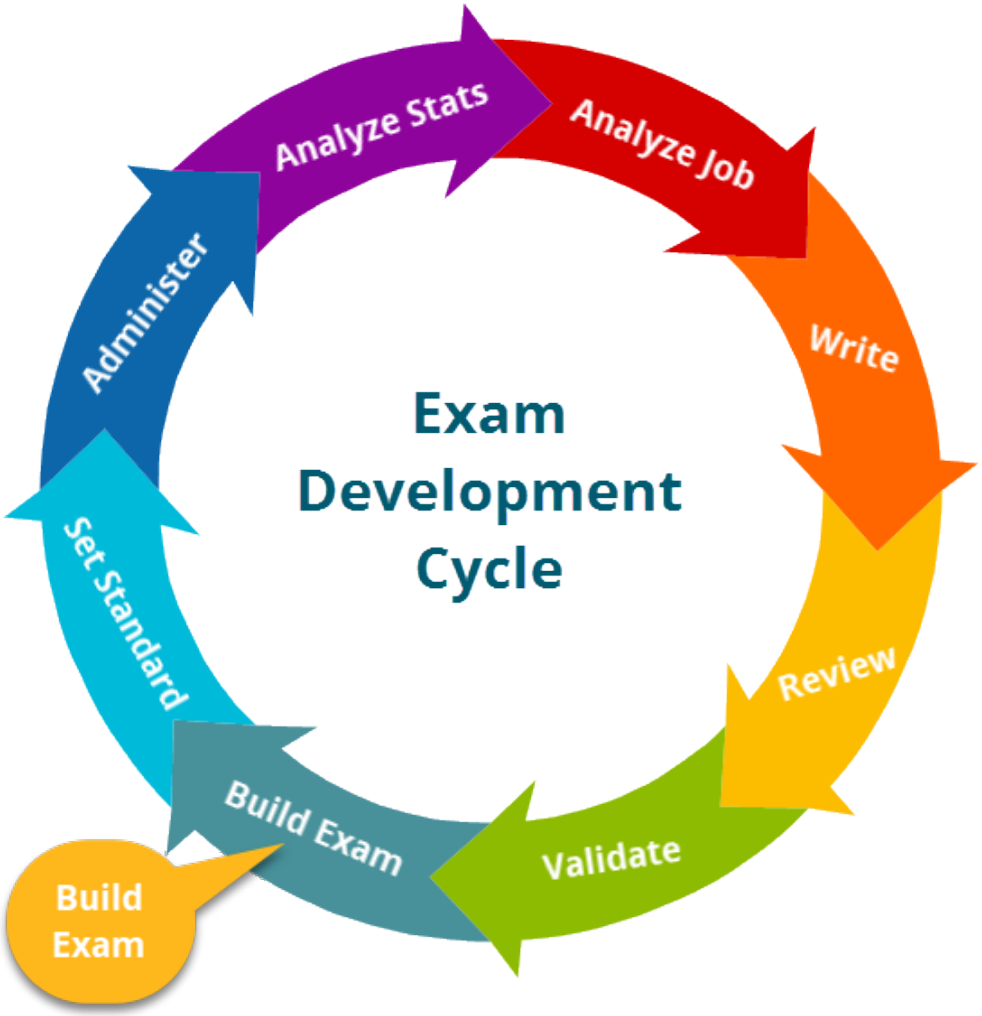
Examination Assembly

Once enough exam content has been developed to fulfill the test specifications, exam assembly can take place. This step also requires the participation of an expert panel.



During this process, the panel selects the best items in each area of the test blueprint. Should any final modifications be required in advance of publishing the exam form, it will likely occur during this meeting.

Building the Exam Versions

Building a high-stakes exam typically involves selecting items that fit certain criteria and placing them appropriately within the exam, much like a complicated jigsaw puzzle. Typically, subject matter experts spread a stack of items on a conference table and manually select the items based on criteria such as item quality, content area, cognitive complexity level, item difficulty, and category type. Due to the inefficient and error-prone nature of this process, organizations can look to online tools such as Exam Design's ExamDeveloper for assistance.

Often times multiple versions of exams are built simultaneously, where different items are selected for each version so that candidates do not always receive the same exam version. When this happens, the difficulty of each exam version is calculated to ensure candidates are not unfairly

penalized by receiving a more difficult exam version.

There are several important considerations when determining how many versions of an exam are needed:

How many candidates are being tested annually? With an annual candidate population of 500, fewer versions are required than if the annual candidate population is 10,000. A general rule of thumb is to use a new exam form after it has been administered to 1,000 candidates, or at least annually.

How often can candidates retake the exam? Candidates who fail a high-stakes exam are often given a chance to retake it after a waiting period (usually 6-12 months after the initial administration). Programs with a short waiting period require additional exam versions to reduce the chance of candidates taking the same exam version.

What are the consequences of failing the exam? Exams with substantial consequences for failure increase the likelihood that dishonesty or cheating may occur. Exams that preclude individuals from practicing in a particular field carry the greatest consequences for failure. Other examples include those that tie compensation to performance on the exam. In these cases, it is important to build and administer multiple versions of the exam to reduce the likelihood of cheating or other dishonest behavior.

Selecting the Right Content

One of the most important considerations when building versions of an exam is choosing the criteria by which items will be selected. Organizations are required to select exam items according to the duties and requirement of the job. In addition, several criteria can be combined to build multi-dimensional exam blueprints.

Job/Task Analysis Specifications. At a minimum, high-stakes exams should be built according to the tasks, knowledge, and skills identified through the JTA process. In order to establish content validity for the exam, a systematic process must be used to select items based on the relative importance, criticality, and frequency of the duties and requirements for competent job performance. In other words, job tasks which are twice as important or frequently performed in comparison to others should receive twice as much weight on the exam.

Cognitive Complexity Level. Organizations can also choose items so that there is not an overwhelming majority of items which target one level of cognitive complexity. Typically, this is used to prevent the exam from consisting of predominately recall-based questions, as item writers find that type of item the easiest to develop.

Content Categories. Organizations may also choose to identify a separate set of content categories to build a two-dimensional matrix of content area by tasks performed. As an example, imagine an exam for veterinary medicine which seeks to evaluate competence in regards to the following professional tasks: gathering patient histories, conducting assessments, diagnosing symptoms, and developing treatment plans. Items could be selected using a two-dimensional matrix according to species so that there would be three items related to gathering patient histories for dogs, three items related to developing a treatment plan for cats, and so on.

Keywords. The exam can also be assembled according to keywords. Keywords are typically important phrases or concepts that are contained within the item. Organizations typically use keywords as a filtering mechanism when building exams so that there are not several items covering the exact same concepts on an exam version.

Difficulty. Another factor used in building exams is an item's difficulty. Organizations should select items such that the difficulty of one exam version is set to be equal to that of other versions. Difficulty can be measured using a variety of statistical techniques. One of the simplest is the percentage of people who answered the item correctly. This method is useful when item difficulty is already known through the use of those items during previous exam administrations.

Ensure Independence
When reviewing an assembled exam, ensure that each item is independent. That is, ensure that knowing the answer to one item will not allow one to answer another item.

Reviewing a Draft Exam

It is also at this stage that substantial review is required. Recall the types of review we discussed before in the **Peer Review Process** section:

1. Content
2. Sensitivity
3. Psychometric
4. Editorial

However, in this case, we will add a fifth type to the list: **Format Review**

Format review is required to confirm that the display of the exam content is satisfactory. For example, on a paper and pencil exam, this review confirms that the items are laid out correctly on the pages. For an exam being delivered via computer, one would want to confirm that all items are displayed correctly on the screen and that any graphics or attachments are shown clearly.

Should each item count?

Scored versus Unscored Items: Depending on the equating method that is chosen for the program, your exam may have a mix of scored and unscored items. Scored items mean that the candidate's response to an item, right or wrong, counts toward his or her final score. In some higher stakes exam programs, only items that have been tested previously can be selected as scored on an exam. Using new or recently modified items as "Unscored" is a responsible practice when one is unsure of its characteristics. One can have a committee of 10 Subject Matter Experts review an item and miss that there are two possible correct answers. It isn't until the exam is administered to thousands that one learns of this.

Balance the Key

When a draft exam of multiple-choice items has been assembled, one should review the key location on each item to ensure that the distribution is relatively balanced. The number of items with the key in the "A" position does not have to match exactly the number of items with the key in the B, C or D position (on a four-option examination). However, it should be close.