TEST ITEM WRITING
BEST PRACTICES

Useful and Practical Considerations

January 2007

©2007 The Donath Group, Inc. All rights reserved.
TEST ITEM WRITING BEST PRACTICES

This document is intended to be used as a practitioner’s guide for test item writing. The information used in this document is subject to change without notice and should not be construed as the only correct way to write high quality test items.

Donath brings extensive experience providing professional examination development services, psychometric and evaluative research consulting services. In addition, Donath provides project management services throughout all our exam development efforts. Donath also offers professional certification exam development consulting and training to help clients develop their own high stakes certification programs.

Donath works with clients to offer customized solutions to meet specific test development needs. This guide is based on years of experience in writing, reviewing and evaluating test items and their performance in many high stakes certification examinations. This guide is only available to The Donath Group’s clients, potential clients or designated business partners.

Unauthorized copying or distribution in printed, electronic, or any other form in whole or in part, is prohibited without the prior written permission or by contract with The Donath Group, Inc. (Donath).
# Item Writing Best Practices

## Introduction

- **Purpose** ........................................................................................................... 1
- **What is a test?** ................................................................................................. 1
- **Why test?** ........................................................................................................ 2

## Human Resources

- **Program/Project Manager** .............................................................................. 3
- **Subject Matter Experts (SMEs)** ..................................................................... 3
- **Reviewers** ...................................................................................................... 4
- **Psychometrician** ............................................................................................ 4

## Exam Development Process

- **Job/Practice Analysis** .................................................................................... 4
- **Exam Blueprint** ............................................................................................... 4
- **Item Writing** .................................................................................................. 5
- **Item Review** .................................................................................................. 5
- **Field Testing Items** ......................................................................................... 6
- **Test and Item Performance Analysis** ............................................................. 6
- **Standard Setting** ............................................................................................ 6

## Characteristics of Good Tests

- **Validity** ........................................................................................................... 7
- **Blueprint Adherence** ..................................................................................... 7
- **Test Objectives** ............................................................................................... 8
- **Statistical Soundness** .................................................................................... 8
- **Fairness** .......................................................................................................... 9
- **Professional Standards** .................................................................................. 9

## Test Item Construction

- **Test Item Formats** .......................................................................................... 10
  - **True-False Questions** ................................................................................... 11
  - **Multiple-Choice Questions** .......................................................................... 11
  - **Use of Graphics/Exhibits** ............................................................................. 11
  - **Multiple-Response** ....................................................................................... 11
  - **Simulations** .................................................................................................. 12
- **Cognitive Processing Levels** .......................................................................... 12
  - **Demonstrating Understanding** ..................................................................... 13
  - **Scenario-based questions** ............................................................................. 13
  - **Problem solving given a scenario** ................................................................. 15
INTRODUCTION

Purpose

The Test Item Writing Best Practices guide was created by The Donath Group, Inc. (Donath) to help subject matter experts (SMEs) in their efforts to write fair, reliable and valid test questions.

Donath often trains SME item writers, who have diverse backgrounds and a wide range of skills to build credential examinations. SMEs generally do not have experience or training in writing test items and require reference materials so that they may quickly learn the fundamentals to construct good quality questions.

Test item writing is a skill that takes much effort to develop. Item writing has a lot of the qualities of an art form that is thoroughly steeped in the application of statistical theories and empirical research. A well written and useful test item combines a skilled application of language arts and known statistical parameters.

A brief summary of the test development process is included to provide a general understanding of what is necessary to create a viable examination. This summary description should help SMEs and program managers become aware of what is needed in test development and how to meet the overall objectives of developing high quality test items; thereby creating, fair, reliable and valid examinations.

What is a test?

A test can be considered a systematic process of interviewing examinees to determine whether specific skills and competencies are at a desired level. Unlike a free-form interview, a test accomplishes this in a uniform, standardized format that is fair, efficient, reliable, valid, and cost-effective.

A test is intended to provide a representative sample of examinee behavior for a well-defined domain of content so that accurate decisions can be made concerning the proficiency of an examinee. A test score is a summary of evidence contained in an examinee’s responses to questions. The proper interpretation of responses is critical to the usefulness of an examination’s results in meeting the purpose of the test.

There are several different test development models. Donath utilizes criterion-referenced mastery testing as our primary test development model. A criterion-referenced test measures what a person knows or can do, as opposed to an aptitude test. An aptitude test is generally designed to measure relative proficiency and performance comparing examinees to other groups of examinees.
Introduction

A criterion-referenced test compares a test taker’s exam performance against a well-defined criterion or content area, whereas, norm-referenced tests compare examinees to the performance of the entire population of test takers or to specifically defined comparison groups. Criterion-referenced testing allows us to define the minimum proficiency needed for the examinee to pass the exam and be confident that they have the skills to do a specific job successfully.

Why test?

Tests offer a standardized objective evaluative process for determining a candidate’s proficiency to fulfill a job role or identify mastery of specific knowledge or skill areas. The creation of a certification exam is a cost-effective method for evaluating large numbers of examinees. The alternative is interviewing each candidate for a job position to determine, through a subjective process, if a person is at a proficient level. Tests can supply reliable and valid measures for decision-making, but the use of multiple measures from a variety of sources will always supply a more complete picture of a candidate’s capabilities.

Tests are used to:

• help motivate learning.
• diagnose learning difficulties.
• assess the effectiveness of instruction.
• provide a measure of progress (strengths and weaknesses).
• determine whether or not trainees have the prerequisite training.
• determine whether a person has the knowledge and skills to do a job.
• determine whether or not trainees have attained the training objectives.

Certification programs create tests because there is a major need for decision tools to efficiently evaluate personnel to perform jobs so that when they are given a job, it is done the right way the first time. A certification program’s ultimate objective is to work towards ending rework, correct failed operations, avoid loss of product market share, to protect the public and to increase customer satisfaction.

Testing programs essentially benchmark standards of performance that inevitably set expectations regarding performance and raises standards to motivate the workforce for quality improvement. Benchmarking standards is an important outcome that can be realized by many organizations.
The top reasons candidates seek certification is to:

- increase personal credibility and self-confidence.
- assess knowledge and skills for professional growth.
- demonstrate technology currency.
- fulfill job requirements.
- help find or switch jobs.
- prepare for a new position.
- increase productivity.
- increase compensation.

A certification program must be congruent with the reasons candidates seek certification and what tests provide in decision-making through a well-articulated program. Program decision-making needs to fulfill candidate and organizational expectations.

**HUMAN RESOURCES**

A number of well-trained and experienced individuals in different areas, are needed to develop high quality examinations. Usually a knowledgeable program manager is the key person who initiates the identification and coordination of program activities.

**Program/Project Manager**

The program manager or project manager is responsible for coordinating the personnel associated with a test development effort. Donath provides guidance and assistance to the program/project managers throughout the process. Test development is a collaborative process, requiring constant communication of information to all parties for the test development to proceed in a timely manner. Additionally, the program manager is involved with marketing the program, vendor and stakeholder relationships, and certification management services.

**Subject Matter Experts (SMEs)**

Subject matter experts are the key contributors to every test development project. They are the individual technical experts in the field of knowledge being tested. SMEs help define the composition of tests, write and review test items, and validate that the test items are representative of the domain defined in the test specification (blueprint). SMEs are usually known as highly proficient job performers and are recognized for their skill and knowledge.
The strength of the exam as a whole relies on the contribution of the SMEs. It is especially important to engage SMEs to further develop their item writing skills, their test item ideas, and to allow them to get up to speed quickly when they are asked again to participate in item writing projects.

**Reviewers**

Donath has experienced reviewers that supply feedback to item writers and ask challenging and critical questions concerning their test items. Item writers should also review each other’s work to learn from one another and to critically evaluate how well test items are being written to the test specification.

**Psychometrician**

Donath has a staff trained in the application of statistical theories for evaluating test item responses and test scores for the purpose of building sound measurement instruments. A psychometrician can build equated test forms and utilize a variety of methods to set test score standards of performance.

---

**EXAM DEVELOPMENT PROCESS**

The following is a brief overview of the certification exam development process identifying the key activities performed in order to develop valid and reliable test score interpretations for examinations.

**Job/Practice Analysis**

A job analysis involves gathering information concerning the behaviors that are important for performing a job. Information gathering can be done a number of ways such as email, telephone, face-to-face interviews, focus groups, or structured questionnaires. The purpose of the job analysis is to identify the most critical aspects of a job. The purpose is also to determine scope, range, and depth of the skills and competencies that will be measured by the test and set standards of performance.

**Exam Blueprint**

An exam blueprint is the framework for the structure of the test. The blueprint defines the exam’s scope and content of skills being measured by the test, as well as the hierarchical importance of each content area. The main content area is called the **domain**. Within each domain there are “sub”-content areas or **test objectives** (**enabling objectives**).
The blueprint specification is the engineering design plan for constructing an examination and it clearly defines the enabling subskills that are measured. It provides information to certification candidates of what is on the test and supplies item writers information on the areas where test items are to be written.

**Item Writing**

Quality test items are essential in developing and publishing a successful examination product. To begin the process, SMEs are given training on item writing best practices. Using best practice guidelines, SMEs are asked to create test items that precisely measure the test objectives prescribed by the blueprint. When test items are clearly written to measure a test objective and a number of independent experts can verify this relationship, it is called item-objective congruence.

Not all test items are perfect when first written. Most do not strictly follow templates or best practice and need further refinement and development to work well. Donath seldom finds the initial construction of test items have the ideal elements of a well-written test item. Early drafts of test items need constant attention and revision so they can be made to work as expected.

**Item Review**

Test items should undergo rigorous reviews as well as several separate independent reviews. Group reviews of test items can be useful, but should be conducted carefully. Donath has discovered that independent blind reviews usually supply much more critical review information for test item improvement than group reviews. There is often a tendency within groups to allow certain people to dominate the review process while others acquiesce to their decisions without conducting enough serious evaluation on their own.

Group reviews are good for learning what is a good test item compared to a flawed test item, but should always be balanced with independent contributions. Do not allow a full room of SMEs to sit through a slow review of each test item when they could individually review the test items independently and supply their feedback to a test development vendor or project manager who compiles the comments from a small group on the test item.

Whenever possible, a testing program should attempt to enlist the support of a group of reviewers on a regular basis. Once an organization has a well defined operational process in place and a set of trained reviewers available, test item and test quality will always improve. Test questions will never improve beyond a medium level when an organization constantly changes the people working to write and improve the quality of their test questions.
Exam Development Process

SME Technical Review. This type of review is conducted by those other than the item writers who wrote the test item. The review should verify item-objective congruence; check for ambiguity, difficulty levels, relevance, and technical accuracy.

English Grammar Review. This review should verify that grammar, usage, readability, clarity and consistency of usage occur appropriately for each test item.

Psychometric Review. The psychometric review should verify the test item structure and the items conform to the recognized item writing guidelines, (language, style, sensitivity) and ensure that defined standards are met. If pilot exam data is available, the statistical characteristics of the test items are included in this review process. Psychometric reviews should be conducted throughout the item review process.

Field Testing Items

Test items should always be piloted or at least undergo a tryout with appropriate small groups who have similar characteristics as those who will be assessed with the final published test. A tryout is conducted to see how well the test items work in doing their respective jobs of discriminating between those who score well and those who don’t score well.

Test and Item Performance Analysis

Test item response data is analyzed and the statistical results are used to select the best test items for the final published exams. The standard setting process determines the appropriate cut scores and is generally conducted during this phase of test development through a variety of statistical and judgemental methods.

Standard Setting

A location on the test score scale is determined as a pass point or there can be a series of cut scores that delineate levels of achievement or proficiency. These defined levels across the test score scale are arrived at through a standard setting process. There are many different systematic methods used for selecting a passing standard for examinations. Each method has its own set of activities and procedures to follow; however, each have the same result in defining a cut score or scores. Here is a short list of methods for setting cut scores.

- Angoff method
- Modified Angoff method
- Nedelsky's method
- Ebel's method
The core question in assigning a cut score is the assessment of decision consistency. Are the candidate scores that award mastery truly representative of masters, and are the candidate scores that are determined not to be masters of the content, truly non-masters and not proficient?

CHARACTERISTICS OF GOOD TESTS

Good tests always have a well-defined plan for measuring the important skills and knowledge of credential candidates. Tests should always be developed on a sound scientific basis with a clear rationale for item writing and test score interpretation.

Validity

Test score validity can be thought of as the collection of evidence that supports the purpose of the test and test score interpretation. The documented scientific process of building a test, assessing its statistical characteristics, and how this information is used to determine mastery and non-mastery is evidence of test score validity.

Blueprint Adherence

A test needs to have a specification that depicts a clear relationship with performing a job function, mastering skills, or with a knowledge area. Tests need to sample the full array of only the most important behavioral objectives uniformly and without emphasis in specific areas, unless that emphasis is desired because of importance or frequency in performing a job.

A test has a planned number of questions at a level of difficulty and discrimination to best determine mastery and non-mastery performance states. Examinee’s should clearly understand what is needed in education and training to prepare for the examination and how much experience performing certain activities would help in preparation. This should be the road map that helps item writers create test items and helps test takers understand what will be required of them to pass an examination.
Characteristics of Good Tests

The test blueprint can be thought of as an engineering design plan. It lays out exactly what will be created and essentially what the test looks like in measurement and human performance terms.

Test Objectives

Test objectives are part of a test specification and should directly relate to what is required to successfully perform a job or master a content area. Each test item in a test should match or be congruent with a specific test objective. Useful test objectives should provide enough information so that an item writer understands what to write about but yet not constrict their writing.

Avoid very narrow or detailed defined objectives which can lead to trivial test items that are few in number. Avoid having numerous test objectives within a test blueprint that will never have any test items in the exam. An item writer’s creativity may actually be inhibited by narrow test objectives.

Donath has found that simple and direct test objectives with general lists of enabling sub-skills or content work well to clearly delimit what needs to be covered. This approach in defining what is measured permits the item writer to release their creativity within a broader area. Generally, more interrelationships can be drawn together for problem-solving and critical thinking types of questions, thereby measuring high order thinking skills in the content area.

Test objectives should use action verbs, such as those listed below.

- recognize, identify, demonstrate, translate, describe
- interpret, reorder, distinguish, classify, differentiate, discriminate
- contrast, evaluate, synthesize, assess, appraise, judge

Item writers will find the actions and direct objects of those actions as appropriate to assessing levels of cognitive functions.

Statistical Soundness

A test must:

- measure the examinee’s ability or proficiency in applying knowledge and principles.
- have a low probability of yielding a high score for test takers who have not performed the job or studied the content area.
Characteristics of Good Tests

- have a high probability of yielding a high score for test takers who have been independently determined to be high achievers in the content area or in performing the job.
- contain good test items, that supply dependable (reliable) and valid information about very important content areas. This information is determined through test and item analyses and subject matter expert judgments.

Fairness

Fairness is a social rather than a test and measurements concept. Fairness has no single meaning or single definition. One definition suggests that different groups have similar or equal outcomes in test scores. Another meaning is that all examinees have equitable treatment in terms of testing conditions, such as, access to practice materials and performance feedback, retest opportunities and other test administration conditions. In testing, fairness indicates there should be an absence of bias in test score results and equitable treatment of all examinees in the testing process.

Generally speaking, the majority of those who have successfully completed the learning program or perform the job for which the test was developed must perceive the test as a “fair test.” When examinees don’t see the examination as fair, legal issues may develop. Testing programs should always ask examinees how they perceive the test experience.

Professional Standards

High-stakes tests should always meet professional standards of practice that are appropriate given how test results are used. There are published standards that provide the criteria for the evaluation of tests, testing practices, and the effects of test use. These standards provide a framework of reference to assure that relevant issues are addressed by test development professionals.

TEST ITEM CONSTRUCTION

Test item development is a critical step in building a test that properly meets certain standards. A good test is only as good as the quality of the test items. If the individual test items are not appropriate and do not perform well, how can the test scores be meaningful? Therefore, test items must be developed to precisely measure the objectives prescribed by the blueprint and meet quality standards.

Test items are the building blocks of an exam. All test items are composed of three parts: item stem (question), correct answer(s), and distracters (incorrect responses). The term “options” refers generally to all the choices that are available. The person responding to the question can either select or construct the appropriate response to answer the question depending on the item format presentation. We will only discuss selected response test items in this document. Constructed response questions require different forms of scoring rubrics but still follow similar guidelines in how questions are formulated.

In the development of test items, there are a number of considerations. In this section, we present some item writing guidelines. We will discuss:

- Test Item Formats
- Cognitive Processing Levels

Our primary concern is how to create well-written and useful questions in the most cost effective manner. These guidelines should provide useful information towards this goal; however, it does take time and practice to write top-notch questions that can elicit responses that truly represent a person’s knowledge and skills.

Test Item Formats

There are many test item formats that can be used in a computer-based examination. The most popular formats include, but are not limited to, essay, short answers, multiple-choice, multiple-response, matching, and simulations. These item formats can be scored either objectively by computers or subjectively through judgments by human evaluators. The following discussion solely relates to objectively scored test items where examinees select a response(s) to a question and it is scored by a computer program.

The multiple choice and multiple-response formatted test items are the most popular since they can be scored easily and reliably by machines as compared to examinee constructed responses. Also, these item formats are relatively easy to write and place in published examinations for either paper exams or computer-administered exams.
Readers should not be surprised by what the most popular question formats are since they have probably seen these formats throughout their education. The multiple-choice format is sometimes criticized since some laypeople consider it a poor way to evaluate a person’s knowledge and skills. We often hear this format referred to as “multiple guess.” Yet, research has shown that these items do perform well when they are well constructed.

**TRUE-FALSE QUESTIONS**

This test item format is presented first since it is the simplest selected response format. The true-false question is a statement of fact and gives the option to select True or False as an answer choice.

This true-false item format, if used, should be used infrequently (< 10%). This item format can work well when constructed properly. However, more often than not, this item format fails to perform well statistically. Donath has come to recognize the failure of this item format and always recommends not to use them at all in developing test questions. Time is much better spent on developing multiple choice or multiple response format questions. Most true-false questions can be rewritten into a multiple choice question.

**MULTIPLE-CHOICE QUESTIONS**

Multiple-choice format is made up of an item stem, an answer, and 3 or more distracters when possible. Two distracter questions can work if it is difficult to write another distracter. There is only one correct answer for this format and can be written so that it measures not only knowledge of facts but can be used to evaluate high order thinking that requires problem-solving or critical thinking.

**USE OF GRAPHICS/EXHIBITS**

“Hot Area” Graphics or Exhibits can be used within this format as well as with multiple-response formats. The format requires an examinee to choose or identify where a specific location is on a picture (graphic) by clicking on it. The hot spot graphic has to have areas identified as incorrect choices and an area that is correct.

**MULTIPLE-RESPONSE QUESTIONS**

This item format has an item stem and more than one correct answer. Essentially, this is a combination of two or three multiple choice items in one. It is generally more difficult to answer and also discriminates very well between those who are proficient and those who are not proficient with the subject area being tested.
Donath, in unpublished research, compared multiple-choice, multiple-response, matching, and simulation test items. Donath found that multiple response items can work as well or better in tests as simulation test items using measures of item reliability. Multiple response items can supply a highly useful means to measure skills in comparison to the expensive development costs of simulation item formats.

When constructing a multiple response item, consider the following:

- Plan to have two correct answers out of five choices or three correct out of five or six choices. Always remove distracters that are not being selected by examinees.
- Do not use “Choose all that apply” instead identify the number of choices that are needed to supply a complete correct response. It is important to provide information to examinees as to the number of correct choices as a matter of fairness.
- Identify the number of correct options. Use the phrase “Choose XXX” in the item stem, and present it in parentheses preceded by spaces. For example: (Choose TWO) or (Choose THREE).
- Score test items so that selecting only the correct options count as a correct response. Do not give partial credit or accept examinee selection of two correct options, but also selected a third incorrect option. Score the item as incorrect.

SIMULATIONS

Simulation test items are computer delivered questions that evaluate examinee responses to a simulated experience of a software application or situation where the examinee actually performs the requested task within the simulated activity. Responses can be evaluated much the same as multiple-choice questions. There are correct responses and there are incorrect responses.

Cognitive Processing Levels

Test item writers should always attempt to write test items that measure higher levels of cognitive processing. This is not easy to do. It should be a goal of the writer to ensure their items have cognitive characteristics exemplifying understanding, problem-solving, critical thinking, analysis, synthesis, evaluation and interpreting rather than just declarative knowledge. There are many theories that provide frameworks on levels of thinking which, from this authors experience, has rarely had any serious impact upon writing good items. Bloom’s taxonomy is often cited as a tool to use in item writing. Nevertheless, this taxonomy has yet to have any research to support its effectiveness in helping in item writing. Always stick to writing important questions that represent and can predict that a candidate is proficient at high levels of cognitive processing in doing their jobs proficiently.
DEMONSTRATING UNDERSTANDING

Test items should always follow a consistent design so that the questioning process in itself doesn’t add unnecessary difficulty to answering questions. Therefore a logical and consistent stimulus format for writing test items can help expedite the arduous process of writing test items as well as supply a format for asking basic questions.

A format that provides an initial starting structure to use in writing questions can be valuable for item writers. When these formats are used, test takers can quickly read and understand the questions, since the format is expected. For example, to measure understanding of knowledge or facts, questions can begin with the following:

- What best defines ....?
- What is (un)characteristic of ....?
- What is an example of ....?

SCENARIO-BASED QUESTIONS

A scenario-based question provides a series of statements that outlines a scheme or sequence of events in summary or in synopsis form, followed by a question concerning what should be discovered or deduced from the scenario.

A scenario question essentially lays out a syllogism for an examinee to draw inferences based upon what they know in the subject area. A syllogism is a form of deductive reasoning consisting of major premises, minor premises, and a conclusion. For example: All humans are mortal, the major premise; I am a human, the minor premise; therefore, I am mortal, the conclusion.

A scenario question presents reasoning from the general to the specific within a deductive reasoning process. The deduction leads to selecting the correct answer choice(s).

This item format is very useful, especially for problem solving or troubleshooting types of questions. Scenarios can be written to describe a situation and ask the test taker to determine the correct answer based on the information provided. This question type can easily utilize either multiple choice or multiple response item formats.

Generally speaking, scenario questions assess key cognitive proficiency that are outlined by test objectives. The questions themselves are devices to represent proficiency. Scenario questions should represent one or more of the following to evaluate higher level thinking.

- identifies reasonable points of comparisons
- identifies evidence, key concepts and principles
Test Item Construction

- describes facts, situations and predicted outcomes
- identifies cause-effect or other critical relationships
- recognizes how inferences are related to what is given
- demonstrates clear purpose for grouping of structural components
- demonstrates sound extensions for generalizations that fit specific situations
- identifies, collects and analyzes related evidence to possible outcomes or consequences

The construction of scenario-based test items are likely to experience SME discussion and criticism, as they should. There has to be careful wording in describing the situation, assessment of clarifying the major points presented, and how various outcomes might be interpreted.

The item writer and reviewers should be aware that excellent test items don’t always come from individuals, but come from teams of people working together to improve the quality of their writing.

EXAMPLE SCENARIO QUESTION

Jen was hosting a party when she warned John not to walk on the balcony, explaining he could slip and fall on the balcony ice. John wanted to view the snowy landscape so he disregarded Jen’s warning and was injured when Jen’s balcony collapsed when he stepped on to it.

John filed suit in a jurisdiction that follows common law contributory negligence. Jen acknowledged that she knew the balcony was in need of major repairs, but argued that John was negligent because he disregarded her warning.

What is the most likely outcome from Jen’s defense of contributory negligence?

A Unsuccessful because John’s decision to disregard the warning was not the proximate cause of his injury.

B Unsuccessful because Jen will not be able to prove, beyond a reasonable doubt, that John was negligent.

C Successful because Jen could prove, by a preponderance of the evidence, that John disregarded Jen’s warning.

D Successful because John's negligence in disregarding the warning was the proximate cause of his injury.

“A” is the best choice because, courts, following contributory negligence hold that the plaintiff's negligence must be the proximate cause of the injury. In this case, John may have been negligent with respect to the hazard of slipping on ice. However, he did not voluntarily
expose himself to the risk of the balcony collapsing, which was the proximate cause of his injury.

This is a good question since it takes facts and knowledge of the law to predict a plausible outcome.

**PROBLEM SOLVING GIVEN A SCENARIO**

Scenario-based items can utilize these forms of question construction formats after presenting the situation under investigation.

- What is the nature of the problem?
- What is needed to solve the problem?
- What will occur from......?
- What is a solution?
- If this happens, what should be done?
- What is the most effective or efficient solution?
- Why is … the most effective solution?

**DEMONSTRATING CRITICAL THINKING OR EVALUATING SKILLS**

- What is the most effective (appropriate) method for ….?
- What is better (or worse) ….?
- What is the most critical step in this procedure?
- What is (un)necessary in a procedure?

**DEMONSTRATING CRITICAL THINKING IN PREDICTING**

- What would happen if ….?
- Given …. what is the primary cause?
- On the basis of …, what is the primary cause?
CHARACTERISTICS OF GOOD TEST ITEMS

Tests don’t work without well written test items. Examinees appreciate clearly written questions that do not attempt to trick or confuse them into incorrect responses.

Donath has discovered that when examinees fail a well written test, they don’t usually blame the test and actually understand that they were not prepared enough to take the test at that time. However, when examinees see and believe that the test questions were not written well, and appear ambiguous in what the test as a whole is measuring, testing program administrators will be inundated with complaints. What is worse; disgruntled test takers tell everyone they know that the test is not worth taking or the certification program is not worth the effort since the program did not put the effort into writing good questions.

The following presents the major characteristics of well written test items.

Item Independence and Dependence

Each item must stand alone as a specific measure of proficiency. Test items should be independent of all other test items and not disclose information about any other item in the exam. They should not depend on other test items for information.

Statistical test theories make the test data assumption known as local independence. This means that an examinee’s responses on any particular test item is unaffected and statistically independent from a response to any other test item. Local independence presumes that an examinee approaches each test item as a new problem without any information gained from responding to any other test item. When the local independence assumption is violated, proper test score interpretations fail.

If an item relates to another item, it should be tracked. This ensures that the items are not presented on the same form of an examination since it is a redundant measure of proficiency. When items measure the same information, they are referred to as: variants, or clones, which are defined as:

**Variant** - an item that covers the same content area as another item, although it may be worded differently and may have different distracters. A variant may reverse stem content with response option content.
Characteristics of Good Test Items

Clone - nearly identical in questioning and has nearly identical distracters that may be in a different order.

Another item type that should be tracked so that it doesn’t get into an examination with other test items is known as a dis Poster. This item really should not exist since test items should never provide information that answers other questions.

A Discloser is an item that “gives away” or “discloses” information that another test item can use to successfully answer its question. These types of items aid an examinee in choosing the correct option. Always try to identify these test items and rewrite them so that they don’t provide information that helps examinees. Or, if used, be sure that the test items they disclose information on are located on different forms of an examination.

Item Relevance and Level of Difficulty

Test items should focus only on one specific behavioral objective that is critically important for mastery within a content area. That is to say, a test item should only be a measure of one test objective. Independent observations by subject matter experts should reliably identify that a test item indeed matches the test objective and it is at an appropriate level of difficulty for the audience being certified. Items should always be written to an “appropriate difficulty” level as defined by the test specification. When they are not, the items should be reviewed, rewritten, or eliminated.

Overall Item Construction Strategy

An overall item construction strategy should be defined prior to starting item writing. Not only should a good plan be in place for writing test questions in specific content areas, but test items should be developed with certain characteristics that all may recognized as useful in asking good quality questions. Here a few standards that should be in place for writing.

- Always use the simplest, uncomplicated, free of irrelevant material and clearest language possible.
- Vocabulary should never be a relevant source of difficulty.
- Always use active voice in the present tense without using personal pronouns.
- Never assume a superior stance or attitude in questioning.
- No grammatical cues to the correct or incorrect answer should exist in a test item.
- With scenario questions, use isomorphs (e.g., different questions based on similar formatted situations). This helps in readability for the examinee.
- Always attempt to create answer choices that are homogeneous in content, length and context.
Characteristics of Good Test Items

- Avoid special knowledge issues or too much detail on non-critical points. Never write questions on trivial points. Questions must relate to activities that involve at least 80% or more of the job’s major activities or knowledge area.
- Always avoid writing “trick” questions.
- Avoid using humor in any form in test questions.

Item Stems

Questions should always try to begin with “Who,” “What,” “Where,” “Why,” “How” or “When” whenever possible. There may be some introduction information that outlines a problem in a few statements, but the test item should have a clear question that needs to be answered.

“Which” may not need to be used at all in most tests. Too often “Which” is over used and doesn’t add any extra meaning that did not already exist in the questioning format. Avoid using “Which of the following.” or “Of the following,” these phrases are not needed; which is always implied by the test item format and the examinee will select from the following options is part of the test taking directions. Generally a simple “What” communicates better even when followed by a noun. Examinees are well aware that the possible answers are listed below. It is a direct insult to examinee intelligence to keep saying, “Which” of the following...

Item stems should:

- be written as short and uncomplicated as possible.
- only assess a single piece of knowledge or skill per item.
- always be presented as a complete sentence in a question whenever possible.
- be free of irrelevant material. Only supply enough information to answer the question.
- be well thought-out, clearly worded, and unambiguous.
- clearly define the problem. The examinee should know exactly what is being asked.
- always ask a direct question and avoid any lecture or lengthy discussion in the question.
- always contain the central idea and most of the content for the item. The central idea should not be in the options.
- use “should” instead of “would” or “could,” for example, “What should be done…?”
- use “must” instead of “can,” for example, “What must be done…?”
- avoid using the words “always,” “all,” and “never.”
- present questions positively; avoid negative phrasing.
Characteristics of Good Test Items

- avoid using words such as: “not,” “cannot,” “except,” or “never,” if possible. If they are retained, then they should be represented in UPPERCASE **bold**.
- watch out for double negatives. If a negative must be included in the stem, do not use a negative in the response options. This makes responding to the question very confusing for an examinee.

**Item Responses**

There are some basic guidelines in writing the item response options that an examinee selects as their answer. Here are a few of them.

- The correct answer(s) must be 100% correct, 100% of the time.
- The distracters are clearly not the correct answer(s) when the best answer(s) is selected.
- Responses should always be independent and mutually exclusive.
- Concentrate on quality not quantity of distracters. It is better to have fewer good distracters; it does matter how the distracters work for the quality of the test item.
- All distracters must be plausible to someone who does **NOT** possess the skills being tested.
- There is only one correct answer for each multiple-choice item and two (or more) for multiple-response.
- Avoid using: “Don’t know,” “All of the above,” or “None of the above” as distracters or answers.

The response alternatives of “all of the above” and “none of the above” carry with them subtleties that can be easily overlooked by item writers. They should not be used without considerable forethought. For example, none of the above as the correct answer may not really represent if a person knows the correct response. They would only need to recognize that the options listed are not correct.

**International and Cultural Considerations**

When exams are distributed internationally, either in a single language or translated to other languages, always refrain from the use of slang, geographic references, historical references or dates (holidays) that may not be understood by an international audience. Tests need to be adapted to other languages so that meaning is fully translated correctly.

Other terms to be considered:

- Avoid item content that may bias gender, race or other cultural groups.
Characteristics of Good Test Items

- Avoid use of language slang of any kind.
- Monetary references (not generic; it varies by country).
- Idiomatic expressions that are particular to a country, e.g., “...as a rule of thumb...”
- Measurement, e.g., lengths, weights, etc. Use internationally accepted terms.
- Reference to dates, use the international date style standard of: day, month, and year, e.g., 26-August-01.

Item Writing Style Considerations

Test items should always follow an organization’s editing and style definitions. SMEs should focus on developing high quality test items, but should be cognizant of word usage according to corporate or organizational standards that may be presented. It is not critical for item writers or item reviewers to be concerned with editing considerations, but item editors will need to have this information to perform their jobs appropriately.

GENERAL STYLE CONSIDERATIONS

- Numbers under 10 should be spelled out, i.e. – “nine”, not “9.”
- Whenever a test item refers to an exhibit, begin with “Click the Exhibit button.” Present the question after this statement.
- For “hot spot” graphics, use the following expression: “Click on the area of the graphic that shows ...”
- When there are multiple-response test items, indicate the number of correct answers. Place two spaces at the end of the stem and add: “(Choose XXX).”
- All response options are listed in logical order; for example, numerical options may be listed in ascending or descending order.
- Repetitious words or expressions should be removed from the choices and included in the item stem. This helps to minimize the time needed to read the items.
- Fill-In-The-Blank. When you require the use of “fill in the blank” in the middle of the item stem, use 7 underscores to represent the blank. For example: “It requires _______ to resolve the problem.” If possible, consider rewriting the question as a closed-ended question. This could be written as, “What is required to resolve the problem?”
PUNCTUATION AND CAPITALIZATION

Here are some examples of areas that may need to be defined for item writing styles.

- Include appropriate punctuation, and capitalization, etc.
- If the item stem is a sentence fragment (to be completed) – do not end the statement with a “colon” unless the item stem uses words like “the following” to introduce a list; in which case, use a colon at the end of the phrase.
- Item responses do not require articles such as “the” and “a” to preface the response. If an article is used and repeated, include it in the item stem. In the case of “a” and “an” being used, refer to this as “a(n).”
- Space between sentences. In computer-based exams, put only one space between sentences versus two spaces.
- Try to avoid the use of personal pronouns. Keep the questions as gender neutral as possible. If pronouns are necessary, use “he” or “she” in test items to give both genders “equal time.” Nevertheless, try to eliminate the use of pronouns completely.
- Acronyms and abbreviations should be defined (spelled out) with the acronym in parentheses. However, common acronym usages can be used if it is very fundamental to the job. For example, “ASCII” is a commonly known term, and spelling it out may confuse some.

If the stem is asking the examinee for multiple responses, then the noun should be plural. For example:

- “What tasks are performed (Choose TWO)?”
- “What task is performed?”

ORGANIZATION EXAMINATION STYLE CONSIDERATIONS

Always refer to company-specific style guides to ensure standards are applied for acronyms, syntax, and specific product names so that organizational copyrights and trademarks are protected from erosion.

Always establish the correct presentation of commonly used terms so that they are presented and used properly, and consistently. This includes the following (with related examples shown):

- capitalization (IRQ2)
- naming conventions (checksum is one word, hard drive is two words)
- use of hyphens (RJ-11, SCSI-2)
- abbreviations (MB=megabyte)
Characteristics of Good Test Items

- program names (COMMAND.COM – all capitals)
- file folder/path naming conventions (C:\STYLE\ITEM_WRITING)
- presentation of commonly used terms (MS Windows 2000)
The following citations are among the most significant books and articles published that cover areas of research and application in testing. There are many more sources available for the interested reader. These books will cite many other sources of useful information for understanding psychometric theory and practice.

**GENERAL**


ITEM WRITING


CRITERION-REFERENCED MEASUREMENT


CERTIFICATION


TEST SCORE EQUATING


VALIDITY AND RELIABILITY


FURTHER READING


DIFFERENTIAL ITEM FUNCTIONING


