**HANDOUT #6.1**

**Module 6.1: Alignment Reviews**

**Key Tasks**

* + Items/tasks matching a specific content standard based upon the narrative description of the standard and a professional understanding of the knowledge, skill, and/or concept being described.
  + Items/tasks reflecting the cognitive demand/higher-order thinking skill(s) articulated in the standards, with extended performance tasks typically focused on several, integrated content standards.
  + Item/task distributions representing the emphasis placed on the targeted content standards in terms of “density” and “instructional focus”, while encompassing the range of standards articulated on the test blueprint.
  + Item/task distributions consisting of sufficient opportunities for test-takers to demonstrate skills, knowledge, and concept mastery at the appropriate developmental range.

**Procedural Steps**

* Identify a team of teachers to conduct the alignment review (best accomplished by department or grade-level committees) with technical support from the district.
* Organize items/tasks, operational forms, test blueprint, and targeted content standards.
* Conduct panelist training on the alignment criteria and rating scheme. Use calibration techniques with a “training set” of materials prior to conducting the review.
* Evaluate the following areas:
  + **Content Match (CM) and Cognitive Demand/Depth of Knowledge (DoK)**
    - Read each item/task in terms of matching the standards both in terms of content reflection and cognitive demand/(DoK).
    - For SA, ECR, Extended Performance tasks, ensure that scoring rubrics are focused on specific content-based expectations.
    - After reviewing all items/tasks, including scoring rubrics, count the number of item/task points assigned to each targeted content standard.
  + **Content Pattern**
    - Determine if the items/tasks are sampling the complexity and extensiveness of the targeted content standards. If assessment’s range is too narrowly defined, refine blueprints and replace items/tasks to match the range of skills and knowledge implied within a targeted standards. Refine if patterns do not reflect those in the standards.
  + **Item/Task Sufficiency (ITS)**
    - Determine the number of item/task points per targeted content standard based upon the total available. Using the item/task distributions, determine whether the assessment has at least five (5) points for each targeted content standard. Identify any shortfalls in which too few points are assigned to a standard listed in the test specification table.
    - Ensure the items/tasks are developmentally appropriate for the test-takers. Further, ensure the passages and narrative text do not contain linguistic challenges associated with vocabulary usage and text complexity.
* Record findings and present to the larger group with recommendations for improvements (e.g., new items/tasks, design changes, item/task refinements, etc.).
* Document the group’s findings and prepare for refinement tasks.

**HANDOUT #6.1**

**Alignment Workflow**

**Alignment Results (Example)**

The alignment review committee evaluated assessments used in Visual-Performing Arts based upon targeted content standards, blueprints, specification tables, scoring rubrics, and operational forms. Panelists focused on four alignment criteria: content match (CM), cognitive demand/depth of knowledge (DoK), Content Pattern (CP), and item/task sufficiency (ITS). Each panelist reviewed independently the assessments and then shared their individual findings with the other two panelists. Table 1 (next page) is a partial summary of their consolidated findings.

**HANDOUT #6.1**

**Table 1.** Alignment Results (Example)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Area/Grade | CM | DoK | CP | ITS | Comments |
| Science K | 🗹 | 🗹 | 🗹 | 🗹 | No Findings |
| Science 1 | 🗹 | 🗹 | 🗹 | 🗹 | No Findings |
| Science 2 | 🗹 |  | 🗹 |  | Standard #6 ≠ DoK 2; Test Length ≠ 56 pts. |
| Science 3 | 🗹 |  | 🗹 |  | Standard #6 ≠ DoK 2; Test Length ≠ 58 pts. |
| Science 4 | 🗹 |  | 🗹 |  | Standard #6 ≠ DoK 2; Test Length ≠ 60 pts. |
|  | **Note**: Standard #6 requires tasks to be DoK 2 or higher. The task points do not reflect the specification tables in quantity or distribution sufficiency. | | | | |
| Science 5 | 🗹 | 🗹 | 🗹 | 🗹 | No Findings |
| Biology | 🗹 | 🗹 | 🗹 | 🗹 | No Findings |
| Physical Science | 🗹 | 🗹 | 🗹 | 🗹 | No Findings |
|  |  | | | | |
| Chemistry | 🗹 |  | 🗹 |  | Standard #6 ≠ DoK 3; Test Length ≠ 56 pts. |
| Life Science | 🗹 |  | 🗹 |  | Standard #6 ≠ DoK 3; Test Length ≠ 58 pts. |
| Environmental Science | 🗹 |  | 🗹 |  | Standard #6 ≠ DoK 3; Test Length ≠ 60 pts. |
|  | **Note**: Standard #6 requires tasks to be DoK 3 or higher. The task points do not reflect the specification tables in quantity or distribution sufficiency. | | | | |

**Note:** The example provide above is one of many approaches used to display alignment results.

**HANDOUT #6.2**

**Module 6.2: Refinement Procedures**

**Tasks**

The following provides tasks typically completed during the refinement step:

* Analyze results from the prior assessment to identify areas of improvement
* Consider item/task replacement or augmentation to address areas of concern
* Development at least 20% new items/tasks, or implement an item/task tryout approach
* Evaluate parallel forms (i.e., Form A and B) for comparability
* Review items with p-values greater than 90% (p>.90)
* Clarify any identified scoring guidelines needing further improvements
* Verify performance level “cut-scores” reflect challenging yet attainable expectations
* Develop exemplars and anchor papers based upon student responses
* Conduct professional development in assessment literacy

**Resources**

The following provides an initial list of resources used to support refinement tasks:

* District curricula, scope-and-sequences, supervisor staff
* Content experts (teachers), curriculum developers, university academics
* Assessment literacy content
* Refinement procedures, protocols, materials
* External vendor assessments, technical materials
* State/national content standards, content frameworks
* Third-party, technical consultants
* Training facilities, equipment (e.g., Promethean board, InFocus, etc.)
* Budgetary authority (e.g., Title II), fiscal allocations

**HANDOUT #6.2**

**Analytics**

The following provides a listing of some analytics used in the refinement step:

* + Scoring consistency of SA, ECR, and Extended Performance tasks
  + Item/task difficulty and discrimination
  + Omission rates and distractor analysis
  + Differential Item Function (DIF)
  + Overall score distribution
  + Internal consistency and correlations between items/tasks within domain

**Note**: Module 5 will provide a set of quantitative procedures to evaluate (post hoc) items/tasks, including operational forms. Further, this module will explain the used of the analytics listed above.

**Timeline**

Table 2 below outlines a timeline (including tasks) to conduct refinement activities within a typical assessment cycle using a 24-month approach.

**Table 2.** Refinement 24-Month Timeline (Example)

|  |  |
| --- | --- |
| Date | Task |
| February 2015 | Scope and Procedural Preparation via Webinar |
| March 2015 | Begin Alignment Review-   * Reading/English-language Arts/English I-IV * Mathematics/Algebra I, II, Geometry, etc. * Social Studies/Civics, Geography, U.S. History, World History, etc. |
| April 2015 | Preliminary Findings Review via Webinar |
| May 2015 | Final Report and Recommendations |
| July 2015 | Item/Task Writing Workshops; Assessment literacy (professional development; on-site) |
| August 2015 | Quality control newly developed items/tasks |
| September 2015 | Publish operational forms and specification tables |
| October 2015 | Conduct Process Improvement Review (“Lessons Learned”) via Webinar |
| February 2016 | Scope and Procedural Preparation via Webinar |
| March 2016 | Begin Alignment Review-   * Science/Physical Science, Biology, Chemistry, Physics, etc. * CTE (Selected Pathways) * Art & Music |
| April 2016 | Preliminary Findings Review via Webinar |
| May 2016 | Final Report and Recommendations |
| July 2016 | Item/Task Writing Workshops; Assessment literacy (professional development; on-site) |
| August 2016 | Quality control newly developed items/tasks |
| September 2016 | Publish operational forms and specification tables |

**HANDOUT #6.3**

**Assessment Quality Rubric**

**-Scored Example-**

**DIMENSION I: DESIGN**

| **Task** | **Descriptor** | **Technical Evidence** |
| --- | --- | --- |
| **I.A** | The assessment’s design is appropriate for the intended audience and reflects challenging material needed to develop higher-order thinking skills. The purpose of the performance measure is explicitly stated. | Specification and Blueprint document contains a “generic” purpose statement for each assessment. Grade-level expectations identified in the design document represent key and challenging concepts. |
| **I.B** | The assessment’s design has targeted content standards representing a range of knowledge and skills students are expected to know and demonstrate. | Specification tables and blueprints identify the selected content standards. The standards identified represent the breadth of the strand/domain. |
| **I.C** | Specification tables and blueprints articulate the number of items/tasks, item/task types, passage readability, and other information about the assessment. | Specification and Blueprint document contains several tables. Passage readability or linguistic “load” information was provided. |
| **I.D** | Items/tasks are rigorous (designed to measure a range of cognitive demands/higher-order thinking skills at developmentally appropriate levels) and of sufficient quantities to measure the depth and breadth of the targeted content standards. | Cognitive demand of Level 2 or higher represent over 70% of the overall available points on the assessment. The assessment has at least five points per standard. Items/tasks reflect grade-level expectations. |

**DIMENSION II: BUILD**

| **Task** | **Descriptor** | **Technical Evidence** |
| --- | --- | --- |
| **II.A** | Items/tasks and score keys were developed using standardized procedures, including scoring rubrics for human-scored, open-ended questions. The total time to administer the assessment is developmentally appropriate for the test-takers. | [See completed items and tasks in the operational form.] Scoring guidelines were provided (answer key) along with scoring rubrics. Administration guidelines were found on page 2 of the operational form. Total time for the assessment was 50 minutes (high school students). |
| **II.B** | Items/tasks were created in terms of: (a) match to the targeted content standards, (b) content accuracy, (c) developmental appropriateness, (d) cognitive demand, (e) bias, (f) sensitivity, and (g) fairness. | 100% of all items and tasks were reviewed prior to the development of the test form by an item development team at the district. All identified issues were corrected prior to administration. Administrative guidelines were embedded within the test document. Directions to test-takers were provided for each item type, with clear expectations on “how” the test-taker was to respond to each SA and ECR task. |
| **II.C** | Administrative guidelines contain step-by-step procedures used to administer the assessment in a consistent manner, including scripts to orally communicate directions to students, day and time constraints, and allowable accommodations or adaptations. | Administrative guidelines were embedded within the test document. Directions to test-takers were provided for each item type. Accommodations provided consistent with those authorized in the IEP and used in classroom instruction. |
| **II.D** | Scoring guidelines were developed for human-scored items/tasks to promote score consistency across items/tasks and among different scorers. These guidelines articulate point values for each item/task used to combine results into an overall raw score. | Scoring rubrics articulate the performance continuum in content-based terms. No behavioral dimensions were included. Overall point values were listed in the test-taker directions. |
| **II.E** | Summary scores were reported in terms of raw and standard scores. Performance levels reflect the range of scores possible on the assessment and use statements or symbols to denote each level. | Performance levels were articulated within the scoring rubrics. Students had to earn approximately 67% of the points available to receive a passing score. |

**DIMENSION III: REVIEW**

| **Task** | **Descriptor** | **Technical Evidence** |
| --- | --- | --- |
| **III.A** | The assessment was reviewed in terms of: (a) item/task distribution based upon the design properties found within the specification and blueprint documents, and (b) item/task and form performance (e.g., levels of difficulty, complexity, distracter quality, bias, and other characteristics) using pre-established criteria. | 100% of all items/tasks were reviewed by a department-level committee. Refinements to the test blueprint in the subsequent year will include 10% field test/try-out items. |
| **III.B** | The assessment was reviewed in terms: (a) editorial soundness, (b) document consistency, and (c) linguistic demand. | 100% of the items and tasks were reviewed for quality and content consistency. 70% of all points are obtained from items/tasks requiring at least cognitive demand of Level 2. No information was provided on either readability or linguistic load. |
| **III.C** | The assessment was reviewed in terms of the following alignment characteristics:   * Content Match (CM) * Cognitive Demand/Depth of Knowledge (DoK) * Content Pattern (CP) * Item/Task Sufficiency (ITS) | The items/tasks on the test form were “mapped” back to the Test Specification table, then blueprinted to the operational form. The refinement process document provided additional data in a summary table from the alignment review. |
| **III.D** | Post administration analyses were conducted on the assessment (as part of the refinement process) to examine items/tasks performance, scale functioning, overall score distribution, rater drift, etc. | The refinement process document provided additional information regarding those data collected and analyzed by the district. Item-level statistics (p-values, pt.-biserial, and omission rates). No data were collected on rater agreement within the district (planned for the next refinement cycle). Overall rating assignments for the human-scored tasks were provided. |
| **III.E** | The assessment has score validity evidence that demonstrate item responses were consistent with content specifications. Data suggest that the scores represent the intended construct by using an adequate sample of items/tasks within the targeted content standards. Other sources of validity evidence such as the interrelationship of items/tasks and alignment characteristics of the assessment are collected. | 1P Rasch Model fit indices provided evidence supporting the unidimensionality of the construct. Partial Credit Model (PCM) was used for polytomously scored (SA and ECR items. A log transformation converted raw scores onto the logit scale. INFIT acceptable parameters (*t*= -2 to 2) showed only one item flagged. INFIT SD = .58, which was within acceptable ranges. [Analytics provided by external vendor] |
| **III.F** | The assessment’s reliability coefficients are reported for the assessment, which includes estimating internal consistency. Standard errors are reported for summary scores. When applicable, other reliability statistics such as classification accuracy, rater reliability, etc. are calculated and reviewed. | The refinement process document demonstrated reasonable internal consistency (r = .72). Human-scored tasks had median values of 1.3 points for SA (2 points possible) and 3.2 points for ECR (4 points possible.). The mean standard error (.30 logit scale) was within acceptable ranges. [Analytics provided by external vendor] |