Authentic Testing White Paper

Item Writing for Performance Tests

Wallace Judd, PhD

Performance Test Development Series 05

Authentic Testing White Paper

Item Writing for Performance Tests

Wallace Judd, PhD

Performance Test Development Series 05

Item Writing for Performance Tests

Overview

Multiple-Choice versus Performance Test Development Phases

Many authors are familiar with writing Multiple Choice (MC) items, and are new to the process of writing performance items. So it may be helpful to articulate the differences in the Test Development Process. An overview of the differences appears in Table 1.

Because the types of items you are authoring vary quite a bit by whether the items are multiple-choice or performance items, this section is broken up into these two dimensions.

Multiple Choice

A multiple-choice (MC) item comprises two mandatory parts. The *stem*, which holds the information necessary for responding with a choice, and the options, which are the choices the candidate can select from.

The stem may include text as well as figures such as graphs, charts, and tables.

The options can be binary (as in True or False) or can include three or more choices. When the number of options exceeds five, we may call these list options.

Additional presentation styles for MC items include matching lists, hotspots, and select two or more.

Performance Testing

Performance testing asks the candidate to accomplish a task or series of tasks.

The candidate may or may not be provided with optional hints or a help system.

Topic selection is critical in performance testing, since the arena of performance is potentially so broad. Items used should be paradigms of a category of items. In Excel, for example, if you wanted to test performance on a function that calculates parameters for a range of cells, you would want to select SUM() or AVERAGE() rather than, say, MODE().

In many ways, writing performance items is easier than writing MC items, since you don't have to write credible distractors. On the other hand, quality assurance is much more difficult, since you have to assure

Topic	Multiple-Choice Test	Performance Test
Task Analysis	Knowledge, skills, understand, aptitudes	Paradigm tasks accomplished
Test Construction	Many items	Small number of items
Pilot Sample Size	Large	Small
Item Selection Filter	Item Parameters	Scoring, Administration
Item Pool Size:	Large	Small
Item Parameters	Homogeneous	Heterogeneous
Item Count	Numerous	Few
Administration	Standardized	Idiosyncratic
Quality Assurance	Trivial	Extensive
Archive	Items, item choices	Items, response log

Table 1. Overview of differences between multiple-choice and performance test items.

that all potential correct answers are accommodated.

Phases of Test Development

Item writing begins with the definition of skills in the blueprint developed by the JTA. Typically, in a performance test, several item sketches will be developed and tried out with five or six candidates. This is the alpha test. Beta testing requires a sample of candidates representative of the target audience, in large enough numbers to provide psychometrically viable numbers. If beta testing is not feasible, there are ways to use an initial administration to replace a beta. After the beta, Quality Assurance checks scoring, and finally there is rollout to the purchasing audience. Below are the primary phases of test development. Many of these are elaborated in the following sections.

Other white papers in this series will cover basic test statistics used to evaluate the quality of items, how to interpret them, and examples of calculations.

Concept, Initial Draft

Initial conceptual draft of the item. Instructions may be on paper. Set-up for the item done by an individual. Intent is to be able to see how the instructions should be refined and whether the item works, as it were.

Alpha Test

Closely monitored presentation of one or several draft items to subjects willing to articulate their thoughts as they take the items.

Revision

Revision of the instructions, scope of the item, set-up, and checking possibilities.

Comprehensive Development

Construction of all items necessary to fulfill the blueprint, along with a subset of item alternatives proposed as options for topics.

Code Score
Code that scores the item. This code must not only score all correct solutions to the item, but must also check for lethal side effects that could occur if the item is done correctly but has additionally induced detrimental results.

Beta Test
Administration of test items to a 30- to 50-person sample of candidates representing a distribution of skill levels to be expected in the target test-taking population.

Revision
Final revisions resulting from the beta test administration.

Test Assembly
Assembly of the test into a final form or multiple forms for administration as the final test.

Quality Assurance
Instructions to the QA team include:

- Multiple ways to do the task correctly
- Likely errors to be encountered
- Lethal side effects of likely errors

The QA team may not be comprised of subject matter experts, so the instructions are critcal.

Fatal issues found during QA must be fixed prior to rollout.

Rollout
Final test form or forms are released to candidates in the field.

Item Components
Following are the minimal components of any performance item.

Set-up
Enumeration of the objects that have to exist in the context for the item to be done successfully. Examples are specific configurations of programs, files that must exist, directory structures, or system settings.

Set-up Code
Code that not only verifies the appropriate initial conditions are set for the item, but also reset code that allows the item to be reset if the candidate makes errors that render the environment impossible to continue in.

Instructions
The text that the candidate will see describing the task. This should be unambiguous and work for all variations of operating systems and software releases. It may include instructions concerning the expected results of task completion.

Constraints
The limits to the functionality that is exposed to the examinee. For example, a candidate may not be able to delete the file templates which

are the source of items.

Answer & Draft Score

The likely correct answer and how the author expects the item to be scored.

The two most likely expectations are (1) a log file or list of parameters and (2) an action which will be successful if and only if the item was done correctly.

For example, if the task was to create a subdirectory with a specific name, check (1) would be to list the subdirectories and see if the name appeared; check (2) could be to save a file in the newly created subdirectory and if that action generated an error it is clear the subdirectory doesn't exist.

Control Structure

The item has to be integrated into the administration system that determines the sequence in which items are presented to the candidate. Code that enables restart, skip, next item exit to menu, and exit exam has to be enabled. Optional controls are hint, repeat instructions, mark for review, previous item.

Common Elements

Independent of the type of item you are writing, there are some common elements to all items. The following are common elements of all items.

Context

Creating a context that candidates can see occurring in their lives is definitely motivational. Candidates who can answer the item will feel that the test fairly evaluates their competence. Candidates who cannot answer the item should feel that this is something they should know, and hence will be motivated to learn the required knowledge or skills.

A brief sentence setting the stage for the item will often be sufficient for the candidate to recognize the context in which the item is being asked.

Complexity

High vocabulary and convoluted sentence structure are barriers to answering an item correctly that are not germane to the purpose of the test. In psychometric terms these are "construct-irrelevant" elements of an item. So use the lowest vocabulary that retains the purpose of the item.

Minimizing complexity does not mean that there should only be relevant information in the item. Sometimes part of the purpose of an item is to be able to separate irrelevant information from the information necessary to answer the item.

Clarity

Both the directions for the item and the acceptable response must be as clear as they can be made. Ambiguous or incomplete directions add difficulty to an item. Ambiguous responses that are partially correct obfuscate the appropriate response.

Checking
Evaluating the correctness of task completion. Checking includes evaluating deleterious side effects and unintended consequences of solving the task.

Scoring
Integration of the item checks into total test score.

Archive
Archive of the item as solved by the examinee. This is critical for review of checking and as a basis for evaluating appeals.

Six Item Types
It is helpful to recognize that all performance items are not the same, and require different underlying skills to successfully solve the tasks requested. Table 2 presents six performance item types that are unique and require different levels of skills.

Conceptualizing the Item
A performance-test blueprint typically lists tasks under each domain that the examinee should accomplish.

It is the item writer's task to bring that task to life, to create a context in which the examinee can imagine the task being requested in a real-world setting.

So the instructions should imply or reference a context for the ex-

Table 2. Six item types used in performance testing.

Type	Example	Characteristics
Memory	Crane Signals	The crane operator has to know the signals in order to perform the correct operation on the load.
Concepts	Tab in Word	The text author has to understand the concept of a tab in order to set it appropriately, and use or modify it.
Reference	In Excel: Look up parameters; Trigger; Net Present Value	Reference material may be available to the user, but the user has to be able to translate the information into action. The user also has to recognize when and where to look up specific information.
Competence	Multi-Step process: Rebuild a carburetor	The mechanic has to be able to assemble the parts in an appropriate sequence, then be able to test the functioning of the resultant assembly.
Kinesthetic	Bowling; Using a crane.	The bowling pins and alley are given; the bowler just has to have the kinesthetic skills to knock the pins over. Likewise, the crane operator test may be known – the challenge is to operate the crane effectively.
Problem Solving	Debug a program; "Fix it" in Red Hat	In a version of the Red Hat Certified Masters program, the candidate was given a non-functioning computer and asked to "Fix it." The test required a variable host of skills that could be used in any sequence.

aminee to understand the utility of having the skill. A brief sentence or phrase may establish the context.

Write the item from a user's perspective. What would the user need to do?

Topic: Compare binary files.

Initial draft: Compare these binary files V1 and V3 and list the output in /Dif .

Better: There are two code files, V1 and V3. Copy any elements of V1 that are not in V3 into V3 at the appropriate places.

Topic: Verify the integrity and availability of resources.

Initial Draft: Verify the integrity of the /stats.bin directory.

Better: The files in \etc\bin\were just unpacked. Identify any files were corrupted on unpacking.

Write the whole item – not just part of it.

Topic: Manage template user environment.

What are the parts of 'manage'?

Write an item that clearly requires all parts:

- Add a file to the user template.
- Delete 3 files from the user template.
- Add a directory \shadow to the user template.

Why would the user do that? When would it be necessary? How would it be useful?

Topic: Identify different types of files. .

Initial Draft: Which file type is this? HooDoo.xlsx

Better: Number all the Excel examples in the /Test directory alphabetically.

Phases of Item Construction - Checklist

Phase # Header

Author: Author's Name
Date: Date
Phase completed in this draft:
Draft | Review | Alpha | Review | Beta | Review | Final
ID: Create Item ID DDTTVV Text
DD = Domain
TT = Task # within Domain
VV = Version of this Task #
Text = Brief Item description
Task: Use Task Description in Tasks for Items
Goal: Purpose of the item

Phase # Instructions

Initial pass/draft of instructions.

Phase # Answer

Enter at least one way the item can be answered correctly.

Check that multiple correct answers are accepted and scored correctly.

Phase # Set-up

Instructions for the programmer to set up the system
Requirements for operating system image to accommodate item.
Configuration of the application software
Necessary data, files, logs, etc. to enable item.

Phase # Check

Proof of concept that item can be checked programmatically.
Draft script/macros to check for correct answers to item.
Draft script/macros to check for known incorrect answers to item.
Draft script/macros to check for deleterious side effects.

Phase # Control Structure Integration

Test that item is integrated into test item flow.
Check instructions: Skip, Next, Repeat instructions, Restart item Exit test; Exit to menu.
Optional controls: Hint, Previous, Mark for Review.

Phase # Final Instructions

Final edited text that will be shown to the candidate

Phase # Quality Assurance

QA check for side effects, unintentional errors.
Finalize code of checks that assure the answer has been answered correctly.

Performance Item Alpha Test Review Checklist

Alpha test is the initial, informal evaluation of an item. It can be done with several trial subjects who would presumably know how to do the task correctly. The goal is to see whether the item works, and what elements may not have been included in the initial item plan.

Alpha test review is usually conducted with one observer closely watching one or two subjects. The instructions can be text on paper. The set-up should be done by the observer prior to the subject starting the test. And the check referred to below is a listing of what must be done to verify that the task was completed successfully.

- ✓ **Context**
  Does the item have a clear Context that relates to the work environment?
- ✓ **Instructions**
  Are the instructions clear to the examinee?
- ✓ **Work**
  Does the item Work in the Operating System environment?
- ✓ **Answer**
  Have you supplied the correct Answer to the item?
- ✓ **Restart**

Does the restart button reset the item correctly and completely?

- ✓ **Check**
  Does the check accept multiple correct answers?
- ✓ **Check**
  Does the check score partial or completely wrong responses as incorrect?
- ✓ **Check**
  Does the check evaluate unintentional or detrimental side effects of an otherwise correct answer?
- ✓ **Independence**
  Is the item independent of other items?
- ✓ **Complete**
  Is it clear to the examinee when they've completed the item??

Quality Assurance

How much QA is required? … Ask Samsung. With their Note 7 fiasco.

If you're doing a process evaluation, you could QA forever and not exclude all possibilities.

NCCA: Standard 23: Quality Assurance

The certification program must have a quality-assurance program that provides consistent application and periodic review of policies and procedures.

> 7. Suggested evidence may include quality-assurance policies, meeting minutes, calendars or schedules, and training materials/logs.

ASTM E2849

The standard gives the following guidance.

> 5.2.5.1 QA shall be conducted to minimize the chance that any candidate will encounter a response path that has not been adequately submitted to QA.
>
> 5.2.5.2 All reasonable successful trajectories through the solution space as defined by an SME panel shall be tested, except that any logically isomorphic paths need not be executed during QA.
>
> 5.2.5.3 The set of unsuccessful trajectories tried by a minimum of 5% of beta test candidates shall be included in the QA trials.

If you've got that much QA, then you can reasonably conduct a recall without penalty.

ISO/IEC 17024

> 6.1 Audit Trail:
>
> 6.1.1 For every examination, the test shall record sufficient data to reconstruct sufficient parameters of the end state of any item to allow assessment of its correctness and reconstruct any intermediate results that are scored.
>
> 6.1.2 For any examination for which the process is evaluated, the test shall record sufficient information to document the process each candidate followed in the performance of the item.

3349 Monroe Avenue, Suite 336
Rochester, N.Y. 14618-5513
www.AuthenticTesting.com
+1-703-777-7321