one form to the next).

Part V: Evidence

This section lists the evidence that districts should produce when completing the template for their submission for a validity evaluation. This section has been organized by the main three segments of an assessment: test design, achievement standards and technical quality. Within each segment, the evidence has been divided into what is required and what is only recommended for further information, but not required. Submitting suggested evidence may raise the overall evaluation score and provide a more well-rounded synthesis of the local assessment system.

Evidence of Alignment

Required Evidence of Alignment

A1. Test blueprint/Specifications
A2. Item specifications
A3. Written instructions for item writers
A4. Written instructions for item reviewers
A5. Results from alignment study done by district or school staff or from external alignment study
A6. Explanation of how components are combined (required for local assessments that contain more than one component or that are to be combined with Keystone results)

Optional Evidence of Alignment

A7. Sample tasks of high DOK
A8. Research studies examining the constructs tested by each item, such as think-aloud studies

Evidence of Fairness

Required Evidence of Fairness

F1. Description of policies and procedures used to ensure tests are not biased against any student group
F2. Written criteria for taking local assessments versus alternative pathways to graduation
F3. Evidence that local accommodations policy follows PDE policy
F4. Test administration and security protocols
F5. Sample score reports

Optional Evidence of Fairness

F6. Evidence of a bias review committee meeting
F7. Evidence of universal design for learning (UDL) procedures in item development
F8. Statistical analyses of item difficulty across various student groups
F9. Statistical analysis of distractor choice across student groups
F10. Test administration and security monitoring plans
F11. User guides to score reports relating scores to graduation requirements
F12. Reports communicating expectations to students, teachers and parents
F13. Policy regarding appeals process for disputed student scores

Evidence of Proficiency Levels

Required Evidence of Proficiency Levels

P1. PLDs (If Keystone PLDs were not adopted, then a description of the development process is required)
P2. Standard-setting technical report
P3. Evidence for how multiple components of the assessment will be combined (if applicable)

Optional Evidence of Proficiency Levels

P4. Description of how PLDs were adopted (e.g., board minutes)
P5. Evaluation forms from teachers regarding proficiency levels

Evidence of Consistency

Required Evidence of Consistency

C1. Description of scoring procedures
C2. Description of procedures used to ensure comparable difficult of items and forms over time

Optional Evidence of Consistency

C3. Data showing inter-rater agreement on scoring of open-ended items
C4. Calculations of internal consistency on multiple-choice items
C5. Other reliability statistics

Exhibit A1: Test Blueprint

	Multiple-choice items		Open-ended items						
Literature-Fiction	**DOK 2**	**DOK 3**	**DOK 2**	**DOK 3**	**DOK 4**	**Item Total**	**% of Total Test**	**Point Total**	**% of Total Test**
L.F.1.1 *Use appropriate strategies to analyze an author's purpose and how it is achieved in literature*	2	0	0	1	1	4	**5.0%**	9	**8%**
L.F.1.2 *Use appropriate strategies to determine and clarify meaning of vocabulary in literature.*	3	1	0	0	0	4	**5.0%**	4	**3%**
L.F.1.3 *Use appropriate strategies to comprehend literature during the reading process.*	3	0	1	1	0	5	**6.3%**	8	**7%**
L.F. 2.1 *Use appropriate strategies to make and support interpretations of literature*	2	1	1	0	1	5	**6.3%**	9	**8%**
L.F.2.2 *Use appropriate strategies to compare, analyze, and evaluate literary forms*	2	1	0	1	1	5	**6.3%**	10	**8%**
L.f.2.3 *Use appropriate strategies to compare, analyze, evaluate, literary elements*	2	1	0	1	1	5	**6.3%**	10	**8%**
L.F.2.4 *Use appropriate strategies to interpret and analyze the universal significance of literary fiction.*	3	1	1	0	1	6	**7.5%**	10	**8%**
L.f.2.5 *Use appropriate strategies to identify and analyze literary devices and patterns in literary fiction.*	2	0	1	1	0	4	**5.0%**	7	**6%**
Fiction Totals	15	6	5	7	5	38	**47.5%**	67	**56%**
[Add non-fiction here]									
Reading Test Totals						80	**100%**	120	**100%**

Exhibit A2: Item Specifications

This section will be provided in the next update of this Handbook.

Exhibit A3: Instructions to Item Writers

Item Writer Training

Item writers were selected and trained for the content areas of mathematics, reading, science and writing. Qualified writers were college graduates with teaching experience and a demonstrated base of knowledge in the content area. Many of these writers were content assessment specialists and curriculum specialists. The writers were trained individually and had previous experience in writing multiple-choice and open-ended items. Prior to developing items for the PSSA and the Keystone Exams, the cadre of item writers was trained with regard to the following:

- Pennsylvania Academic Standards, Assessment Anchors and Eligible Content
- Webb's Four Levels of Cognitive Complexity: Recall, Basic Application of Skill/Concept, Strategic Thinking and Extended Thinking
- General Scoring Guidelines for Each Content Area
- Specific and General Guidelines for Item Writing
- Bias, Fairness and Sensitivity Guidelines
- Principles of Universal Design
- Item Quality Technical Style Guidelines
- Reference Information
- Sample Items

Alignment to the Assessment Anchors and Eligible Content, grade-level appropriateness (reading/interest level, etc.), depth of knowledge, cognitive level, item/task level of complexity, estimated difficulty level, relevancy of context, rationale for distractors, style, accuracy, and correct terminology were major considerations in the item development process. The Standards for Educational and Psychological Testing (AERA, APA, NCME, 1999) and the Principles of Universal Design (Thompson, Johnstone, & Thurlow, 2002) guided the development process. In addition, Data Recognition Corporation's Bias, Fairness, and Sensitivity Guidelines were used for developing items.

DRC's guidelines for bias, fairness, and sensitivity includes instruction concerning how to eliminate language, symbols, words, phrases, and content that might be considered offensive by members of racial, ethnic, gender, or other groups. Areas of bias that are specifically targeted include, but are not limited to, stereotyping, gender, regional/geographic, ethnic/cultural, socioeconomic/class, religious, experiential, and biases against a particular age group (ageism) or persons with disabilities. DRC catalogues topics that should be avoided and maintains balance in gender and ethnic emphasis within the pool of available items and passages.

As stated above, the Principles of Universal Design were incorporated throughout the item development process to allow participation of the widest possible range of students in the PSSA. The following checklist was used as a guideline:

- Items measure what they are intended to measure
- Items respect the diversity of the assessment population

- Items have a clear format for text
- Stimuli and items have clear pictures and graphics
- Items have concise and readable text
- Items allow changes to other formats, such as Braille, without changing meaning or difficulty
- The arrangement of the items on the test has an overall appearance that is clean and well organized

An important element in statewide assessment is the alignment between the overall assessment system and the state's standards. A methodology developed by Norman Webb (1999) offers a comprehensive model that can be applied to a wide variety of contexts. With regard to the alignment between standards statements and the assessment instruments, Webb's criteria include five categories, one of which deals with content. Within the content category is a useful set of levels for evaluating DOK. According to Webb (1999), "depth-of-knowledge consistency between standards and assessments indicates alignment if what is elicited from students on the assessment is as demanding cognitively as what students are expected to know and do as stated in the standards" (p. 7-8). The four levels of cognitive complexity (i.e., depth of knowledge) are as follows:

- Level 1: Recall
- Level 2: Application of Skill/Concept
- Level 3: Strategic Thinking
- Level 4: Extended Thinking

DOK levels were incorporated in the item writing and review process, and items were coded with respect to the level they represented. Generally, multiple-choice items are written to DOK levels 2 and 3, and open-ended items are written to DOK level 3.

Exhibit C1: Scoring Procedures

Our scoring procedure for the multiple-choice items simply involves running the Scantron form through the machine. The key is checked by three reviewers before it is entered. Any item that appears to have multiple incorrect answers all selecting the same distractor is examined by our district English language arts (ELA) coordinator to be sure the key is correct and that the alternate answer is truly incorrect.

Our scoring procedures for the essays include rangefinding, training of scorers, and quality insurance. All procedures are described here.

Rangefinding

After student answer documents were received, our lead ELA coordinator assembled groups of responses that exemplified the different score points represented in the mode-specific and conventions scoring guidelines.

Once examples for all the score points were identified, five ELA teachers from the district gathered for rangefinding. After an introductory general session, copies of the student example sets were presented to the teachers. The teachers reviewed and scored the student samples together to ensure that everyone was interpreting the scoring guidelines consistently. Teachers then went on to score responses independently and those scores were discussed until a consensus was reached. Only responses for which a good agreement rate was attained were used in training the final scorers. Discussions of the responses used the language of the scoring guidelines, assuring that the score point examples clearly illustrated the specific requirements of each score level.

Training

After rangefinding was completed, the district lead ELA coordinator compiled the scoring guidelines and responses that were relevant in terms of the key scoring concepts (e.g., elements of author's purpose or cause and effect) they illustrated were annotated for use in a scoring guide. The scoring guide for each mode served as the reader's constant reference. Scorers were instructed on how to apply the guidelines and were required to demonstrate a clear comprehension of each academic standard set by performing well on the training materials that were presented for each grade and mode. Training and qualifying sets consisted entirely of examples of student responses chosen by the rangefinding committee.

Scorer training began with the ELA coordinator providing an intensive review of the scoring guides and anchor papers to all readers. Next, the scorers "practiced" by independently scoring the responses in the training sets. Afterwards, there was a thorough discussion of each set of responses. Once the scoring guides and all the training sets were discussed, scorers were required to apply the scoring criteria by qualifying (i.e., scoring with acceptable agreement to the "true" scores) on at least one of the qualifying sets. Readers who failed to achieve a 75 percent level of exact agreement were given additional training to acquire the highest degree of accuracy possible. Readers who did not perform at the required level of agreement by the end of the qualifying process were not allowed to score "live" student work and were released from the project.

Hand-scoring Process

Student written responses were scored independently and by multiple readers. All essays were read twice to ensure reliability. Scorers were seated at tables where they read each response and keyed in the scores. Score mismatches were routed to the scoring director (ELA coordinator) for review and resolution.

Quality Control

Scorer accuracy was monitored throughout the scoring session, ensuring that an acceptable level of scoring accuracy was maintained. Inter-reader reliability was tracked and monitored with multiple quality control reports that were reviewed by quality assurance analysts. The following reports were used in scoring the essays:

- The "Reader Monitor Report" monitored how often readers were in exact agreement and ensured that an acceptable agreement rate was maintained. This report provided daily and cumulative exact and adjacent inter-reader agreement and the percentage of responses requiring resolution.
- The "Score Point Distribution" Report monitored the percentage of responses given each of the score points. For example, this daily and cumulative report showed how many 0s, 1s, 2s, 3s, and 4s a reader had given to all the responses he or she had scored at the time the report was produced. It also indicated the number of responses read by each reader so that production rates could be monitored.
- The" Item Status Report" monitored the progress of hand-scoring. This report tracked each response and indicated the status (e.g., "needs second reading" or "complete"). This report ensured that all discrepancies were resolved by the end of the project.
- The "Response Read by Reader Report" identified all responses scored by an individual reader. This report was useful if any responses needed rescoring because of reader drift.
- "Validity Reports" tracked how the readers performed by comparing predetermined scored responses to readers' scores for the same set of responses. If the readers fell outside of a determined percentage of agreement, remediation occurred and additional validity responses were given to individuals who needed to be monitored more closely.

Recalibration sets were used throughout the scoring sessions to monitor scoring by comparing each reader's scores with the true scores and to refocus scorers on the rubric. This check made sure there was no change in the scoring pattern as the project progressed. Scorers failing to achieve a certain percent of agreement with the recalibration true scores were given additional training to achieve the highest degree of accuracy possible.

Exhibit P3: Proficiency Table

This table shows how we combine the supplemental essay component with the Keystone exam. Note that the Keystone carries slightly more weight, so anytime a student is between levels, the final judgment is in favor of the Keystone judgment. However, a student may not reach proficiency with a score of Below Basic on either component. The table shows the Keystone proficiency levels across the top and the essay proficiency levels along the side. The levels within the table show the final proficiency levels for every possible combination of the two components.

	Keystone English Composition Score			
Essay Score	**Below Basic**	**Basic**	**Proficient**	**Advanced**
Below Basic	Below Basic	Basic	Basic	Basic
Basic	Below Basic	Basic	Proficient	Proficient
Proficient	Basic	Basic	Proficient	Advanced
Advanced	Basic	Proficient	Proficient	Advanced

Exhibit P4: Sample Board Minutes Adopting Proficiency Standards

INDEX TO MINUTES
STATE BOARD OF EDUCATION
April 7, 2010

MINUTES OF THE
STATE BOARD OF EDUCATION MEETING
Atlanta, Georgia April 7, 2010
Wanda Barrs, Chairman **Kathy Cox, Superintendent**

The State Board of Education met on Wednesday, April 7, 2010, at 8:30 a.m. in the State Board Room for an official one-day meeting, combining the Committee of the Whole with the regular State Board meeting.

Chair Barrs pointed out and it was agreed that based on the agenda for the month of April, it was determined that a one day meeting was a more efficient use of time.

Roll Call

Mrs. Wanda Barrs
Dr. Jim Bostic
Mr. Brad Bryant
Mr. Brian Burdette
Mr. Al Hodge Mr. Allen Rice
Dr. Mary Sue Murray Mr. Larry Winter
Mr. Jose Perez Mrs. Linda Zechmann
Dr. Elizabeth Ragsdale

Absent: Mr. Buzz Law
District three: vacant

The Board Committee meetings, Budget/Finance Committee, Charter Committee and Policy/Rules Committee, were followed by the meeting of the Committee of the Whole.

At this juncture, the Board went into recess.

Call To Order

The business portion of the State Board Meeting was called to order at 3:10 p.m.

By motion of Mr. Bryant, seconded by Dr. Ragsdale a unanimous affirmative vote was given to suspend State Board By-Law 5-1, for the purpose of deviating from the Annual State Board meeting calendar, for April 7 and 8.

Executive Session

At 3:15 p.m., by motion of Dr. Bostic, seconded by Mr. Burdette, a unanimous affirmative vote was given to go into Executive Session for the purpose of discussing waivers, appeals, legal and personnel matters.

At 4:05 p.m., the State Board reconvened. By motion of Dr. Ragsdale, seconded by Mr. Burdette, a unanimous affirmative vote was given to come out of Executive Session.

ACTION ITEMS

AGENDA

By motion of Dr. Bostic, seconded by Dr. Ragsdale, a unanimous affirmative vote was given to approve the agenda as amended for the April 7, 2010 State Board Meeting.

OTHER ITEMS

1. Assessment and Accountability – Criterion Referenced Competency Tests Standard Setting. By motion of Mrs. Zechmann, seconded by Dr. Bostic, a unanimous affirmative vote was given to approve the recommended Georgia Performance Standards (GPS)-based test performance standards or cut scores for the Criterion-Referenced Competency Tests (CRCT) in Social Studies for grades 6 and 7.

2. Assessment and Accountability – Georgia High School Graduation Test – Standard Setting. By motion of Dr. Ragsdale, seconded by Dr. Bostic, a unanimous affirmative vote was given to approve the recommended Georgia Performance Standards (GPS)-based test performance standards or cut scores for the Georgia High School Graduation Tests (GHSGT) in Social Studies.

3. Supplemental Educational Services (SES) Provider Subject Removal – FY10. By motion of Mrs. Zechmann, seconded by Dr. Bostic, a unanimous affirmative vote was given to authorize the State Superintendent of Schools to remove four (4) Supplemental Education Service (SES) providers from the 2009-2010 State-Approved Providers List for reading and/or language arts.

Adjournment

At 4:22 p.m., by motion of Dr. Ragsdale, seconded by Mr. Hodge, an affirmative vote was given to adjourn the State Board meeting.

Next Scheduled Meeting of the State Board

The next State Board Meeting is scheduled for May 13, 2010.

P4: Sample Evaluation Forms from Teachers Regarding Proficiency Level Setting

Final Evaluation of the Standard Setting Study

The purpose of this final evaluation form is to obtain your feedback about the cut score study. Your feedback will provide a basis for evaluating the training, methods and materials. Please complete the information below. **Do not put your name on the form.** We want your feedback to be anonymous.

Subject: 11 ELA (4 content specialists, 7 sped teachers)

Grade Level: 5 Elementary 3 Middle 3 High

Gender: 1 Male 10 Female

Race/ethnicity: 10 White 1 African American or Black

For items 1–6 below, please rate each statement using the scale given in the item. Place a check mark (√) in the appropriate box for each statement.

1. Please read each of the following statements carefully and indicate the degree to which you agree with each statement.

	Strongly Agree	Agree	Disagree	Strongly Disagree
I understood the purpose of this workshop.	11	☐	☐	☐
The training was adequate to give me the information I needed to complete my assignment.	10	1	☐	☐
I understood how to make the judgments.	9	2	☐	☐
I understood how to use the data provided.	9	2	☐	☐
I understood how the cut scores were calculated.	9	2	☐	☐

2. Please rate the clarity of the following instructions provided.

	Very Clear	Mostly Clear	Mostly Unclear	Very Unclear
Instructions provided in the material	9	2	☐	☐
Instructions provided by the facilitator	11	☐	☐	☐

3. How <u>useful</u> was each of the following in making your judgments?

	Very Useful	Somewhat Useful	Not At All Useful
Reviewing the performance level descriptors	8	2	1
Discussion of weights	10	1	☐
Consideration of rubric rules	10	1	☐
Discussions with other participants about their rules	11	☐	☐
Using the spreadsheet to calculate cut scores	10	1	☐
Reviewing the samples of student work	8	2	1
Discussions with other participants about the works samples	10	1	☐
Impact data (% of students in each proficiency level)	11	☐	☐

4. How influential was each of the following in making your judgments?

	Very Influential	Somewhat Influential	Not At All Influential
Proficiency Level Descriptors	9	1	1
Discussions of borderline performance	10	1	☐
The performance rubric	10	1	☐
The complexity rubric	11	☐	☐
My experiences with students	9	2	☐
My experiences with the content area	7	4	☐
Discussions with other participants	10	1	☐
Cut scores of other participants	7	4	☐
Samples of student work	8	2	1
Percent of test takers who will be in each proficiency level (impact data)	7	3	1

5. Were there any materials or procedures that became more (or less) influential during the course of the cut score study? If so, which ones? Why?

- ✓ *Reviewing student work made the process confusing because of our group's judgment not matching the given score. (Also because we saw falsification of dates in evidence.) [marked sample student work as somewhat useful/influential]*
- ✓ *PLDs became less [from panelist who indicated PLDs were not at all influential]*
- ✓ *Reviewing PLDs – we found that quick discussion and highlighting differences between Basic/Proficient and Proficient/Advanced was effective [from panelist who indicated PLDs were very influential]*
- ✓ *Samples of student work became confusing—some of the scores didn't match what I would have thought [panelist marked sample work as very useful/influential]*
- ✓ *The students' work samples became more influential because we gained a better understanding through analysis of those. Unfortunately, I decided to reduce the importance of these due to sample size and rating discrepancies/inaccuracies [panelists changed rating from very to somewhat useful/influential]*

6. How appropriate was the amount of time you were given to complete the different components of the cut score study?

	Too Much Time	About Right	Too Little Time
Training on how to set cut scores	☐	11	☐
Discussion of the performance level descriptors	☐	11	☐
Determining the appropriate weights for the dimensions	☐	11	☐
Using the rubric to set preliminary cut scores	☐	11	☐
Reviewing and discussing the sample evidence	☐	10	1
Reviewing the impact data	☐	11	☐

7. Do you have additional comments about this process or suggestions on how to improve the training and/or implementation of the cut score study?

- ✓ *I enjoyed learning about the MAAECF this week, and I look forward to working with my SCD teachers and SPED teachers at my school.*

- ✓ *<u>All</u> teachers administering the alternate assessment need to be trained. Teachers also need everything in writing.*

- ✓ *This was very informative training—very non-threatening. When a teacher completes an alternate assessment: Watch complexity of baseline assessment to final assessment. (Final assessment task should be more complex than baseline assessment task.)*

Appendix A: Glossary

Academic standards: Statements of the knowledge and skills that students are expected to learn. Provide consistent targets for students, teachers and districts. Also known as content standards.

Accommodation: Changes in the administration of an assessment (such as setting, scheduling, timing, presentation format, response mode, or others) to provide better access to the assessment in a manner that does not change the construct intended to be measured by the assessment or the meaning of the resulting scores.

Accountability: The systematic use of assessment data and other information to evaluate the effectiveness of a program, such as an education system, for the purpose of rewarding desired outcomes and sanctioning undesirable outcomes.

Alternate assessment: An instrument used in gathering information on the performance and progress of students whose disabilities preclude them from valid and reliable participation in the general state assessment. Alternate assessments may be developed to measure alternate achievement standards, modified achievement standards or grade-level achievement standards.

Assessment: Any systematic method of obtaining evidence to draw inferences about people or programs. Assessment may include both formal methods, such as large-scale state assessments, and less formal classroom-based procedures, such as quizzes, class projects, and teacher questioning.

Assessment anchor: Statements that clarify the standards assessed in Pennsylvania which are intended for use by educators to help prepare their students for the assessments. The assessment anchors target a specific band of standards, enabling a higher level of clarity and improve articulation between instruction and assessment.

Bias: In a statistical context, a systematic error in a test score. In discussing test fairness, bias may refer to construct under-representation or construct-irrelevant components of test scores that differentially affect the performance of certain groups of test takers.

Blueprint: Detailed documentation of the intended characteristics of a test including, but not limited to, the content and skills to be measured, the numbers and types of questions, the level of difficulty and other statistical characteristics, the timing, and the layout.

Classification Errors: (aka Type I/Type II errors). Errors made when the application of a cut score or other determinant results in “failing” a student, school or district when they should have passed (Type I error) or “passing” someone who should have failed (Type II error).

Cognition: Essentially, thinking or learning that encompasses how people develop knowledge, skills and other forms of competence in a subject matter domain.

Cognitive complexity: A description of the type of thinking a student would need to do in order to correctly answer an item complete a task. This includes the number of mental structures a student would have to use, how abstract the item structures were and how elaborately the structures interacted with each other.

Comparability: The degree to which similar inferences can be made from the outcomes of two or more assessments.

Content validity: Evidence regarding the extent to which a test provides an appropriate sampling of a content domain of interest—e.g., assessable portions of a state's Algebra I curriculum in terms of the knowledge, skills, objectives and processes sampled.

Construct: As applied to assessment, the complete set of knowledge, skills, abilities, or traits representing a particular domain of knowledge (such as American history, reading comprehension, study skills, writing ability, logical reasoning, honesty, intelligence, and so forth). Constructs are not observed directly, but are inferred from observations of examined performance on tasks thought to be representative of this hypothesized trait.

Construct validity: Evidence regarding the extent to which a test measures the theoretical construct or trait it is intended to measure. Such evidence can be demonstrated by examining the interrelationships of the scores (i.e., correlations) on one test with scores on other tests that are theorized to measure either the same traits, or unrelated traits, and determining if the results are in the expected direction (e.g., high correlations with same trait measures and low correlations with unrelated trait measures).

Content domain: The set of behaviors, knowledge, and skills to be measured by a test, represented in a detailed specification and often organized into categories by which items are classified.

Curriculum: The knowledge and skills in subject matter areas that teachers are supposed to teach and students are supposed to learn, including a scope or breadth of content in a given subject area and a sequence for learning.

Cut score: A point on a score scale at or above which test takers are classified in one way and below which they are classified in a different way. For example, if a cut score is set at 60, then people who score 60 and above may be classified as "passing" and people who score 59 and below classified as "failing."

Decision consistency: A measure of the reliability of the classification decision. Decision consistency estimates the extent to which, if an examinee were administered a test on two separate occasions, the same classification decision (whether pass or fail) would be made.

Depth-of-knowledge: Related to cognitive complexity, it is the degree of depth or performance complexity required to understand/perform academic content/process found in content standards or assessment items; a description of different ways students interact with content measured by how deeply students must understand the content in order to respond.

Difficulty: In assessment, the proportion of respondents answering the item correctly. Conceptually, it is based on underlying knowledge and cognitive processes required to answer an item correctly.

Differential Item Functioning (DIF): A statistical property of a test items in which different groups of test takers who have the same total test score have different performance on particular items.

Distractor: An incorrect option presented to an examinee in a multiple choice item.

Domain Sampling: The process of selecting test items to represent a specified universe of performance tied to a hypothesized construct.

Dynamic Evaluation: As used in the context of validity, dynamic evaluation refers to the notion that evaluative judgments will be updated as new information about the assessment system is presented. In other words, dynamic evaluation refers to the idea that the evaluation continues to move (or adjust) as new information is gathered.

Eligible content: Often known as the "assessment limits" this helps teachers identify how deeply they need to cover an assessment anchor and/or the range of the content they should teach to best prepare their students for the assessment. Not all of the eligible content is assessed, but it shows the range of knowledge from which the test is designed.

Equating: The strongest of several "linking" methods used to establish comparability between scores from multiple tests. Equated test scores should be considered exchangeable. Consequently, the criteria needed to refer to a linkage as 'equating' are strong and somewhat complex (equal construct and precision, equity and invariance). In practical terms, it is often stated that it should be a 'matter of indifference' to a student if he/she takes any of the equated tests.

Interpretative argument: A plan specifying the proposed interpretations and uses of test results by laying out the network of inferences and assumptions leading to the observed performances to the conclusions and decisions based on the performances.

Item difficulty: A measure of the proportion of students answering an item correctly. A p-value is an index calculated as the proportion (sometimes percent) of students in the group who answer an item correctly. P-values range from 0.0 to 1.0 on the proportion scale. Lower values correspond to more difficult items and higher values correspond to easier items. P-values are usually provided for multiple-choice items or other items worth one point. For open-ended items or items worth more than one point, difficulty on a p-value-like scale can be estimated by dividing the item mean score by the maximum number of points possible for the item.

Item format: The variety of test item structures or types that can be used to measure examinees' knowledge, skills, and abilities,; typically including multiple-choice or selected-response, open-ended or constructed-response, essay, or performance task.

Learning progression: Description of successively more sophisticated ways of reasoning within a content domain or the expected pattern of the development of a set of knowledge and skills in a particular subject matter domain.

Measurement error: The differences between observed scores and the theoretical true score; the amount of uncertainty in reporting scores; and the degree of inherent imprecision based on test content, administration, scoring, or examinee conditions within the measurement process that produce errors in the interpretation of student achievement.

Modification: Changes made in both instructional and assessment situations that are individualized to student needs. In the context of assessment, changes are made to the content, format, and/or administrative procedures of a test in order to accommodate test takers who are unable to take the original test under standard test conditions. Unlike *accommodations*, modifications may directly or indirectly compromise the validity of the content standard by changing the construct. Modifications include a much wider range of supports and instructional scaffolding than do accommodations but can be effectively used in combination with accommodations in instructional and assessment situations when individualized to the student's strengths and needs. Modifications are intended to allow for meaningful participation and enhanced learning.

Multiple-choice item: A type of item format that requires the test taker to select a response from a group of possible choices, one of which is the correct answer (or key) to the question posed.

Open-ended item: An open-ended (OE) item—also known as a constructed-response (CR) item—is an item format that requires examinees to create their own responses, which can be expressed in various forms, (e.g., written paragraph, created table/graph, formulated calculation, etc.). Such items are frequently scored using more than two score categories, that is, polytomously (e.g., 0, 1, 2, and 3).

Parallel forms: Two or more assessments that provide similar outcomes (true scores) of the construct being measured.

Portfolio (assessment): An assessment comprising the collection and analysis of examinee work samples, typically consisting of performance tasks gathered over a specific period of time; often used to assess special populations who have difficulty with standard paper-and-pencil assessments. Portfolios often require some form of student self-reflection and evaluation.

Proficiency level: A definition of a level of performance, including both a minimum cut score and a written description that distinguishes the level of performance from other defined levels. Also called a performance standard or an achievement standard.

Raw score: An unadjusted score usually determined by tallying the number of questions answered correctly, or by the sum of item scores (i.e., points). Raw scores typically have little or no meaning by themselves and require additional information—like the number of items on the test, the difficulty of the test items, norm-referenced information, or criterion-referenced information.

Reliability: The characteristic of test scores of being dependable, generally conceptualized as stability or consistency over both time and items. Reliability is the quantification of the amount of measurement error in a test.

Sampling error: The error associated with observations from a sample instead of the whole population, used to quantify the expected range within which the true population value might be located relative to the sample data.

Scale score: A mathematical transformation of a raw score developed through a process called scaling. Scaled scores are most useful when comparing test results over time. Several different methods of scaling exist, but each is intended to provide a continuous and meaningful score scale across different forms of a test.

Standard setting: An activity in which a procedure is applied systematically to gather and analyze human judgment for the purpose of deriving one of more cut scores for a test.

Standards-based individualized education plan: An IEP that specifically refers to instruction of the state's academic standards for the student's enrolled grade and focuses on aligning instruction of students with disabilities to the academic content that all students at that grade level should know and be able to do.

Standard deviation: A statistic that measures the degree of spread or dispersion of a set of scores. The value of this statistic is always greater than or equal to zero. The further the scores are away from each other in value, the greater the standard deviation. This statistic is calculated using the information about the deviations (distances) between each score and the distribution's mean. It is equivalent to the square root of the variance statistic. The standard deviation is a commonly used method of examining a distribution's variability since the standard deviation is expressed in the same units as the data.

Standard error of measurement (SEM): The extent to which test scores from the same person can be expected to vary because of differences in such factors as the specific questions in different forms of the test, or the leniency or rigor of different scorers. As an example, across replications of a measurement procedure, the true score will not differ by more than plus or minus one standard error from the observed score about 68 percent of the time (assuming normally distributed errors). The SEM is frequently used to obtain an idea of the consistency of a person's score in actual score units, or to set a confidence band around a score in terms of the error of measurement.

Student with disabilities (SWD): In the Individuals with Disabilities Act, a student with disabilities is defined as "a child evaluated in accordance with §§300.530-300.536 as having mental retardation, a hearing impairment including deafness, a speech or language impairment, a visual impairment including blindness, serious emotional disturbance (hereafter referred to as emotional disturbance), an orthopedic impairment, autism, traumatic brain injury, another health impairment, a specific learning disability, deaf-blindness, or multiple disabilities, and who, by reason thereof, needs special education and related services."

Technical advisory committee (TAC): A group of individuals, most often professionals in the field of testing, that are either appointed or selected to make recommendations for and to guide the technical development of a given testing program.

Test domain: The portion of all knowledge and skill in a subject matter area that is selected to be assessed because there is consensus that it represents what is important for teachers to teach and for students to learn.

Test specifications: A detailed description for a test that specifies the number or proportion of items that assess each content and process/skill area (aka a test blueprint). Test specifications also describe the types of questions to be included on the tests and the types used to assess specific aspects of the domain.

Theory of action: Originally drawn from sociology and organizational studies, theory of action is used in the education context to refer to higher level view of the interpretative argument. Essentially, it provides an overview of how the specific components of the testing/educational system are intended to work in concert to bring about the desired aims.

Universal design: The creation of products and environments meant to be usable by all people, to the greatest extent possible, without the need for adaptation or specialization.

Validity: The extent to which inferences and actions made on the basis of a set of scores are appropriate and justified by evidence. It is the most important aspect of the quality of a test. Validity is based on how the scores are used and interpreted rather than on the test itself.

Validity argument: An evaluation of the completeness and coherence of proposed interpretations and uses of test results, based on both empirical evidence and logic, as specified by the interpretative argument.

Validity evaluation: The full set of activities related to evaluating the proposed interpretations and uses of test results including the interpretative and validity arguments as well as the validity studies plan and the actual studies themselves.