Test Score Validation Process

June 2010 (Ver 3)

PEARSON

Overview 3
Pearson Teams & Their Quality Assurance Roles 3
Pearson Partners 8
Test Development 9
Content Development 9
Florida Review Procedures 11
Initial Publishing Activities 13
Test Construction Process 13
Test Map (Define) Creation & Management 14
Content Monitoring of Quality during Forms Composition 16
PubOps Electronic Checks 17
Print Procurement 17
Key Verification Process 18
Test Processing, Scoring, and Reporting 19
Performance Scoring 26
Scorer Qualification 26
Pseudoscoring 28
Scoring Standards 28
Psychometric Analysis 33
Calibration, Scaling, and Equating 34
Final Score Replication 38
Checking for Missing Scores and Test Administration Anomalies 38
Document Retrieval 38
Caveon Data Forensics 39

Pearson’s Test Score Validation Process

Overview

Pearson’s quality assurance processes for managing test materials and assigning scores begin at the earliest phase, during test item development, and continue through the generation and distribution of score reports. These processes are characterized by collaboration across work disciplines and multi-stage validation. Pearson’s quality processes include these examples: 1) two content experts from Content Support Services (CSS) take every test without the benefit of seeing the correct answers in order to verify the correct printing of the items and coding of the answers, 2) the Assessment & Information Quality (AIQ) group independently validates that systems correctly manage data, including the machine scoring of student responses, and 3) psychometricians from Psychometric & Research Services (PRS) rescore student responses to multiple-choice items and compare their results to the scoring system results before conducting deeper analyses.

Pearson’s quality assurance processes can be broken into four categories: 1) Test Development, 2) Test Processing and Scoring, 3) Psychometric Analysis, and 4) Score Reporting. This report provides an overview of the various Pearson work teams and their formal roles in these quality processes, and it describes the various activities involved in validating student test scores. The report also identifies the portions of the quality processes directly overseen by the Florida Department of Education (FDOE) and the Test Development Center (TDC).

Pearson Teams & Their Quality Assurance Roles

Assessment & Information Quality (AIQ)

The Assessment & Information Quality (AIQ) group provides independent verification and validation of the software systems. This division of Pearson’s Organizational Quality group is independent from Pearson’s IT organization. The validation techniques of this group include:

- <u>End-to-End Testing</u>

  End-to-End Testing consists of executing test cases (referred to as a “mock data”) through the test environment in the same manner that live data will flow through the system. The system is validated holistically to verify that the system functions and integration points are performing as intended.

  The mock data processed through the system are created in the same manner as live data and test cases are created. For example, all of the data used to validate scoring are hand-gridded on answer documents or entered online through the online testing system in the exact manner that is used to process live data. This process is directly overseen by FDOE.

- Production Validation

  Once the End-to-End testing has been approved by the AIQ group, the software code and configurations are migrated to the production environment and live data are processed through the system. A sampling of the first live run through each of the software systems is validated to confirm that production data are processing through the systems as intended.

Content Support Services (CSS)

The Content Support Services (CSS) team is primarily responsible for test item and test form development. The team includes experienced educators (called Content Specialists) and assessment editors. This group contributes to quality processes in two critical ways:

- Test Item Development

  Pearson's CSS team verifies the content and grammatical accuracy of the text within test items and answer options before the items are presented to the Florida content team (referred to as the Test Development Center, or TDC). CSS then facilitates committee reviews of the items by Florida educators. The educator committees review the test items and answer options with a focus on presenting appropriate item contexts as experienced by Florida students.

- Test Form Publishing

  Test items and their correct answers are checked repeatedly at each step of the publishing phase to make sure they appear correctly on test forms. To provide independent checks on the content validity of the keyed responses, the test questions are answered by content experts who conduct a "final forms" review during the publishing phase.

IT Assessment Validation Group (AV)

The IT Assessment Validation (AV) group provides verification and validation of all software systems. This group is aligned and reports through Pearson's IT organization. The validation technique of this group is administration configuration testing (also known as *system testing*).

A full system test is conducted at the administration level that evaluates whether the system is fully compliant with the requirements and is functioning as outlined in the Functional Specifications. This testing is performed in the "test" environment where the new software is staged upon completion of software development and unit testing is complete. IT AV creates test cases with expected output definitions based on the functionality being performed and creates data based on test cases and requirement specifications. This testing includes:

- Generating test cases based on requirements and functionality.
- Generating test data based on test cases and expected results.
- Baselining and promoting software code and to "test" environment.
- Executing and validating test cases against expected results.
- Basing defects on test case failure and/or other requirement deviations.
- Executing a regression test suite after all software code is fully tested and all defects are closed.
- Updating IT Assessment Validation checklists and confirming all requirements are completed.

Software code is approved by the IT Assessment Validation group.

Performance Scoring Center (PSC)

The Performance Scoring Center (PSC) is responsible for scoring Florida's performance tasks. The PSC uses industry-leading score- and scorer-monitoring techniques. The PSC's quality processes are ISO 9001:2008 certified, which means that they meet the global standard for quality requirements. Florida's Test Development Center (TDC) staff continuously oversee the scoring process at each scoring site, have full visibility into the data, and are consulted on all decisions. The score monitoring techniques are as follows:

- Score Monitoring

  The quality of scores is monitored through several real-time quality metrics that are checked periodically throughout the day and tracked for trends. These metrics provide information on validity (or accuracy) and reliability (or consistency). Tolerance limits for these metrics are established with the TDC/FDOE and with industry standards as governed through the PSC's ISO 9001:2008 certification.

- Scorer (Rater) Monitoring

  Individual scorers must meet demanding minimum quality metrics during the training process and maintain those metrics throughout the scoring assignment. Scorers must meet these criteria before they are allowed to score actual student papers. Once they have begun scoring actual papers they are monitored continuously, and dips in scoring quality trigger actions such as more frequent monitoring, retraining, and potentially expunging of assigned scores.

Print Procurement

Pearson's Print Procurement group is responsible for monitoring selected print vendors and verifying that effective print quality procedures are in place. Their responsibilities include:

- Providing strategic direction, print production expertise, and test material procurement for Pearson's Assessment and Information group.
- Coordinating with Global Sourcing, suppliers, and internal/external customers for optimum quality, service, and value.
- Developing and implementing continuous improvement initiatives through supplier-quality forums, documented procedures, forecasting, cross-functional teams, supplier certification and audits, pre- and post-program lessons learned, digital workflow, and electronic printer proofs.
- Overseeing procurement of test materials for all custom state contract programs and catalog products.

Program Management (PM)

Pearson's Program Management (PM) group serves as the primary liaison between Pearson and the FDOE. This group manages Pearson's contractual relationship with Florida and oversees all production activities. The PM team works directly and collaboratively with FDOE/TDC and Florida educators to validate the quality of Pearson's work and respond to problems. The PM team implements the following quality processes:

- <u>Requirements Documentation</u>

  Quality begins with accurate and complete documentation of Florida's requirements. The PM team collects these requirements from the FDOE through interviews and review of documentation. The requirements are validated with FDOE before systems are configured.

- <u>Oversight and Approval</u>

  The PM team oversees implementation of Florida's requirements and confirms that processes are completed using program management techniques and principles. At various stages throughout test publishing and processing, the PM team uses the first processed batch of materials to verify that processes have been completed accurately. This production systems test is called a "blue dot."

Psychometric & Research Services (PRS)

The Psychometric & Research Services (PRS) group is composed of doctoral-level psychometricians and highly-trained staff who are responsible for test design and psychometric analysis. All psychometric work is directly overseen by FDOE doctoral-level psychometric staff. Pearson's PRS team is directly involved in quality processes in four critical ways:

- Statistical Validation of Answer Keys

  Before items are allowed to contribute to a reported score, the PRS team evaluates the statistical properties of those items in order to provide substantive evidence of the correct answer. The first time the statistical key check occurs is after field-testing. Before the question is added to the item bank, it must pass this evaluation. Each time an item appears on a test, the PRS team evaluates it again using the same statistical procedures.

- Certification that Test Forms Meet Psychometric Requirements

  The highest quality test forms meet blueprint requirements that dictate the span and breadth of content, and they must match demanding psychometric requirements for difficulty and parallelism with previously administered test forms. The PRS team verifies these requirements are met for each form and then approves the form before it is composed and printed.

- Validation of Scoring Before Equating

  The PRS team receives scored data from IT systems in order to conduct test analyses before students receive their results. As a standard practice, the PRS team confirms the correct application of scoring keys and the correct summation of total scores before conducting more in-depth analysis.

- Validation of Final Scoring Process

  Final scores for Florida students are assigned using a well-researched procedure that considers the items students answered both correctly and incorrectly. This procedure uses an algorithm and incorporates statistical features about the items administered to students. Pearson's PRS team developed the software used to assign these scores, and they oversee the implementation of that software during each reporting period.

Publishing Operations (PubOps)

The Publishing Operations (PubOps) group is primarily responsible for composing and publishing test content in both paper and online formats and overseeing the quality of printing from Pearson's print vendors. Pearson PubOps is ISO 9001:2000 certified for quality forms development and meets 20 ISO-required elements that are primarily aimed at achieving customer satisfaction through consistency and uniformity.

PubOps staff works directly with Pearson content and editorial staff to verify that all content is composed according to Florida specifications for style and quality. Since test form development is an iterative process and edits to composed material are a reality of the industry, PubOps uses flexible strategies that allow valid changes to be successfully and accurately made at any stage of the process.

The Test Map Team (TMT)

The document used by Pearson to score student responses on the test and by FDOE/TDC to validate the correct construction and scoring of the test is called a test map (historically referred to by FDOE as a "test define"). Updates to each test map document are managed by a subteam within PRS called the Test Map Team (TMT), which contributes to quality processes in the following ways:

- <u>Inspection of Contents</u>

  Test maps are created by content and psychometric teams by extracting data directly out of the Florida item bank. The data is then put into the format specified by FDOE. This serves as the source document for all Pearson scoring and reporting activities. The TMT inspects every value of this report, and if discrepancies are discovered, requires that these be resolved before releasing it for any purpose. Documents created by this inspection are detailed in a section below.

- <u>Change Management</u>

  Changes to test maps are a reality of the business and must always be approved by the FDOE. For example, items may be moved in order to conserve pages or field test items may be dropped during the composition phase. Change management is therefore critical. It is handled through a "two sets of eyes" process, whereby each change is carried out by one team member and then confirmed by another. All changes are further confirmed through an electronic compare process.

- <u>Quality Inspections at Touch-Points</u>

  The TMT conducts repeatable and thorough inspections of test maps and supporting data at various points through the process starting when publishing begins and proceeding through the last change requested to the document. These procedures include electronic methods in order to reduce the opportunity for human errors. The TMT also supports the quality efforts of other groups, such as CSS, by providing reports to verify answer correctness.

Pearson Partners

Caveon

Caveon is an independent subcontractor to Pearson responsible for investigating test score anomalies as described in the final section of this document. Caveon specializes in helping assessment programs enhance existing security measures and in analyzing test data for statistical evidence that might be associated with test fraud.

EdExcel

EdExcel is an independent subcontractor to Pearson responsible for the external proofreading of all test materials. EdExcel proofreading reports are presented directly to FDOE/TDC during the second round of forms review for each test administration.

HumRRO

HumRRO is an independent subcontractor to Pearson responsible for validating all psychometric activities that lead to student scores. HumRRO also independently validates the final scoring of student results and provides ad hoc replication and consultation services to both Pearson and FDOE. HumRRO uses its own procedures and software and collects requirements from FDOE independently of Pearson. HumRRO reports its findings directly to FDOE.

Test Development

Content Development

During the item development process, numerous steps are followed so that items will be developed to meet Florida's high standards. Multiple-choice items go through special steps to verify that they have a single correct answer and plausible alternate options (referred to as "distractors"). Open-ended question types, such as performance tasks and gridded-response items (or fill-in response items for online tests), also have special steps to verify that the rules for scoring the questions accurately identify all responses that are deserving of either partial or full credit.

A new development cycle is initiated for FCAT reading, math, science, and End-of-Course each year (with End-of-Course development beginning in 2009). Development cycles for writing are initiated once every two or three years. Before an item is used on an operational test form, it goes through 18–24 months of validation activities, including multiple reviews by content experts and validation by psychometricians. Florida FDOE/TDC staff and Florida educators (in most cases) participate in these reviews. Students are never scored on any item unless Florida educators and FDOE/TDC have already approved its wording, correct answer, and scoring rules.

Prior to the commencement of each development cycle, Pearson creates the Item Development Plan (IDP). This plan breaks out the number of items by subject, grade, reporting category, and benchmark that will be developed in that given cycle. This document is reviewed and approved by TDC to verify that it aligns with what is needed to build future tests given the content that already exists within the item bank. A variety of content foci are targeted when developing items so that there is diversity in item contexts.

As part of its quality processes, Pearson develops Item Writer Training Materials that are submitted to FDOE/TDC for review and approval. These materials and item specifications documents are provided to item writers in a yearly live training held by Pearson and overseen by TDC. This training facilitates the initial submission of quality multiple-choice items with solid distractors, or, in the case of constructed- and gridded-response items, clearly articulated scoring rules. All item writers are screened by Pearson and TDC to confirm that they meet the necessary requirements of expertise and experience required to write items for Florida.

Items are submitted by writers via the Test Developer's Studio (TDS), an online item-authoring system that enforces many of the rules governing Florida items. Items are then submitted to a first pass review by Pearson content specialists and are accepted, rejected, or marked for revision and resubmission by Pearson content specialists. Once an item is accepted by the content specialist, the item is reviewed again and revised by Florida's subject matter experts to make the item adhere to Florida standards for item format (e.g., for multiple-choice questions, a single correct answer must exist and the incorrect options must be plausible given the varying levels of difficulty that the items are written to meet).

Before items are submitted to the TDC for review, they go through a series of internal Pearson quality control and verification steps. Members of the Pearson internal review team generally do not write the initial version of items, which allows them to objectively evaluate the accuracy and quality of each item and its accompanying scoring rules. Items are senior-reviewed by the lead content specialist for quality of overall item, accuracy of content, benchmark match, grade appropriateness, and scoring rules (e.g., key or rubric). In the case of science, items are submitted with an authoritative reference to the scientific theory/application and evidence of the correct answer (or scoring rules) so the scientific correctness and relevance can be verified. To the greatest extent possible, Pearson submits items so that a variety of meaningful scientific contexts are addressed for each benchmark. Cognitive levels are addressed in roughly the same proportion that they appear in the test-design summary for each content area.

Items, passages, and context-dependent sets are also sent through two additional groups, Research Librarian (RL) and Universal Design Review (UDR), for review. Science context-dependent (CD) sets are used in a manner similar to reading passages. The "context" is a rich scenario for which several items are developed. The "set" is the combination of the items and the scenario. Pearson's research librarians investigate the items, verify sources, and ascertain that the contents of passages, items, and CD sets are valid. Items, passages, and CD sets are also reviewed by the UDR group to verify that they meet industry best practices for universal design.

Throughout these internal reviews, feedback is captured within the TDS system. Correctable flaws, such as UDR and RL violations, are corrected by the content specialists. Pearson's editorial team reviews items for grammar, punctuation, clarity, consistency, and concision and verifies that items adhere to Florida's requisite style guidelines. These edits are later incorporated by the content specialist if the edits are determined to be logical and do not impact content within the item. No item is submitted to the TDC for review until all members of the content review team approve it. Numerous test questions do not pass the scrutiny of this internal review and are never submitted to the TDC.

Florida Review Procedures

Once items have been initially accepted using Pearson's internal review process, they are formally submitted to the TDC for review, which takes place in the TDS system. The TDS software provides TDC content staff with secure, web-based access to the items and reduces or eliminates the need to print, package, and securely ship items from Pearson to the TDC for review. Training is provided to TDC staff members prior to Pearson's first item submissions. TDC staff members may request additional one-on-one training assistance from Pearson staff should the need arise. The training includes instructions for accessing, reviewing, and approving items; inputting feedback on items needing revision; and submitting results to Pearson for subsequent action. TDC content teams review items for benchmark match, content and grade appropriateness, single answer and plausible distractors, and correct sources for science. Items that require further action are reviewed and resubmitted and subsequently approved or rejected for committee review. TDC editors may also review and comment on items in TDS during this review period. Their recommended edits should be approved by TDC content staff before edits are incorporated.

The FDOE/TDC, with support from Pearson, conducts numerous review meetings each year with Florida educators. The purpose of these meetings is to receive feedback on the quality, accuracy, alignment, and appropriateness of the passages, prompts, scenarios and test items that are developed for FCAT Writing, FCAT 2.0, and the newly mandated Florida End-of-Course (FL EOC) assessments in Algebra 1, Geometry, Biology 1, and U.S. History. Item Review and Content Advisory committees are composed of Florida educators. The Bias and Sensitivity Review committees and the Science Expert Review committee are composed of educators, university professors (for Science Expert Review), and other Florida citizens selected by the TDC staff. The meetings are held at various venues throughout Florida. The roles of these committees are described below.

Community Sensitivity Committee

Community Sensitivity Committee members, representing various communities throughout Florida, are asked to consider and identify issues that might be sensitive in passages and items based on the wide range of cultural, regional, philosophical, and religious backgrounds of students throughout Florida. After a brief training session conducted by a TDC representative, committee members evaluate "sets" of passages and/or items provided to them in black-strip-bound review books.

Multiple copies of each set are provided to allow the simultaneous review of any given set by up to five reviewers. The reading assignments for each reviewer are organized so that each set is reviewed by a demographically representative sample of the committee members. Committee members sign out their review books (for security purposes), review the content according to the training, and identify concerns on a feedback form. Feedback is transcribed throughout the day into an electronic file by TDC staff. Upon completion of the review assignments, reviewers sign in and return their review books. Reviewers also complete an affidavit indicating which sets of passages and items they reviewed.

Bias Committee

Bias Committees include participants from both genders and a variety of multicultural backgrounds. Bias Committee members are asked to consider and identify passages or items that inadvertently demonstrate bias (e.g., racial, ethnic, gender, geographic, etc.) in some way. After a brief training session conducted by a Pearson representative, committee members evaluate reading, mathematics, science, and U.S. history items and reading passages and science context-dependent sets provided to them in black-strip-bound review books.

During each Bias and Sensitivity meeting, Pearson staff help track the number of reviews completed for each set of items or passages. An electronic file of reviewer comments is compiled, organized, and reviewed by TDC staff prior to or during Content Review Committees. The data inform decision-making about suitability of passages and items for placement on future tests.

Item Review Committees

Items at each grade level and content area are presented to committees of Florida educators in three-ring binders and in an electronic format projected on a screen. For each Item Review Committee meeting, a member of the Pearson staff keeps an electronic record of decisions made and documents any changes requested to item stems, options, art, or changes made to maintain alignment to the Sunshine State Standards (and, beginning with the current development, the Next Generation Sunshine State Standards). This electronic record is reconciled daily with the written record kept by the TDC staff member in charge of facilitating the meeting. Items are categorized as accepted, accepted with revisions, revised during the meeting, or rejected. Some items may be revised by an on-site Pearson staff member during the meeting for re-review by the committee. These revised items will be presented for review and approval to the TDC representative prior to the presentation to committee members unless arrangements are otherwise agreed upon by the TDC Content Lead and Pearson Content Lead. Item binders for committee members do not have correct answers indicated. This allows committee members to “take the test” for each question. This strategy is designed to help identify any items with miskeyed answers, multiple correct answers, or ambiguous wording. TDC and Pearson staff members are provided with the correct answers in their notebooks.

Science Expert Review

After the Science Item Review Committee meetings each year, a science Expert Review meeting is held. Florida universities and science research institutions are represented in this committee, and they review only the science items approved by Item Review Committees. The purpose of this additional review is to confirm scientific accuracy and appropriateness. After a brief training session conducted by a TDC representative, committee members evaluate items provided to them in binders. To verify that the items in each binder are reviewed by two different members, Pearson uses a sign-out sheet to track the reviews. Any inaccuracies or concerns identified by committee members are identified directly on the item and the feedback is compiled by TDC representatives. Items that have inaccuracies or concerns are later addressed and corrected by Pearson and TDC staff, or the item is rejected. An electronic file of reviewer comments is compiled, organized, and reviewed by TDC staff and shared with Pearson staff prior to the first round of test composition.

Initial Publishing Activities

After Pearson content staff have applied corrections to the items as indicated by Item Review Committees, the items are "composed," meaning that the item is moved from the TDS item-authoring tool and changed to a format for online or print publication. An additional review of the items occurs at this time. This review permits Pearson and TDC content specialists and editors to perform a final review of the content, art, correct answer, and plausible distractors. It also allows the content and editorial teams to verify the correct rendering of the content. TDC provides Pearson with official approval of all questions, passages, and context-dependent sets once items and passages appear in this composed format.

At this phase, PubOps also conducts a quality step called "preflight" that electronically verifies that the new items comply with Florida's style requirements. Preflight also verifies that the embedded art objects conform to system requirements so they can be accurately transformed to publishable format.

Test Construction Process

The selection of passages, context-dependent sets, and test questions appearing on all Florida tests are guided by Florida's test specifications (available on the FDOE website) and Test Construction Specifications that are commissioned and approved by FDOE/TDC. The Test Construction Specifications are drafted by the senior Pearson staff working on the project, including expert assessment content staff and doctoral-level psychometricians. This document extends the test specifications by providing the detailed guidelines, process, and clarification needed to select and place content on each test. The document is reviewed and edited by FDOE leadership, content, and psychometric staff. In the process of building a form, content specialists look at the parameters and then evaluate the existing item bank. Content specialists consult the test blueprint to make sure that reporting categories, benchmarks, and content foci are adequately represented within the assessment they are building as outlined by the test blueprint and item development plan. Items are reviewed as part of the proposed form to avoid clueing. Once Pearson content specialists have built a form, it is evaluated by a Pearson psychometrician. The Pearson content staff and psychometricians then engage in a collaborative and iterative process of refining the test form. Draft forms are not submitted to FDOE/TDC until they are approved by both Pearson content and psychometricians.

The FDOE/TDC review process is multi-staged in a similar manner as Pearson's. Iterative reviews take place first between TDC and Pearson, and then between TDC, FDOE psychometric staff and Pearson. FDOE/TDC provides tentative approval for forms in preparation for FDOE leadership review. A face-to-face review meeting is held to finalize the test build. FDOE and TDC staff, including content specialists, psychometricians, and FDOE leadership, confer with Pearson staff to complete the test builds. Refinement occurs until the best possible test forms are selected. Test forms must meet demanding expectations for match to blueprint and psychometric criteria. All items appearing on a form, the sequence on the form, and the scoring rules, are explicitly approved by FDOE/TDC.

Test Map (Define) Creation & Management

Throughout this process, Pearson and FDOE/TDC scrutinize each item for correctness and verify that the correct answers/scoring rules are indicated in the item database. Once tests are approved by FDOE/TDC, Pearson prepares a test map (previously referred to as "test define" by FDOE), which is an electronic record of the items, their position on the test form, the key and other scoring rules, as well as alignment to Florida's Sunshine State Standards. The test map is created by extracting the data elements from the electronic item bank, which is the original source of the scoring and alignment information. The test map is then verified manually against the composed test form.

Once complete, the test map is passed to the Pearson Test Map Team (TMT). The TMT is responsible for inspecting the test maps, normalizing contents for use in Pearson systems, document control, and change management. The test map created for Florida serves as the source document for all Pearson publishing and scoring activities.

Inspection Process

Prior to receipt of the first test map, the TMT interviews the Pearson teams creating and using the test maps to obtain detailed internal requirements about the project. These requirements range from the data elements appearing in the test map, to how test forms will be composed, to the data elements that will be used by the various scoring and reporting systems. This information is used to verify that the test maps contain the information needed for internal users to fulfill their objectives.

Each test map is thoroughly inspected for valid values and the elements needed for internal Pearson groups to conduct their work. Note that some data elements prescribed by FDOE/TDC to be included in the test map are for Florida use only, and are not used by Pearson systems. A series of reports are generated during the inspection process that are reviewed by the TMT and given to CSS for content specialists to validate:

- <u>Benchmark Report</u>

  This report lists the alignment codes reported in the test map and the number of items and points associated with those codes. This report is used to verify that the alignment codes conform to the specification, and that items and points conform to the blueprint.

- <u>Distinct Values Report</u>

  This report lists every value found in each column of the test map. This is valuable for identifying invalid characters. For example, data such as '1' (one) and 'l' (L) look very similar, and in some cases, '1' may be a valid character while 'l' may not be. If both characters are found in the same column they will both be listed this report, thus making obvious for visual inspection. This report is reviewed by the TMT to validate values needed by internal Pearson groups. CSS uses this report to verify the valid data elements uniquely used by FDOE/TDC.

- <u>Repeat Items Report</u>

  This report lists every test item that appears more than one time in the test map. In many cases, it is valid for the item to appear more than once. For example, a field test item may appear on two forms. The repeat items report is used to by TMT and CSS to verify that the scoring and alignment information for repeated items is identical in all places. It is also used to confirm that changes made to content occur wherever the content is used.

Document Control and Change Management Process

Each new document that the TMT receives first goes through the inspection process in order to verify the contents. Once the content is verified, the document is made read-only and given a version number. From this point on, the TMT makes all changes to the test map. When a change is requested, a copy of the original version of the test map is saved to an archive folder for historical preservation. The requested changes are made to the original, and an electronic comparison between the original version and the archived copy is made. A change report is generated and inspected to verify that the change was made correctly and to confirm that no other changes were inadvertently made to the test map. The change report is the posted, as is the change request, for a second TMT member to validate. A message is sent to the change requester, providing a link on Pearson's network to the test map and change report so the change requester can verify completion.

Quality throughout Test Map Production

Special quality control steps are conducted at three separate times during test map production. The first of these inspections is detailed above. The next two events occur after the PubOps group receives final approval to print the test forms. Step two is an electronic comparison of the test map to a data extract made from the test form. Embedded in the electronic files of Pearson's published content are data that are suppressed from the printed material. These data include several pieces of information about the test items on the form: the sequence number, the unique identifier, and the keyed response. Using a special script, the data are extracted from the electronic files and then compared to the test map. This verifies that the test book and test map are in sync with one another. Any discrepancies are resolved immediately.

The third step is an electronic comparison between the test map and the results of the final "Taking the Test" step from the Key Verification Process detailed is below. This step verifies that changes made to content that affect the correct answers to the test, if any, are identified before printing, and the test maps are appropriately modified. Discrepancies are immediately resolved.

Only after these steps have been successfully completed are the test maps released for use by Pearson's scoring systems.

Content Monitoring of Quality during Forms Composition

Once items are selected for a test form, Pearson content specialists review each publication-ready test form and compare the composed item in the test book against the previously-approved version of the item and the test map. Pearson editors also review the composed items against the previously-approved items and item bank versions of the items at the initial round and compare the files to the test map to make sure that any changes requested by TDC have been incorporated correctly. In subsequent rounds, Pearson editors compare edited items to the previous review round and continue verifying changes made to test maps. After each internal review at Pearson, the PDF files of the test forms are posted for TDC staff along with the test map. TDC editorial staff and content specialists review the composed forms along with the test map, and mark up changes to items and test maps, as appropriate.

Pearson also sends the test forms to a subcontractor, EdExcel, during the review process so the materials will receive an external review. EdExcel specializes in conducting editorial reviews of assessment materials, and the company has nearly 25 years of assessment review experience. EdExcel perform an unbiased review of all Florida assessments materials, including test forms, interpretive products, and test administration manuals. EdExcel reviews each test form and provides feedback about the items and suggested edits. These edit suggestions are then sent to TDC for review and approval. Upon approval by TDC staff, the edits are incorporated into the forms by TDC editors.

Throughout this process, Pearson and TDC editorial staff perform careful cross-checks to make sure edits have been applied across test forms for items that appear on multiple forms and that edits appear on both the PDF files and test maps, if necessary. TDC's edits are then reviewed by Pearson's content and editorial team, and any outstanding queries are forwarded to TDC content specialists for resolution prior to sending the files to the final publishing preparation. If there are any changes to test maps, these are applied by the TMT at each round during the composition process. This is an iterative process, which does not stop until TDC provides approval for both the test forms and test maps.

PubOps Electronic Checks

At the beginning of forms composition, Pearson designers (desktop publishing experts) once again execute "preflight" checks that verify correct application of Florida style, such as line weights, RGB color model, and fonts. If any potential quality issues are found, they are corrected using the processes for notification outlined in the FCAT Production Specifications and Editorial Style Guide. This guide is created with TDC and FDOE input and updated yearly.

After FDOE/TDC approves a test form to be printed, Pearson designers create the final print-ready file, by running the form through an automated print-ready file creation system that outputs a composite proof, a registration proof, and a separation proof. PubOps staff verify that each of these documents match client specifications.

Print Procurement

Quality Performance for Accurate Test Documents

Because of the absolute need for accurate test documents, Pearson conducts an annual review of print vendors. Variances are tracked and reviewed as part of a collaborative support program aimed at maximizing quality, accuracy, and performance.

Security (General)

Pearson has a long-established audit program and partnership with a select pool of print suppliers. These suppliers must meet a series of stringent security protocols in order to work on Pearson material. All files, film negatives, and plates are maintained in secure locations at the print supplier, with only authorized personnel permitted access to the material. All plates and film negatives are securely destroyed by the print supplier upon completion of a contract. At the end of each day's print run, authorized personnel securely shred all press overages and waste material. Each production run is made under close supervision of the printing supervisor. Test material is maintained in a secure manner at all times to preserve its integrity.

Quality Assurance (Printing)

Pearson has specifically chosen suppliers that will enhance the program quality. As a Pearson requirement, each supplier has implemented several additional quality procedures, including, but not limited to inspecting additional sheets/forms off press throughout each print run and installing an electronic Signature Recognition System into their binding process. These additional quality steps are designed to mitigate defective products by eliminating miscollations. Pearson's internal print facility is ISO certified. Pearson also works with outside print suppliers that are ISO certified.

Printing

Once the printer proofs are approved, another quality check is performed during the creation of press plates and at press to verify that no data has dropped from the document. The press operators are required to pull a specified number of print sheets to verify registration, pagination, print quality, and color consistency. Signatures coming off the press are stacked on a skid and tagged with a color-coded tag that identifies all signatures to verify a positive identification during transport to the next production station.

Bindery

Each Pearson-certified print supplier, including Pearson's internal print facility, has installed an electronic signature recognition system to prevent miscollations within a test booklet. The supplier's system will either electronically read a small bar code on the first page of each signature of a booklet uniquely coded to that specific signature or will read a predetermined image zone of a page to confirm that the correct signature is being processed. Any booklet that contains miscollated pages or a missing signature will cause an automatic shutdown of the bindery equipment for proper corrective action. During the binding process, a set number of books coming off the conveyor belt are pulled to confirm that proper quality control standards are met. As an additional quality measure, corresponding colored tags are placed on each individual pocket on the bindery equipment for visual check for accurate collation of materials. These materials are physically inspected by Quality Checkers within the manufacturing facility, and a predetermined number of these inspected booklets are sent to Pearson for review and approval. Pearson has developed these quality assurance standards, which exceed standard industry practice, to provide the highest confidence in the quality of Pearson's products.

Key Verification Process

Pearson mandates a four-phase key verification process for all test construction and publishing. These steps are essential to verifying that correct answers are identified in test maps.

Taking the Test during Test Construction

The first verification phase in the process occurs during test construction. While the test questions are being selected, the content staff working on the project scrutinize every selected question for the identified scoring information. The correct answer is confirmed for every question, and each incorrect option is verified as incorrect. If discrepancies are discovered, the issue is first reconciled with the item bank to verify that the comparison data is correct. Additional follow-up is made with FDOE/TDC if the question is flawed. In some cases the question is corrected, while in others, FDOE and TDC decide to change the status of the item to "do not use."

Creating the Test Map from the Item Bank

The second verification phase occurs when the test maps are created. Test maps are created electronically using the master database of information about the test questions rather than key-entering data into a spreadsheet. This reduces or eliminates the opportunity for key entry problems. The test maps are then checked by the Test Map Team (TMT) and content specialists for accuracy.

Taking the Print-Ready Test

When the electronic publishing files are sent to the printer, examination copies are provided to a team in Pearson's CSS group. This is the third phase of verification. Two content experts who do not work on the Florida project are provided with a test and an answer document. They read the test questions and respond to them by entering their answers on the answer document. If correct answer discrepancies are found during this process, they are reported immediately to Pearson's Florida CSS team and reconciled with the TMT. The TMT conducts an electronic comparison between the test map and the answer document used for this exercise. This is to verify that all issues have been reported. The TMT does not release the test map for scoring until evidence is documented that resolution has taken place.

Statistical Key Check

Finally, after the test is administered, student responses to the items are evaluated statistically to identify the presence of statistical flaws in the outcomes. See the section below on Statistical Key Check for details on this fourth phase of the verification process.

Test Processing, Scoring, and Reporting

Data Capture and Processing

During the Data Capture and Processing phase, paper headers and answer documents are scanned and processed through the editing systems prior to scoring item responses.

IT Assessment Validation

Test cases are designed to test each component (scanning, editing, etc.) independently to verify that all data are being captured properly and every editing rule is tested thoroughly.

Scanning

Expected results are created in the format that is expected to be received from the scanning system. Once the test deck is scanned, expected results are compared to the scan file to verify there are no discrepancies. The following are included in this validation:

1. Every response (bubbles and write-in boxes) on every header and answer document is hand-gridded using pre-defined patterns to verify that all data is captured properly and each response has no impact on other responses.
2. Multiple marks (double-grids) are hand-gridded to verify that the scanners correctly identify when more than one response is received.
3. Pages are extracted for every header and answer document. This verifies that the scanner outputs the data properly when pages are missing.
4. Every header and answer document is scanned with no responses gridded. This ensures that data is captured properly when processing blank documents.
5. Every bar code is scanned for each header and answer document. This verifies that all bar codes are being captured properly.
6. An answer document that cannot be scanned (cut corner of sheet) is included. This verifies that the scanner will recognize when an answer document cannot be scanned.

Editing

Data files are created to simulate the data coming out of the scanning system. This data contains both positive and negative (including “edge” conditions) test cases for every edit rule. Expected results are created for each test case. The data files are processed through the editing system and the output is compared to the expected results to verify there are no discrepancies.

- Assessment & Information Quality

  Test cases are designed to verify processing and editing of paper material is performing as intended. All software and interfaces are utilized and executed in the same manner that will process live data. The data processed through this system is generated from the material distribution phase. The edited data that is generated out of this system is used to test all the downstream systems (scoring and report distribution). The following are within scope of these testing activities:

- End-to-End Testing

1. Every field for every answer document and header is hand-gridded with pre-defined patterns to verify that all data is being captured, edited, and reported accurately and that no field has an impact on another.
2. Every answer document is hand-gridded using the max field lengths (all values populated for each field) to verify that all gridded values are picked up by the scanners, edited, scored, and reported properly.
3. Every answer document is hand-gridded with all item responses to verify that all gridded responses are being captured, edited, scored, and reported properly.
4. Every answer document is gridded with no item responses to verify data is captured, edited, scored, and reported properly when processing blank item responses.
5. Every field is hand-gridded with "edge" values (e.g., 0 and 9; A and Z) to verify proper data capture, editing, scoring, and reporting.
6. Cases to check the document count form processing (covering all grids, including special document types).
7. Cases to check errors on the document count form (blank DCFs, double-grids).
8. Cases where all bubbles are gridded to be sure they are being scanned properly, including each student demographic field, form field, accommodation fields, and item responses (see scannable document configuration chart and examples).
9. Cases to check multi-marks as well as missing data for each field.
10. Cases loaded based on the PreID files as well as cases that are added into the computer-based system by the school coordinators.
11. Cases to check procedures for duplicate testers in the PreID file (a student with the same last name and SID with two answer documents for one subject and grade level).
12. Cases to check all other score flag scenarios.
13. For each grade level and subject, cases to check that the PreID label information, if not blank, is overriding any gridded information for the birth date, gender, race and ethnic codes, primary exceptionality, ELL, and Section 504 data.
14. Cases to check procedures for duplicate testers after processing across both paper and computer versions if both are administered (may be computer generated duplicates).
15. Cases to check RMS rules for not meeting attemptedness, i.e., a student who answers fewer than 6 questions (may be computer generated; Score Flag = 2).
16. Cases for a variety of score ranges including raw scores of zero, perfect scores, and scores on each side of cut scores (may be computer generated).
17. Cases for each achievement level (may be computer generated; 100% in any one achievement level within a school).
18. Cases to check all aspects of reporting student scores and aggregated scores, including all flags used on the data files (may be computer generated, using

enough student records that results are displayed and not suppressed on mock reports; more than 9 reportable students in school).

19. Cases to check each item response area, including tracking of changed responses; some items should have erasures with another response gridded and some with just erasures.
20. For each grade level and subject, cases to check that history data are properly merged for each grade level (may be electronically generated; includes grades 3–9 only).
21. For each grade level and subject where the reading answer document is separate from the mathematics answer document, cases to check that the reading and mathematics records are merged properly (see scannable document configuration chart).
22. Cases to check that all PreID and security barcode information are being recorded properly (verifies Pearson access is bringing in correct data).
23. Retake cases to check that gridded grade level can override the grade level on a PreID label (check cases where retake reading and retake math grade levels conflict, and where DCF grade level and scannable documents grade levels conflict).
24. For each grade level and subject, cases to check that the aggregated suppression rules are being applied correctly.
25. For each grade level and subject, cases to check proper rounding on aggregated reports (may be computer generated).
26. For each grade level and subject, cases to check that special school types and district numbers are being processed correctly (may be computer generated).
27. For each grade level and subject, cases to check that school data are being properly aggregated to district data (see item 20).
28. When appropriate, cases to check that items answered in the computer-based tests in a non-sequential manner have responses associated with the correct item number in the test.
29. When appropriate, cases to check that paper-based and computer-based records for the same student are merged and reported properly.
30. For retakes, cases to check that the APS file is being used properly. A mock APS file may be generated (may be computer generated).
31. Cases to check all handscoring condition codes and rules (may be computer generated).
32. Cases to check hand edits resulting from torn books, unreadable barcodes, missing pages, pages out of order, etc.
33. All scanning and editing outputs are validated against expected results to verify the material is being processed correctly.

- Production Validation

1. Once live material is received from the systems, a sample of the material is selected. This sampling contains all answer documents and headers.

2. Once this live material has been scanned, a sampling of each header and answer document (both hand-gridded and pre-ID) is selected for validation.
3. Every response for every field for each of the samples is manually compared to the scanning output to verify that all data are captured properly.
4. Once the scanning validation is complete, the Assessment & Information Quality tester approves the documents to be edited.
5. Once editing is complete, the edits performed on the headers and answer documents are manually validated to verify that all edits performed are accurately reflected in the edited file.

Scoring and Report Distribution

IT Assessment Validation

Test cases are designed to test each component of scoring independently to verify that all data are captured properly and every scoring rule is tested thoroughly.

Online Testing

Test cases and automated compare tools are used to test the online forms creation process independently, to verify that all forms are authored correctly, and to confirm that the forms perform as expected.

All test forms for the given administration are authored in the “QC” region of Pearson’s production division. Each form is compared to its test map. A check is performed of all correct and incorrect option letters and fill-in response areas to verify that field positions match the expected keys. Comparisons are performed between the outputs of the test and the expected results to check for accuracy. This validates keys, item categories, score groups and reporting categories. Once testing is complete in the QC region, the forms are locked and published to production to prevent any changes from being made to the forms without testing. The following are within scope of the testing activities:

- <u>Scoring</u>

    1. Validate every possible raw and derived score for every form.
    2. Validate all attemptedness rules as defined by the customer (used to determine whether a student is considered to have taken the test.) This includes boundary condition testing.
    3. Validate Objective/Strand/Domain/Reporting Categories scoring (minimum to maximum).
    4. Validate all possible scores, including multiple-choice and opened-ended.
    5. Validate all possible incorrect answers, including a double gridded response and all possible correct answers, if applicable.

6. Validate that record changes are processed correctly from Schoolhouse into the scoring engine.
7. Validate that all demographic updates are correct and processed.
8. Item responses for all correct records on the Online Tests (eMS) Extract match OSA keys extracted for all forms.

- Reporting

1. All Reports are present, e.g., rosters, ISRs, and summary.
2. All scoring results displayed on reports match the student file extract.
3. All demographic fields are being displayed on reports and match the student file extract.
4. All aggregation rules determined by the customer are correct and are functioning properly.
5. All footnotes, legends, page number, page breaks, logos, etc., are correct.
6. All packing lists contain the appropriate content as per specifications.
7. All report-level rollups are validated, e.g., student, class, etc.
8. All inclusion and exclusion rules are correct.

- Assessment & Information Quality

  Test cases are designed to verify that scoring and reporting of both paper material and online testing records are performing as intended. All software and interfaces are utilized and executed in the same manner as used for live data. The data processed through this system are generated from the material distribution and data capture and processing phases. In addition to the test cases previously identified in the data capture and editing phases, the following are within scope of these testing activities:

- End-to-End Testing

1. Answer documents are hand-gridded with all correct item responses for multiple forms (both regular and accommodation) within each grade/subject combination to verify scoring is working properly and data are displaying correctly on reports.
2. Answer documents are hand-gridded with all incorrect item responses (including the minimum score for all open-ended and essay responses) for multiple forms (both regular and accommodation) within each grade/subject combination to verify scoring is working properly and data are displaying correctly on reports.
3. Every possible open-ended and essay response score point value is entered in the scoring system for every subject to verify scoring is working properly and data are displaying correctly on reports.

4. Every open-ended and essay response condition code is utilized for each subject to verify scoring is working properly and data are displaying correctly on reports.
5. Answer documents are hand-gridded for each grade/subject combination with item responses (including open-ended and essay) one below and one above the attempted calculation (edge conditions) to verify scoring is working properly and data are displaying correctly on reports.
6. Test cases 1–5 will be created for online records (electronic testing) to verify proper scoring of online tests.
7. All reporting exclusion rules (do not include non-attempted students, etc.) will be hand-gridded and/or included in online testing included to verify reports are accurate.
8. The quantity of student records will include enough to test all reporting page breaks.
9. If matching applies, multiple answer documents will be hand-gridded and/or online records created to verify the matching functionality and reporting results are accurate.
10. All scoring and reporting outputs (including data files, paper reports, and electronic reports) are validated against expected results to verify scoring and reporting are accurate.

- Production Validation

1. Once the live data are processed through the scoring system, the scoring is validated by selecting a sampling of the data to include multiple records for each grade and subject combination.
2. All demographic data are validated to verify that the final values are accurate.
3. Validate that reader scores for constructed-response items are correctly transferred to the student's record.
4. Validate that the correct number of readers scored each response.
5. Validate that the final scores on hand-scored tasks are correctly calculated.
6. Validate that all achievement levels are accurately assigned.
7. Validate that all pass/fail indicators are accurately assigned.
8. Validate that all aggregated scores are correctly rounded and reported.
9. Validate that records are merged properly for reading and mathematics.
10. Validate that computer-based and paper-based records are merged properly.
11. Validate that all records have a unique identifier across all grade levels and subjects tested in an administration.
12. Identify duplicate student records within and across districts.

Performance Scoring

Pearson's Performance Scoring Center (PSC) is composed of highly trained and experienced scoring directors, supervisors, and project managers, many of whom have several years of direct experience on Florida projects. Quality activities begin with the creation of detailed scoring specifications that conform to established ISO 9001:2008 criteria. The specifications are tailored to meet the needs of Florida and are approved by FDOE/TDC.

The materials used in performance scoring training, typically example student responses, are pre-selected by PSC staff and approved by TDC. The materials are identified for use in training, validity, or inter-rater reliability. Those used for validity and inter-rater reliability are introduced to and retired from use based on the number of times they have been used to maintain their effectiveness.

Individual raters must first demonstrate they have learned and can successfully apply Florida's scoring rules to carefully selected example student responses. Then they are allowed to score live student responses. The process and rules for reaching this point, called qualification, is overseen by TDC staff. After qualifying, raters must continuously meet demanding metrics for scoring quality to remain on the project. Raters can serve in various roles on the project based on their performance and experience, including Scoring Director (SD), Assistant Scoring Director (ASD), Supervisor, and Scorer.

Scorer Qualification

- <u>Supervisor and Scoring Director Qualification</u>
    - Scoring directors and assistant scoring directors are approved by FDOE/TDC to work on the project. They attend all validation meetings and assemble qualification sets.
    - Supervisor training and qualification occurs prior to the scorers' start date. This allows qualified supervisors to assist the scorers during the training and qualification window.
    - Qualifying criteria are higher for supervisory staff. This higher standard verifies that supervisors understand the content well enough to help scorers improve.
    - For scoring writing, supervisors must achieve an average of 75% perfect agreement and 100% adjacent agreement on the best two of three qualification sets. Supervisors must take all sets and average 75% on the best two of three sets, receive no score below 60% on any set, and receive only one nonadjacent score across all three sets. The SD or ASD will confer with any supervisor who assigns a nonadjacent score during qualification.
    - For scoring math and reading extended-response items, supervisors must achieve an average of 70% perfect agreement and 100% adjacent agreement across the three qualification sets. Supervisors must take all sets.

    - For scoring math and reading short-response items, supervisors must achieve an average of 80% perfect agreement and 100% adjacent agreement across the three qualification sets. Supervisors must take all sets.
    - Supervisory staff not meeting these criteria may be offered the opportunity to participate on the project in another capacity (e.g., as a scorer).

- <u>Scorer Qualification Standards</u>
    - Scorers in all content areas must meet the following qualification requirements prior to scoring:
        - Writing—an average of 70% on the first two sets with 100% adjacency and no set below 60% perfect agreement. (If scorers qualify on sets one and two, they do not take set three.) Scorers taking the third set must achieve an average of 70% on the best two of three sets with no set below 60% perfect agreement on the two sets that are used to qualify, and total nonadjacent scores across all sets cannot exceed one.
        - Math and reading extended-response—an average of 70% across two of three sets. There may be no nonadjacent scores on sets used to determine qualification. (If scorers qualify on sets one and two, they do not take set three.)
        - Math and reading short-response—an average of 80% across two of three sets. There may be no nonadjacent scores on sets used to determine qualification. (If scorers qualify on sets one and two, they do not take set three.)
    - Scorers who qualify above these rates and demonstrate a high level of accuracy during operational scoring may be promoted to Supervisor during the scoring window.
    - Scorers will be monitored throughout training to confirm that they complete the training within a reasonable time frame. Estimates for completing the training are:
        - Writing—24 hours.
        - Math and reading extended-response—eight hours, not to exceed nine hours.
        - Math and reading short-response—five hours, not to exceed seven hours.
    - Scorers that do not meet these minimum standards are terminated from the project.
    - If the number of qualified scorers is not sufficient to meet the schedule for the project, an action plan is implemented by the PSC Project Manager and PSC Functional Manager.
    - Writing scorers meeting the following criteria may be asked to continue the project in probationary status with increased supervision.
        - Scorers who average 65/100 across all three sets.

    - Scorers who average 70 across the three sets with no more than three nonadjacent scores.

Pseudoscoring

Pseudoscoring is a process where actual papers are scored as an extension to scorer training. This session provides scorers the opportunity to apply rubric criteria and established standards to live student responses. Pseudoscoring also allows the TDC and scoring directors to verify a scorer's understanding of the application of the scoring criteria and established standards to operational responses. Scorers are not aware that pseudoscoring is taking place so that this does not influence their performance. Pseudoscoring continues until an individual scorer completes a certain number of validity responses. (The number of responses varies by item, and is approved by the TDC.) Pearson's ePEN system computes the scorer's validity agreement and compares to the target value. If the scorer's validity agreement meets or exceeds the required value, scores applied by that scorer are retained. If the scorer's validity agreement falls below the required value, the scores applied by that scorer during the pseudoscoring session are expunged from the system. For scorers that do not meet the required level, additional supervision or training is provided and overseen by TDC. If additional supervision and training are not effective after a period of time agreed to by TDC, the scorer may be removed from the project.

Scoring Standards

The Scoring Standards used to monitor scorer performance and score validity and to address scoring quality issues include:

- Validity Delivery Rate (frequency or percentage can be used)
    - Writing: one for every seven responses scored
    - Math: one for every 25 responses scored
    - Reading: one for every 25 responses scored
    - Validity delivery rates may be adjusted throughout the life of the project by the PSC Project Manager in conjunction with the TDC.

- Validity Standard
    - Writing—60% daily perfect agreement; 70% cumulative perfect agreement; 95% perfect plus adjacent agreement
    - Math and reading extended-response—70% perfect agreement
    - Math and reading short-response—80% perfect agreement

- Scorer – Validity Intervention Standards
    - Writing—warning issued to scorers with less than 50% cumulative exact agreement after a minimum of 11 validity responses.
    - Math extended-response—warning issued to scorers with less than 60% cumulative exact agreement on validity after a minimum of:

Grade 5 at 30 validity reads
Grade 8 at 30 validity reads
Grade 10 at 24 validity reads

- Reading extended-response— warning issued to scorers with less than 60% cumulative exact agreement on validity after a minimum of:

  Grade 4 at 24 validity reads
  Grade 8 at 24 validity reads
  Grade 10 at 12 validity reads

- Math short-response—warning issued to scorers with less than 70% cumulative exact agreement on validity after a minimum of:

  Grade 5 at 48 validity reads
  Grade 8 at 48 validity reads
  Grade 10 at 48 validity reads

- Reading Short-Response—warning issued to scorers with less than 70% cumulative exact agreement on validity after a minimum of:

  Grade 4 at 30 validity reads
  Grade 8 at 36 validity reads
  Grade 10 at 24 validity reads

  - Scorers who receive warnings will be given additional feedback/training to aide them in improving validity responses.

- Final Warning and Termination
  - Writing
    - If the scorer has not met or exceeded quality expectations after another ten validity papers, a ten-paper targeted calibration set will be administered. The scorer must achieve at least 70% exact agreement on this calibration set in order to continue on the project. If the scorer passes the targeted calibration, the project standards for validity will still need to be met, and Pearson will continue to check validity statistics after every eight validity papers. If the validity agreement falls below-standard at one of these subsequent intervals, the scorer will be released from the project and all scores assigned by the released scorer will be reset.

  - Math Extended-Response
    - Grade 5—If the scorer still is not improving after another 30 validity papers, a ten-paper targeted calibration set will be administered. The scorer must achieve at least 70% exact agreement on this calibration set in order to continue on the project. If the scorer passes the targeted calibration, the project standards for validity will still need to be met, and Pearson will continue to check validity statistics after every 30 validity papers. If the validity agreement falls below-standard at one of these subsequent intervals, the scorer will be released from the

    project and all scores assigned by the released scorer will be reset.

    - Grade 8—If the scorer still is not improving after another 30 validity papers, a ten-paper targeted calibration set will be administered. The scorer must achieve at least 70% exact agreement on this calibration set in order to continue on the project. If the scorer passes the targeted calibration, the project standards for validity will still need to be met, and Pearson will continue to check validity statistics after every 30 validity papers. If the validity agreement falls below-standard at one of these subsequent intervals, the scorer will be released from the project and all scores assigned by the released scorer will be reset.
    - Grade 10—If the scorer still is not improving after another 24 validity papers, a ten-paper targeted calibration set will be administered. The scorer must achieve at least 70% exact agreement on this calibration set in order to continue on the project. If the scorer passes the targeted calibration, the project standards for validity will still need to be met, and Pearson will continue to check validity statistics after every 24 validity papers. If the validity agreement falls below-standard at one of these subsequent intervals, the scorer will be released from the project and all scores assigned by the released scorer will be reset.

- Reading Extended-Response
    - Grade 4—If the scorer still is not improving after another 24 validity papers, a ten-paper targeted calibration set will be administered. The scorer must achieve at least 70% exact agreement on this calibration set in order to continue on the project. If the scorer passes the targeted calibration, the project standards for validity will still need to be met, and Pearson will continue to check validity statistics after every 24 validity papers. If the validity agreement falls below-standard at one of these subsequent intervals, the scorer will be released from the project and all scores assigned by the released scorer will be reset.
    - Grade 8—If the scorer still is not improving after another 24 validity papers, a ten-paper targeted calibration set will be administered. The scorer must achieve at least 70% exact agreement on this calibration set in order to continue on the project. If the scorer passes the targeted calibration, the project standards for validity will still need to be met, and Pearson will continue to check validity statistics after every 24 validity papers. If the validity agreement falls below-standard at one of these subsequent intervals, the scorer will be released from the

    - project and all scores assigned by the released scorer will be reset.
    - Grade 10—If the scorer still is not improving after another 12 validity papers, a ten-paper targeted calibration set will be administered. The scorer must achieve at least 70% exact agreement on this calibration set in order to continue on the project. If the scorer passes the targeted calibration, the project standards for validity will still need to be met, and Pearson will continue to check validity statistics after every 12 validity papers. If the validity agreement falls below-standard at one of these subsequent intervals, the scorer will be released from the project and all scores assigned by the released scorer will be reset.

- Math Short-Response
    - All grades—If the scorer still is not improving after 48 validity papers, a ten-paper targeted calibration set will be administered. The scorer must achieve at least 80% exact agreement on this calibration set in order to continue on the project. If the scorer passes the targeted calibration, the project standards for validity will still need to be met, and Pearson will continue to check validity statistics after every 48 validity papers. If the validity agreement falls below-standard at one of these subsequent intervals, the scorer will be released from the project and all scores assigned by the released scorer will be reset.

- Reading Short-Response
    - Grade 4—If the scorer still is not improving after another 30 validity papers, a ten-paper targeted calibration set will be administered. The scorer must achieve at least 80% exact agreement. If the scorer passes the targeted calibration, the project standards for validity will still need to be met, and Pearson will continue to check validity statistics after every 30 validity papers. If the validity agreement falls below-standard at one of these subsequent intervals, the scorer will be released from the project and all scores assigned by the released scorer will be reset.
    - Grade 8—If the scorer is still not improving after another 36 validity papers, a ten-paper targeted calibration set will be administered. The scorer must achieve at least 80% exact agreement. If the scorer passes the targeted calibration, the project standards for validity will still need to be met, and Pearson will continue to check validity statistics after every 36 validity papers. If the validity agreement falls below-standard at one of these subsequent intervals, the scorer will be released

from the project and all scores assigned by the released scorer will be reset.

        - Grade 10—If the scorer has still not improved after another 24 validity papers, a ten-paper targeted calibration set will be administered. The scorer must achieve at least 80% exact agreement. If the scorer passes the targeted calibration, the project standards for validity will still need to be met, and Pearson will continue to check validity statistics after every 24 validity papers. If the validity agreement falls below-standard at one of these subsequent intervals, the scorer will be released from the project and all scores assigned by the released scorer will be reset.

- Inter-Rater Reliability (IRR) Standard
    - ▪ Summary of IRR guidelines or requirements in the Project Plan:
        - Writing— 60% perfect agreement
        - Math and reading extended=response—70% perfect agreement
        - Math and reading short-response—80% perfect agreement

- Inter-Rater Reliability (IRR) Intervention Standards
    - ▪ Scorers' IRR will be monitored, and interventions may occur based on exact plus adjacent IRR (not exact IRR alone, as a poor scorer could negatively affect a good scorer's IRR). Minimally, scorers will be monitored if they fall below the following standards:
        - Writing—less than 95% cumulative exact plus adjacent agreement after a minimum of 100 operational papers.
        - Math and reading extended-response— less than 95% cumulative exact plus adjacent agreement after a minimum of 100 operational papers.
        - Math and reading short-response— less than 95% cumulative exact plus adjacent agreement after a minimum of 200 operational papers.

- Supervisory Scoring and Backreading Requirements
    - ▪ Supervisors are expected to score one and a half hours per day for purposes of monitoring their scoring accuracy. Supervisors need to backread scorers based on the project plan. However, once validity stats provide enough information to evaluate scorers' performance, backreading is focused on those scorers who need more targeted monitoring.
    - ▪ Supervisors need to backread targeted scorers who are scoring outside of the parameters in this plan

- Reasons for Supervisor Interventions
    - ▪ Supervisors who do not maintain at least 65% for writing, 85% for math and reading short-response, 75% exact agreement for math and

reading extended-response and 95% for adjacent agreement on calibration, must score and achieve acceptable validity statistics to resume supervisory responsibilities.
    - Acceptable validity statistics for these supervisors are writing – 65%, math and reading SR – 85%, and math and reading ER – 75% for exact agreement and 90% adjacent agreement after scoring a minimum of ten validity papers.
    - Supervisors are expected to fulfill their roles. If a supervisor does not complete these tasks, a documented intervention will result.

- Scoring Directors and Assistants
    - Scoring Directors and assistants are expected to score one hour per day for the purpose of monitoring their scoring accuracy.
    - Scoring Directors need to backread their supervisors' work.
    - Additional backreading of supervisors or scorers who are not meeting quality standards may be required.

- Targeted Calibration
    - Targeted calibration will be used for scorer intervention and/or individual coaching.
    - Sets will be built as needed according to the individual intervention plan and in concert with TDC oversight.

- General Calibration
    - General calibration will be required a minimum of once a day and as needed.
    - Each general calibration will address a single issue and consist of one to three responses.
    - General calibration will be online after TDC approval or consent.

- Escalation of Validity and Calibration Papers
    - Escalation of papers will occur daily to verify validity. The calibration pool will also be continuously updated and papers will be approved by TDC staff.

Psychometric Analysis

Psychometric analysis for Florida tests is a highly collaborative activity involving doctoral-level psychometric staff from four independent organizations: FDOE, Pearson, HumRRO, and Buros. The procedures for psychometric analysis for Florida tests are well established, stretching back to the inaugural year of the FCAT program, 1998. The procedures Pearson uses for Florida incorporate elements required by FDOE, elements gained through experience, and developments in the science of educational measurement. Pearson's psychometricians and data analysts are engaged in almost every aspect of the program and have major roles in test construction, scoring, and reporting. The psychometricians also consult with Pearson's PSC staff on a routine basis.

Calibration, Scaling, and Equating

Calibration, scaling, and equating activities are the scientifically based analyses that lead to reported scores. The general activities, and outcomes, are documented each year in the FCAT Technical Report that is available on the FDOE website. Pearson psychometricians create detailed specifications several months in advance in order to facilitate planning and communication among the various participants. These specifications originate from the prior year's documentation and are carefully scrutinized for any special considerations that must be attended to for the upcoming administrations. Details included in the specifications include: file naming conventions, secure file posting directions, formats for input and output files, detailed directions and formulations for statistical analyses, and the responsibilities of each participant. All parties to the psychometric analyses are given ample opportunity to review and comment on the specifications, with clarifications made as needed. Further planning and preparation does not begin until all parties agree to the processes and details described.

Every process used by Pearson's psychometric team is transparent to the FDOE, as is the source code for software that is uniquely written for Florida to manage and analyze results. Pearson, HumRRO, and FDOE all use a commercial statistical software platform called SAS® to manage and conduct many of the analyses. Both Pearson and HumRRO provide actual SAS source code to FDOE psychometricians for inspection and validation each year. In addition, the commercially available MULTILOG© software is used to estimate (calibrate) Item Response Theory (IRT) item parameters that are used for item analysis and equating.

Each winter, after the specifications have been thoroughly vetted, the calibration teams (Pearson and HumRRO) deliver preliminary versions of the software to FDOE and all parties engage in a process practice session. This session uses mock student data that are formatted to the FDOE file specifications and generated to be a realistic facsimile of student responses in order to validate the processes and software that will be used in the analysis of the following spring student results. This practice session provides all parties with an opportunity to fine-tune their roles and systems for the upcoming spring, and gives FDOE an opportunity to evaluate the effectiveness of the process and each participant. No work with actual student data commences until the parties resolve any discrepancies encountered in the practice session, and FDOE has approved the final readiness of the process and software.

In addition to the external replication conducted by HumRRO, there are three replication/oversight activities conducted. First, the Buros Institute from the University of Nebraska at Lincoln observes all communication and reviews all documentation, including the specifications and results. They provide real-time consultation to the FDOE, if needed, and a full report on the equating activities. Second, FDOE psychometricians execute the software provided by HumRRO and Pearson to verify that it ran correctly, and FDOE psychometricians themselves engage in additional analyses to validate outcomes. Third, Pearson psychometricians conduct a series of replication and diagnostic activities for quality assurance:

- All student responses are rescored by Pearson's data analyst staff using the source keys. These rescored responses are compared to the official scores. If discrepancies are discovered, they are reconciled or corrected before data are turned over to FDOE or HumRRO.
- A statistical key check is conducted to identify items that may require closer review. This check is an extension of the key verification process discussed in the Test Development section. The goal of the statistical key check is to use statistical procedures and generally accepted criteria to identify items that are not performing to expectations. See below for details on the criteria used and the follow up procedure. The statistical key check and all follow-up are completed before equating analyses begin.
- Key elements of psychometric processes are conducted independently by two Pearson psychometricians, and discrepancies, if discovered, are reconciled before results are posted or discussed with FDOE.
- After the program MULTLOG is run, a plot of the empirical item response distribution is overlayed on the plot of the model fitted item characteristic curve. If the two plots are dissimilar, the MULTILOG implementation is checked to verify it was executed correctly. If it was, and very poor model-data fit is encountered, the issue is brought to the calibration team to discuss action.
- The item statistics from the current administration are compared to their item bank values. If differences are large, psychometricians consult with CSS staff to determine if the question was changed in some way, or if the position of the item is very difference between administrations. If issues are encountered, it is shared with the calibration team to discuss action.
- The percentage of students classified in each level of achievement is compared to the historical trend as a final step in validating the results. If the current year appears to be off-trend, the steps are retraced to verify they were conducted correctly, and cohort comparisons are made to determine if the trend is visible from a different data view. If issues are encountered, it is shared with the calibration team to discuss.

Statistical Key Check

The statistical key check is an extension of the key review process that uses empirical data to identify items that may have key or printing problems. There are two types of statistical key check procedures conducted for Florida. The first is done for multiple-choice questions where the following statistics are calculated: p-value (the proportion of students responding correctly), item-total correlation, and percentage of students choosing each option. Florida uses multiple base-forms so that items can be field tested during the regular administration instead of having a special administration that would inconvenience schools. On each base form there are a number of field test items that are not included in student scores. The scored items are the same. Because a printing problem could occur on one form and not another, the statistics are computed for each form, and questions can be flagged on one form but not others. A report is generated that flags items that have the following characteristics:

- P-value < 0.15
- P-value > 0.90
- Item-total score correlation < 0.20
- Incorrect option selected by 40% or more students
- P-value on any one form differs from the population p-value by |0.08 | for operational items that appear on multiple forms

Pearson psychometricians review the data for all scored items, regardless of whether the flagging criteria were met. The items flagged, the statistics computed, the key used in scoring, and the form the item was flagged on (if not on all forms) is provided to both Pearson and TDC content staff who first verify that the key used in scoring is the official key. They then review the question itself for correct key and correct printing. If the item is determined to be incorrectly keyed, and a single correct key exists, the test maps are corrected and a rescoring is conducted. If the item is printed incorrectly, FDOE leadership is informed and a policy process is engaged to resolve the problem.

For gridded-response questions, which appear in mathematics and science, the format is more open and students may provide a large number of different responses. These responses are normalized into a format that can be scored through an automated system. Many of these responses may be considered correct. During the test development process the expected correct answers are identified, confirmed by Florida educators, and placed in the test map. However, because of the open-ended nature of the question format, it is possible that students could provide answers that are correct, yet were not anticipated during test development process. Once the answer documents are processed each response is normalized and cataloged. A frequency distribution of each unique response is made and evaluated by TDC staff in order to identify any response provided by students that should be considered correct. If additional correct answers are discovered, the test map is changed, and a rescoring is conducted. As with multiple-choice questions, if the resulting data are otherwise unexpected, closer scrutiny is made, and they may be reported to FDOE leadership for decision-making.

Independence and Oversight

The four parties participating in the psychometric analyses work both independently and collaboratively. Final decisions are the exclusive purview of FDOE. The detailed roles of each participant are:

- <u>Pearson</u>

  1. Writes and maintains specifications and SAS code.
  2. Conducts the primary inspection of the data to determine if the information is accurate so that psychometric activities can be initiated.

3. Conducts the statistical key check.
4. Responsible for communication management. Sets up and leads daily conference calls during analysis window. Makes sure all parties are aware of issues and decisions.
5. Posts data to secure FTP, and maintains secure FTP.
6. Conducts and posts all computational results first.
7. Conducts interpretive analysis, and provides professional judgments about various solutions.
8. Investigates anomalous or unexpected results to verify correctness of data or outcomes.
9. Produces the official production scoring documents.
10. Archives all final documents, beginning specifications for next year as needed.

- FDOE

1. Oversees all Pearson and HumRRO work; approves all specifications.
2. Evaluates all outcomes. Judiciously replicates computations and explores alternative solutions to validate final decisions.
3. Seeks professional advice from Pearson, HumRRO, and Buros Institute.
4. Confirms Pearson and HumRRO results match. Oversees resolution process when they do not.
5. FDOE leadership, psychometricians, and content experts comprehensively evaluate all of the scaling and equating solutions, considering numerous factors related to content and statistics, and make a final decision.Verifies that the official scoring documents created by Pearson are accurate and reflect the decisions made by FDOE.
6. Replicates the final reported score computations using Pearson's and HumRRO's scoring programs before scores are reported.
7. Documents rationale for final decisions.

- HumRRO

1. Reports directly to FDOE. Provides independent advice, review, and replication of results using independently created systems.
2. Reviews and provides feedback on specifications.
3. Conducts data inspection and statistical key check.
4. Replicates all computational activities and posts after Pearson.
5. Conducts independent professional evaluation and provides professional judgment to FDOE.
6. Compares official scoring files to independently generated files.
7. Replicates the final reported score computations before scores are reported.

- Buros Institute

1. Reviews and provides feedback to FDOE on the specifications.

2. Attends conference calls.
3. Evaluates the process, and provides a written report to FDOE about the effectiveness of the process.

Triangulation and Depth of Investigation

Throughout the process the psychometric parties provide their own solutions and professional judgment, presenting solutions to each other and to FDOE. Computational procedures are compared and they are all required to meet to a demanding level of tolerance. Results not matching are painstakingly explored until identical results are achieved (in almost all cases), or the reason for the mismatch is ruled as immaterial by the entire team (this is a rare outcome). The independence of each of the three parties engaged directly in computational activities provides confidence that the best solutions are actualized. The collegial and collaborative approach that the parties take brings resolutions more quickly and minimizes communication problems.

Thorough the equating process, when judgments must be made, all participants thoroughly investigate the possible solutions in order to provide FDOE with the most complete information with which to make final decisions.

Final Score Replication

After final scores are assigned by Pearson's scoring system and validated by the IT AV and AIQ groups, two additional validations are completed by the psychometric groups. First, the Pearson psychometric group reassigns the reported scores using the source final scoring files (file approved for use by FDOE), and PC version of the scoring program used by the scoring system. This provides an internal Pearson replication of the reported scores. Third, HumRRO receives the same data and reassigns the final scores using their final scoring files, and an independently written scoring program. HumRRO results are reported directly to FDOE. Third, FDOE replicates scoring using both Pearson's and HumRRO's scoring programs. If discrepancies are found, FDOE and HumRRO confirm findings. The differences in scale score calculations are evaluated by FDOE's psychometric team and resolved before scores are reported.

Checking for Missing Scores and Test Administration Anomalies

Pearson uses two procedures to assist FDOE in this area: 1) document retrieval, and, 2) statistical investigations with Caveon Data Forensics.

Document Retrieval

FDOE often requests that Pearson provide student answer documents for the department to investigate missing data and other issues reported by districts. Pearson's system allows easy retrieval of these materials since all scanned documents are stored electronically. When requested, Pearson investigates the administration data and status of requested documents, and provides these documents to FDOE via overnight mail, if physical documents are required.

Caveon Data Forensics

Caveon employs Caveon Data Forensics statistical analysis methods to verify honorable student and administrator testing activities and to identify those test takers and groups whose score results must be considered indeterminate because of extremely aberrant test taking behaviors.

Initial Caveon Data Forensics analysis is conducted on the State Student Results (SSR) file provided by Pearson for all grades where the test group code is collected. Caveon completes its analysis and returns the identified inconsistencies and irregularities to Pearson in electronic format with outliers clearly identified. Pearson transmits these data directly to FDOE via the secure FTP.

Caveon works with the FDOE to develop and help interpret reports focused on the information of most interest to the FDOE and in a format that is easy to follow and use. Caveon Data Forensics analyzes test response data using patent-pending algorithms to identify instances of potential testing irregularities. The statistical algorithms detect anomalous[1] test response data for both schools and students. These algorithms are specifically designed to detect patterns that correlate highly with different types of potential testing irregularities[2]. Very conservative statistical thresholds are chosen in consultation with FDOE staff so that results will stand up under scrutiny.

Caveon Data Forensics uses four specific statistics to detect potential testing irregularities:

Possible testing irregularity	Detection statistics
Evidence of collusion among test takers	Looking for pairs of tests with large numbers of identical answers (for detecting answer copying and answer sharing)
Inconsistent student response across the test materials	Detecting response aberrance (answering difficult questions correctly and missing easy questions)
Erasures to answer choices	Analyzing erasure counts (for paper-and-

[1] An observation is statistically anomalous when the measured attributes are seen to be extremely different than the expected values for those attributes. A common euphemism to describe anomalous observations is “outlier.” Statistical practice for outlier detection or declaring an observation anomalous is usually based upon statistical tests where the probability value of the test statistic is extremely small.

[2] There are many types of testing irregularities. Some of these are due to natural disruptions such as power outages or extreme weather; others are due to behaviors that unfairly allow students, teachers, or administrators to obtain higher test scores. These include cheating (unfair access to the test content), answer-copying, collusion (efforts that result in two or more tests being more similar than would be seen by chance alone), and test coaching (sharing the test content or teaching the actual test items).

	pencil testing)
Changes in performance from test event to test event	Identifying gain scores (for detecting students or schools with improbable gains)

All initial reports include summary and detailed results as follows:

- The incidence of potential testing irregularities by school.
- If longitudinal data is provided, longitudinal effects on test performance.
- Anomalous tests that are associated with individual students.

For each test administration, Caveon delivers illustrations containing analyses of extremely similar tests. Additional case analyses or illustrations of anomalous tests may be made available as these analyses are refined within Caveon. The intent of these case analyses is to provide more detailed information to FDOE in understanding the nature of potential testing irregularities. The exact formats of these data vary depending upon the analyst’s investigation of the data.