TVAP Online

Test Validation & Analysis Program

Version 1.1

May 2014

Biddle Consulting Group, Inc.
193 Blue Ravine, Suite 270
Folsom, CA 95630 / 916.294.4250 ext. 113
www.biddle.com

Contents

Program Setup and Overview

System Requirements and Setup Instructions

Before installing the Test Validation & Analysis Program, be sure that your computer is equipped with at least the following:

- Active Internet connection
- Internet Explorer, Mozilla Firefox, or Google Chrome
- Microsoft Excel or Adobe Reader

Program Overview

This program is designed for use by human resources professionals to aid in **validating and analyzing written multiple-choice and/or true/false test items** (the program may also be used for other types of tests—please call for assistance). This program uses a *content validity* approach for validation (see Section 14C and 15C of the *Uniform Guidelines*, 1978) as it is most appropriate for validating tests designed to measure knowledge, skills, and abilities ("KSAs")[1]. Note that the *Uniform Guidelines* specify that tests measuring *abstract* traits or constructs that cannot be "operationally defined" in terms of observable aspects of job behavior (see Sections 14C[1] and Questions & Answer #75) should not be validated using a content validation strategy.

The program was developed by integrating concepts and requirements from professional standards, the *Uniform Guidelines* (1978), and relevant court cases. While efforts were made to automate these processes as much as possible, professional judgment should be used when operating this program and evaluating its results.

The program includes two separate portals, the **Administrative** and **Job Expert**. The **Administrative** portal allows human resources professionals to either manually input or import the relevant knowledge, skills, and abilities, as well as the test content itself, into the system. They are also able to invite Job Experts via email to take the validation survey online and view the results. The **Job Expert** portal allows Job Experts[2] who have been selected by the human resources professionals to take the survey online and provide ratings to the knowledge, skills, and abilities, as well as the test items.

To use the two portals in a successful validation process, the user needs two documents in advance. The first is a **job analysis** document that includes both a list of important and/or critical job duties and a list of KSAs for the target position. The second is the **written test—**which should be developed based on the job analysis conducted by both subject-matter experts ("Job

[1] KSAs are also sometimes referred to as KSAPCs, which stands for knowledge, skills, abilities, and personal characteristics. Personal characteristics in that instance, while generally not as *concrete* as individual knowledge, skills, or abilities, are similar to abilities and should be operationally defined in terms of observable work behaviors and/or work products as required under Uniform Guidelines Section 14C[4].

[2] "Job Experts" are also referred to by some in the human resources field as "Subject Matter Experts" or SMEs.

Experts") and human resources staff. If these two documents are available, the test validation and analysis process can begin by using the two workbooks in the following manner.

The Job Expert portal includes the tools for validating a test before it is administered to applicants. This portal includes a survey that is given to a panel of Job Experts online. The Job Experts use the survey to rate each question contained in the written test. These ratings are then automatically analyzed in the Administrative portal where the results are provided.

In summary, the Administrative portal is used to set up an online validation survey and the Job Expert portal is used to collect KSA ratings and test content validity evidence as well as to set cutoff scores based on various factors.

This manual is designed to provide the user with instructions for operating both portals and interpreting their results. It should be noted that some background in statistics and test development/validation is needed to operate these workbooks and effectively analyze their results. There is a **Glossary** at the end of this manual that defines key terms.

There are three attachments in this manual:

- **Attachment A - Instructions for Completing the Test Item Survey** can be used as a guide for administering the Test Item Survey to Job Experts and facilitating a test item validation workshop.
- **Attachment B - Test Item Writing Guidelines** provides information for working with Job Experts and human resources staff to write effective test items.
- **Attachment C - Upward Rating Bias on Angoff Panels** provides a brief description of the tools and methods in the program for detecting (and correcting where appropriate) an upward rating bias that sometimes occurs when establishing cutoff scores for tests.

Historical/Legal Background of the Program

Written tests have been the focus of litigation for decades. Completing a thorough validation process, such as the one facilitated by the TVAP Program, offers two key benefits to the employer. First, it helps insure that the test used for selection or promotion is sufficiently related to the job and includes only test items that Job Experts have deemed fair and effective. Second, a validation process generates documentation that can be used as evidence should the test ever be challenged in an arbitration or civil rights litigation setting.

Some of the standards used in the TVAP Program were adopted from court cases where the criteria pertaining to test validation have been litigated. Two of these court cases are Contreras v. City of Los Angeles (656 F.2d 1267. 9th Cir. 1981) and U.S. v. South Carolina (434 US 1026, 1978).

In the Contreras v. City of Los Angeles case, a three-phase process was used to develop and validate an examination for an Auditor position. In the final validation phase, where the Job Experts were asked to identify a knowledge, skill, or ability that was measured by the test item, a "5 out of 7" rule (71%) was used to screen items for inclusion on the final test. After extensive litigation, the Ninth Circuit approved the validation process of constructing a written test using items that had been linked to the knowledge, skills, and abilities (KSAs) of a job analysis by at least five members of a seven-member Job Expert panel.

In U.S. v. South Carolina (1978), Job Experts were convened into ten-member panels and asked to provide certain judgments to evaluate whether each question on several tests (which included 19 subtests on a National Teacher Exam used in the state) involved subject matter that was a part of the curriculum at his or her teacher training institution, and therefore appropriate for testing. These review panels determined that between 63% and 98% of the items on the various tests were content valid and relevant for use in South Carolina. The U.S. Supreme Court endorsed this process as "sufficiently valid" to be upheld.

These cases provide guidelines for establishing minimum thresholds (71% and 63% respectively) for the levels of Job Expert endorsement necessary for screening test items for inclusion on a final test to be used for selection or promotion purposes. In either case, it is important to note that at least an "obvious majority" of the Job Experts is required to justify that the items are sufficiently related to the job to be selected for inclusion on the test.

Following the reasonable precedence established by these two cases, this program uses a *>65% job duty or KSA linkage* criteria for classifying items as "acceptable" for inclusion on a test. If less than 65% of the Job Experts link the item to an important/critical job duty or KSA, the program simply "red flags" the item for closer evaluation. It should be noted, however, that because this program function requires ***at least*** a 65% endorsement level for each item, the ***collective*** endorsement level for all of the items used for an actual test is likely to be much higher. This is because several of the items are likely to have much higher Job Expert endorsement levels.

Setting up a Test Validation and TVAP Online Tool Administrator Portal

As previously mentioned, a Job Analysis document and a written test are needed before beginning the test validation and analysis process. It is necessary that the Job Analysis used for this process consists of important and/or critical job duties and KSAs (knowledge, skills, and abilities). We strongly recommend that you sequentially number both the job duties and KSAs, so that, when loaded into the TVAP Online system, the same sequence of numbers will appear. The written test being validated should consist of multiple choice and/or true/false questions, and also be arranged in numerical order for Job Expert reference during the rating process. After these two documents have been compiled, the steps below can be completed using the TVAP Program.

I. Logging into TVAP Online

The TVAP Online administrator login can be accessed at:

https://www.onlinetestvalidation.com/Admin

Enter your username (typically your email address) and password to access the system. Upon logging in, you will see the "Client Home Page." Underneath the "Client" tab, there are three options: Home, User, and Validation.

- **Home:** The "Home" button will bring you back to the "Client Home Page."
- **User:** The "User" button allows you to add, edit, and inactivate current users in the administrative portal.
- **Validation:** The "Validation" button allows you to create a new test validation process or edit existing test validation content.

II. Creating and Editing Test Validation Processes

Step 1: Creating New Test Validations

The main function of TVAP Online is to create an online survey for Job Experts to complete in order to content validate a written assessment. To do so, a new validation process must be created in the TVAP Online system. From the "Home" screen, click the "Validation" button. At this screen, you can manage your existing validation processes. If no validation processes have been created, click the "+ Create Validation" button to begin creating a new validation process.

There are four (4) initial pieces of information that are required for creating a new test validation: Position Name, Test Name, Validation Test Type, and Use Full-Length Survey Form.

Position Name: Position Name refers to the job/position that the test references to select and/or certify candidates.

Test Name: Test Name refers to the name or type of test that is being validated. This could be a math test, reading comprehension test, etc.

Validation Test Type: The TVAP Online tool allows for the validation of three types of tests. The instructions given to the Job Experts vary slightly, according to the type of test that is chosen by the administrator.

- **Entry-Level**: This type of test measures an applicant's knowledge, skills, or abilities that are performed on the first day of the job (i.e., prior to any training a person might receive *after* being hired).
- **Promotional**: This type of test measures whether or not an applicant possesses the knowledge, skills, or abilities that are required to promote to the job level being tested.
- **Certification**: This type of test measures whether or not an applicant meets the minimum knowledge, skill, or ability criteria to obtain a certificate for the position (i.e., demonstrate competency in the certification area being tested).

Use Full-Length Survey Form: By selecting this option, all 14 of the validation survey questions are presented to the Job Experts for each test item (including survey questions related to test item-writing basics). Do not click this box if you intend for the Job Experts to be initially presented with only the five (5) most basic validation questions for each of the test questions. Note that the Job Experts who are given the shortened survey have the option to view the additional survey questions related to test item-writing basics.

Step 2: Managing KSAs

Section 15C[5] of the federal *Uniform Guidelines* requires a validation report to provide evidence that demonstrates that the selection procedure is a representative sample of a knowledge, skill, or ability (KSA) used as a part of a work behavior and necessary for that behavior. TVAP Online has the ability to have Job Experts rate the Importance and differentiation between Best Workers for each KSAs. The *Importance* rating refers to how useful or meaningful the KSA is to the job. Average and standard deviations for the Importance ratings must range between 1.00 and 5.00. Those KSAs that are rated high on the *Best Worker* rating are those that, when performed above the "bare minimum," distinguish the "best" performers from those who are "minimally proficient." Average and standard deviations of the Best Worker ratings must also range between 1.00 and 5.00.

You should only include KSAs that are measured by the test in the TVAP Online survey. If a job analysis has been previously completed, we strongly recommend that you use the same KSA numbers that are listed in the job analysis. Once again, we recommend that you *only* include the KSAs that are relevant to the test being validated.

IMPORTANT: If a job analysis has already been completed and the KSA Importance and Best Worker averages and standard deviations have been calculated during that process, the KSA rating portion of the survey will not be administered to the Job Experts during the validation process.

Add KSA: To manually add a KSA, click the "+ Add KSA" button. A pop-up window will appear and prompt you to enter a KSA number as well as the KSA itself. The KSA number is not the sequence number. A job analysis will often contain at least some KSAs that are not relevant to this particular test. Using the same number for the KSAs as on the

job analysis will help track them to the job analysis, if necessary. To reorder the sequence of KSAs, click the up and down arrows on the *Manage KSAs* page.

Import KSAs: To import a list of KSAs, click the "Import KSAs" button. A pop-up window will appear. The KSAs must be imported following our KSA import specifications. A template for importing KSAs can be downloaded by clicking the white question mark in the blue circle at the top right-hand corner of the pop-up window. There are two templates: (1) if importance and best worker ratings have not been established in advance of the importing of the KSAs, and (2) if the importance and best worker rating have been established in advance of the importing of the KSAs. An additional four columns are included once the importance and best worker rating have been established, either by importing that information into the system or when the TVAP Online program computes that data based on the responses from the Job Experts.

- **KSA Number**: The KSA number presented to the Job Experts will often not be in continuous sequence. This is because a job analysis will frequently contain KSAs that are not relevant to this particular test. By using the same KSA number as in the job analysis, you can easily reference KSAs presented during the TVAP program back to the job analysis document. To reorder the sequence of KSAs, click the up and down arrows on the Manage KSAs page.
- **KSA description**: The KSA description is the actual KSA itself.
- **Imp.Avg (only appears if you choose the w/ratings template for import)**: This is the average of the importance criteria ratings for this KSA. This value ranges between 1.00 - 5.00 and includes two decimal places.
- **Imp.SD (only appears if you choose the w/ratings template for import)**: This is the standard deviation of the importance ratings for each KSA. This value includes two decimal places.
- **Best.Avg (only appears if you choose the w/ratings template for import)**: This is the average of the best worker ratings for this KSA. This value ranges between 1.00 - 5.00 and includes two decimal places.
- **Best.SD (only appears if you choose the w/ratings template for import)**: This is the standard deviation of the best worker ratings for this KSA. This value includes two decimal places.

After saving the import file to a local drive, click the "Browse…" button in the Import KSA pop-up. Once the KSAs have been imported, you will be given a chance to review the KSAs before accepting the import.

IMPORTANT: When importing using the KSA template, commas (",") cannot be used in the description. This is because the import file is created in a .csv format and will read the comma as a delimiter. Please replace all commas in the description with a semi-colon or other form of punctuation. Once the file has been imported, you can add the commas by clicking on the pencil and notepad icon next to the KSA underneath the "Action" header.

IMPORTANT: If a job analysis has already been completed, you can also import the KSA Importance and Best Worker averages and standard deviations. If you do this, the

Job Experts will not be required to rate the KSAs and the program will instead use the averages and standard deviations that have been imported when computing the validation analyses.

IMPORTANT: Importing KSAs will overwrite any existing KSAs you may have manually entered.

Step 3: Managing Test Items

Once the KSAs have been entered into the TVAP Online tool, the next step is to either input/import the test items, or only display the item numbers in the survey. In certain cases, such as when there are test security concerns, it may not be prudent to input/import the actual test content into the TVAP Online tool. In these instances, you can still use the survey tool accompanied by a separate "hard" copy of the test.

Add New Test Item: To manually add a new test item, click the "+ Add New Test Item" button. A pop-up window will appear in which you can input the relevant information for the test item.

- **Item #**: This is the item number as it would appear on the test; this is <u>not</u> necessarily the sequence number that appears in the survey. This is because some test items might have been removed during the development phase of the process, causing the actual item numbers to not be sequential. By using the same test item number during the survey as contained on the hard copy of the test, you can easily reference each test item being validated to the hard copy of the test.

- **Number of alternatives for this item**: This refers to the number of choices a candidate is presented when answering an item. For example, "Fill in the blank" refers to only one (1) choice. For true/false items, you should select two (2) from the dropdown. For multiple-choice test items, choose the number of alternatives offered to the test taker for each test item. The system is designed to accept a *maximum* of six (6) choices or alternatives (labeled A through F in the TVAP program).

- **Is this item a job knowledge question?** According to the federal Uniform Guidelines, knowledge being measured is "that body of learned information which is used in and is a necessary prerequisite for observable aspects of work behavior of the job." (Please see Section 14C[4] for additional information about job knowledge specifications). Specifically, if the test question is measuring a person's *knowledge about something*, rather than their ability to perform a task or demonstrate a skill (such as performing mathematical computations or solving problems), you should indicate "Yes" for this survey option. There are a set of additional questions given to Job Experts during the survey that relate to job knowledge items.

- **Status**: Administrators are provided the option to activate/inactivate a test item on the survey. Inactive items are not presented to Job Experts during the survey. Please note that if test items are inactivated *while* Job Experts are responding to the TVAP survey, those Job Experts will likely need to exit and re-enter the TVAP survey.

- **Stem**: This is the question being asked for each test item.

- **Alternatives**: Alternatives are the choices from which a test taker selects when asked a test question. Based on the number of alternatives selected by the Administrator in the TVAP program, the appropriate number of text boxes will appear so that all of the alternatives can be entered. If you are going to show the test takers the actual test questions using the TVAP survey, you should enter the choices exactly as they will appear on the test. Once the choices have been input, indicate the *correct* answer by clicking the radio button next to the item marked "Key." There can only be *one correct answer* for each test item.

Import Test Items: To import a list of test items, click the "Import Test Item" button. A pop-up window will appear. The test items must be imported following our test item import specifications. A template can be downloaded by clicking the white question mark in the blue circle at the top right-hand corner of the pop-up window.

- **Item Number**: This is the item number as it would appear on the test; this is not necessarily the sequence number that appears in the survey, due to some test items having been removed during the development phase of the process. By using the same test item number during the survey as contained on the hard copy of the test, you can easily reference each test item being validated to the hard copy of the test. To reorder the sequence of test items, click the up and down arrows on the Manage Test Items page.

- **Item Stem**: This is the question being asked for each test item.

- **Number of Alternatives**: This refers to the number of choices a candidate is presented when answering an item. For example, "Fill in the blank" refers to only one (1) choice. For true/false items, you should select two (2) from the dropdown. For multiple-choice test items, choose the number of alternatives offered to the test taker for each test item. The system is designed to accept a *maximum* of six (6) choices or alternatives (labeled A through F in the TVAP program).

- **Alternatives A-F**: Based on the number of alternatives entered, type in the alternative from the test into the corresponding cell. Leave any cells that are not used blank. Enter the choices exactly as they would appear on the test.

- **Key**: Enter the alternative that will be designated as the correct answer. The acceptable values are the letters A - F.

- **Knowledge**: If the item is assessing job-related knowledge, type "Yes" (without the quotes) in this column. According to the federal Uniform Guidelines, knowledge being measured is "that body of learned information which is used in and is a necessary prerequisite for observable aspects of work behavior of the job." There are two (2) additional questions given during the survey that relate to job knowledge items. If the item is not assessing job-related knowledge, type "No" (without the quotes) in this column.

After saving the import file to a local drive, click the "Browse…" button in the Import Test Items pop-up. Once the test items have been imported you are given a chance to review the test items before accepting the import.

IMPORTANT: When importing using the Import Test Items template, commas (",") cannot be used in the description. The import file is saved as a .csv format and will read the comma as a delimiter. Please replace all commas in the description with a semi-colon or other form of punctuation. Once the file has been imported, you can add the commas by clicking on the pencil and notepad icon next to the item underneath the "Action" header.

IMPORTANT: Importing test items will overwrite any existing test items you may have manually entered.

Only Display Item Numbers: If you wish to *only* display the items numbers in a survey *without* the item stem or alternatives, click the check box next to "Only Display Item Number" above the test item box. You will be prompted to input the number of items contained in the test. This will create a list of items in the manage items box, but will not populate any of the test question information. You can click the pencil and notepad icon underneath the "Action" header to change the "Knowledge" type and "Status" of the item.

Step 4: Sending/Resending Invitations

Once the KSAs and test items have been entered into the TVAP Online tool, the next step is to email invitations to Job Experts to take the online test validation survey. The TVAP Online tool will manage email invitations to the survey and display a Job Expert's completion status.

Job Expert Selection Criteria: A panel of qualified Job Experts should be selected for the job analysis study. The following criteria are presented as guidelines for selecting the members of the panel. They should:

(1) collectively represent the demographics of the employee population (with respect to gender, age, race, years of experience, etc.). We recommend slightly over-sampling gender and ethnic groups to insure adequate representation in the job analysis process;

(2) be experienced and active in the position they represent (e.g., Job Experts should not be on probationary status or temporarily assigned to the position). While seasoned Job Experts will often have a good understanding of the position, it is also beneficial

to include some relatively inexperienced Job Experts to integrate the "newcomer's perspective." However, we typically suggest that Job Experts selected for the panel have at least one year of active job experience;

(3) represent the various "functional areas" and/or shifts of the position. Many positions have more than one division or "work area" or even different shifts, where job duties and KSAs may differ; and,

(4) consist of between 10 – 20% supervisors for a given position. For example, if a Job Expert panel contains seven to ten Job Experts, we suggest including one to two supervisors on the panel.

How many Job Experts should participate in the job analysis process? Some court cases have relied on as few as seven to ten Job Experts for providing judgments and ratings about job and test characteristics.

Inviting Job Experts: You can have a maximum of 50 Job Experts invited to a single test validation process. A Job Expert can *only* be assigned to validate each test once, but may be invited to validate a number of different tests for each position. For example, a Job Expert may be assigned to validate two tests for an accounting position: math and reading comprehension. However, that same Job Expert cannot be assigned to provide more than one set of ratings for the same math test or the same reading comprehension test. There are three methods for inviting Job Experts to participate in the test validation survey.

- **Manual input:** To manually invite a Job Expert, click the "+ Add Job Expert" button. You will be required to input the Job Expert's email address, name, and password of your choosing (optional). Otherwise, a random default password is automatically created, but can be changed if desired. Edits can be made by clicking on the pencil and notepad icon next to the Job Expert under the "Action" header.

- **Enter Email String:** You can quickly enter a list of email addresses into the invitation manager by clicking the "Enter Email String" button, using a semi-colon (";") to separate them. This is especially effective if you are copying an email list from another program, such as Microsoft Outlook. When entering an email string, the name of the Job Expert is not populated and a random password will be generated for each Job Expert. To edit the name or password, click on the pencil and notepad icon next to the Job Expert under the "Action" header.

- **Import Job Experts:** To import a list of Job Experts, click the "Import Job Expert" button. A pop-up window will appear. The Job Experts must be imported following our test item import specifications. A template can be downloaded by clicking the white question mark in the blue circle at the top right-hand corner of the pop-up window. All three columns must include a value.

    - **JE Name**: This is the Job Expert's name as it will appear in the survey.
    - **JE Email**: This is the email address of the Job Expert being invited to take the survey.

- **JE Password**: This is the password that the Job Expert will use to access the test validation survey.

After saving the import file to a local drive, click the "Browse…" button in the Import Test Items pop-up. Once the test items have been imported, a chance to review the test items before accepting the import will be presented.

IMPORTANT: When importing using the Import Job Expert template, commas (",") cannot be used in the description. The import file is a .csv and will read the comma as a delimiter. Please replace all commas in the description with a semi-colon or other form of punctuation. Once the file has been imported, commas can be added by clicking on the pencil and notepad icon next to the item underneath the "Action" header.

IMPORTANT: Importing Job Experts will overwrite any existing Job Experts that were manually entered.

Editing the Message for Job Experts: The TVAP Online tool allows the administrative user to customize the email invitation message to the Job Expert. A default message is provided on the Send/Resend Invitations page in blue text. To edit the message, click the edit button underneath the blue text. Once editing is completed, be sure to save the updated message.

Sending the Invitations: Once Job Expert emails have been input/imported, manually select each box under "Send/Resend" next the Job Expert's information for each Job Expert that will receive the survey. (All Job Experts can be selected at once by clicking the box next to "Send/Resend" in the grey header bar.) Once the appropriate Job Experts are selected, click the "Send Invitations" button on the bottom of the screen to send out the surveys.

Step 5: Exporting Survey Data

Once the validation surveys have been completed by the Job Experts, various summaries and detailed data can be exported by the administrator. All exports can be either exported to Excel or PDF. In order to access these files, Microsoft Excel or Adobe Acrobat Reader/Plugin is required.

Job Expert Demographic Summary: This export provides a frequency count of the Job Experts' ethnicity and gender as well as averages for years of experience, supervising, and training.

Job Expert Demographic Data: This export provides Job Expert responses for ethnicity and gender as well as *averages* for years of experience, supervising, and training for each individual Job Expert.

Individual Job Expert Data: This export provides Job Expert responses for all validation survey questions.

KSA Summary: This export provides averages and standard deviations for the KSA Importance and Best Worker Ratings. This export is selected for each KSA.

Test-Item Data: This export provides the average ratings of the Job Experts who responded to the surveys. Use this export if the non full-length survey form was used for data collection. (Use the Test Item TVAP export if the full-length survey form was used).

Angoff Data: This export provides Job Expert responses regarding the percentage of minimally qualified applicants that would be expected to answer each of the test items correctly. The results are provided for all active applicants across all active items.

Test Form: This export provides a printable copy of the test items that have been either manually input or imported.

KSA List: This export provides a printable copy of the KSAs that have been either manually input or imported.

Test Item TVAP: This export provides the average ratings of the Job Experts who responded to the surveys. This export is most appropriately used if the "Use Full-Length Survey Form" option was chosen.

III. Evaluating the Test Item Validity in the Test-Item Data Export

The Test-Item Data export contains the relevant information to determine whether or not a test item is valid for use on a particular test. In order for an item to be valid, a specific set of criteria must be met.

Column Descriptions

Test Item: This is the actual item number on the test. This is not necessarily the sequence number.

Test Item Status: "Active" indicates that this item was rated during the test validation process. "Inactive" indicates that the item was not rated during the test validation process.

Valid Responses: Indicates the number of Job Experts who have rated this item.

 # Red Flags: The number shown in this column represents the number of potential problem areas with each corresponding test item. Test items that are red-flagged should be *considered* for removal from the test, or at least revised and re-evaluated by Job Experts (some minor changes can be made to the test items without requiring a re-evaluation of the items by Job Experts; however, items that are substantially re-worked to address the problem areas identified by Job Experts should be re-evaluated by Job Experts). Because some survey criteria are more significant than others, professional judgment should be used during this process.

KSA: This column provides the KSA numbers that were linked to the item. A list of the KSAs can be downloaded from the KSA List export from the Export Survey Data screen.

First Day: Tests should measure an aspect of the targeted KSA that is needed before on-the-job training. This column provides the "Yes/No" ratio of the "Necessary on the First Day

of the Job" rating; a "Yes > 65%" criteria is used for determining the validity of this rating. This survey question is designed to address Section 14C(1) and 5F of the *Uniform Guidelines*. While the job analysis process may insure that only KSAs that are needed the first day on the job are selected for measurement on the written test, this survey question helps to also insure that the *specific aspects* of the KSA measured on the test are actually needed before on-the-job training.

Fair: This column shows the ratio of Job Experts who agree that the item is fair to all groups of applicants (based on race, gender, and age), and therefore free from unnecessary bias or culturally-loaded content. A "Yes > 86%" criteria is used for determining the validity of this rating.

Minimally Qualified %: This column shows the average Angoff rating (minimum competency rating) for the item. The Angoff rating is the percentage of minimally-qualified test takers who would be expected to answer each test item correctly. No "Red Flag" criteria are used for this rating; however, items that are rated very low (near a "chance score" for the item, which means the item is extremely difficult) or very high (e.g., >95%; which means that the item is likely to be extremely easy) should be closely evaluated. This survey question is designed to address Sections 5H, 14C(7), and 15C(7) of the *Uniform Guidelines*. See Attachment A for detailed instructions on how to gather these ratings from Job Experts.

Fundamental: This column provides the "Yes/No" ratio of the Job Expert (e.g., 67% indicates that two-thirds of the Job Expert answered "Yes" to the following survey question) ratings regarding whether or not the item has fundamental issues that are outside of generally-accepted test writing practices. *(A detailed explanation of the ratings associated with each of these considerations follows).*

- Is the question being asked clear and understandable?
- Does the question being asked provide sufficient information?
- Are the incorrect choices similar in difficulty?
- Are the incorrect choices similar and distinct?
- Are the incorrect choices incorrect, yet plausible?
- Are the incorrect choices similar in length to the correct answer?
- Are the incorrect choices correctly matching to the question being asked?
- Is the correct answer correct in all circumstances?
- Is the item free of providing clues to other items?
- Is the item free of unnecessary complexities?

Memorized (Included only for job knowledge items): Tests measuring job knowledge should only measure job knowledge areas that are needed in memory while a person is performing the job (rather than job knowledge areas that can be easily looked up while performing the job without a potential negative consequence). For example, sometimes job knowledge tests measure areas of job knowledge that are provided on a reading list that applicants are directed to study before taking the test. Tests measuring these

"directed" areas of job knowledge are acceptable, provided that the majority of Job Experts agree that the item is measuring a specific area of knowledge that is necessary to have in memory and cannot be looked up without some potential negative consequence on the job. This column provides the average "Memorized" rating (using a 0-2 scale). A >=1.0 criteria is used for this question.

Consequences (Included only for job knowledge items): This survey question is designed to flag items that measure only trivial aspects of job knowledge (i.e., those items which would have little or no consequence if the test taker did not know the correct response). This rating is useful for screening out items that may be linked to a job knowledge domain (survey question 9) that may be (globally) critical to job performance; however the specific aspect of the knowledge measured by the item is not critically important. This survey question is designed to address Section 14C(4) and Question & Answer #62 of the *Uniform Guidelines*. This column provides the average for the "Consequences" rating (using a 0-2 scale). A >=1.0 criteria is used for this rating.

Important Note

The data shown in the columns for some of the following categories will fundamentally differ depending on whether the *full-length* survey form or the *non full-length* (shorter) survey form of the TVAP Online survey is administered to the Job Experts. Specifically, if the *full-length* survey form of the TVAP Online survey is administered, the percentages in these columns will be based on the responses from all of the Job Experts. If the *non full-length* of the TVAP Online survey is used, the percentages will indicate only the percentage of the responses from those Job Experts who indicated "Yes" to the following survey question: "Are there any Fundamental issues with the question that would prevent it from being used for testing." For example, if ten Job Experts complete a TVAP survey for a single test using the *non full-length* of TVAP Online survey, and only two of those Job Experts select "Yes" when responding to the "Fundamental" survey item, and of those two only one indicates that a Distractor is not plausible, the percentage that will be shown in the Distractor Plausible column would be 50% (i.e., one Job Expert out of two; not 10%, signifying one Job Expert out of ten).

Clear (Please see the "Important Note" above for more details about how to interpret the data in this column): This column provides the "Yes/No" ratio of the Job Experts' (e.g., 67% indicates that two-thirds of the Job Experts answered "Yes" to the survey question) ratings regarding the quality and readability of the item stem (the part of the test item that asks the question) and the alternatives. A "Yes > 65%" criteria is used for this column.

Sufficient Info (Please see the "Important Note" above for more details about how to interpret the data in this column): This column provides the "Yes/No" ratio of the Job Experts' (e.g., 67% indicates that two-thirds of the Job Experts answered "Yes" to the survey question) ratings regarding whether or not the item stem (the part of the test item that asks the question) contains sufficient information for the test taker to answer to question correctly. A "Yes > 65%" criteria is used for this column.

Distractor Difficulty (Please see the "Important Note" above for more details about how to interpret the data in this column): This column provides the "Yes/No" ratio of the Job Experts' (e.g., 67% indicates that two-thirds of the Job Experts answered "Yes" to the survey question) ratings regarding whether or not the incorrect answers are similar to

one another in difficulty and to the correct response for someone who does not know the correct response. Distractors should be presented in a manner whereby no incorrect answer is *obviously* incorrect (i.e., the test taker should not know the response is incorrect based on "common sense" or knowledge of some knowledge area unrelated to the area being tested). A "Yes > 65%" criteria is used for this column.

Distractor Distinct (Please see the "Important Note" above for more details about how to interpret the data in this column): This column provides the "Yes/No" ratio of the Job Experts' (e.g., 67% indicates that two-thirds of the Job Experts answered "Yes" to the survey question) ratings regarding whether or not the incorrect answers are similar and distinct. Distractors should be distinct from one another (i.e., presented in a manner whereby the incorrect alternatives are not simply variations of the same response). A "Yes > 65%" criteria is used for this column.

Distractor Plausible (Please see the "Important Note" above for more details about how to interpret the data in this column): This column provides the "Yes/No" ratio of the Job Experts' (e.g., 67% indicates that two-thirds of the Job Experts answered "Yes" to the survey question) ratings regarding whether or not the incorrect answers are plausible responses. Distractors should be presented in a manner whereby all the incorrect alternatives are plausible, yet incorrect, answers to the question. A "Yes > 65%" criteria is used for this column.

Distractor Length (Please see the "Important Note" above for more details about how to interpret the data in this column): This column provides the "Yes/No" ratio of the Job Experts' (e.g., 67% indicates that two-thirds of the Job Experts answered "Yes" to the survey question) ratings regarding whether or not the incorrect alternatives are similar in length to the correct alternative. Distractors should be presented in a manner whereby all the alternatives are roughly the same length. This is to avoid the systematic hint that the longest alternative contains the most information and therefore is typically the correct answer by default. Likewise, this will also keep the longest or shortest alternative from consistently being the correct alternative throughout the test. A "Yes > 65%" criteria is used for this column.

Distractor Matching (Please see the "Important Note" above for more details about how to interpret the data in this column): This column provides the "Yes/No" ratio of the Job Experts' (e.g., 67% indicates that two-thirds of the Job Experts answered "Yes" to the survey question) ratings regarding whether or not the incorrect alternatives match the question. Distractors should be presented in a manner whereby all of the incorrect answers are relevant choices to the question and are not too easily eliminated. In addition, the Job Experts should answer "No" to this survey question if there are grammatical (or other) clues that can be used to eliminate distractors without the test taker actually knowing the correct answer to the question. A "Yes > 65%" criteria is used for this column.

Correct (Please see the "Important Note" above for more details about how to interpret the data in this column): This column provides the "Yes/No" ratio of the Job Experts' (e.g., 67% indicates that two-thirds of the Job Experts answered "Yes" to the survey question) rating that the key (correct alternative) is appropriate for this question in all circumstances. A "Yes > 65%" criteria is used for this column.

Clues (Please see the "Important Note" above for more details about how to interpret the data in this column): This column provides the "Yes/No" ratio of the Job Experts' (e.g., 67% indicates that two-thirds of the Job Experts answered "Yes" to the survey question) ratings regarding whether or not the stem (the part of the test item that asks the question) or the alternatives provide clues or hints to other questions. If an item provides clues or hints to another item, it reduces that item's relative difficulty. A "Yes > 65%" criteria is used for this column. *We note that care should be taken concerning the interpretation of the responses to this question whenever the test items being rated are part of an "item bank" of test items from which test items will be chosen for inclusion on one or more shorter versions of the test. This is because larger banks of test questions can sometimes legitimately contain items that might provide clues about how to answer other test items. The key is to use the information from this survey item to make certain that test items that might provide clues to other test items are not contained within the same shorter version of the test.*

Difficult/Complex (Please see the "Important Note" above for more details about how to interpret the data in this column): This column provides the "Yes/No" ratio of the Job Experts' (e.g., 67% indicates that two-thirds of the Job Experts answered "Yes" to the survey question) ratings regarding whether or not the item is unnecessarily complex. Items should be written so they do not contain extraneous information. Instead, they should contain only the information necessary for answering the item. A "Yes > 65%" criteria is used for this column.

Current Info (Only for job knowledge items): This column provides the ratio of Job Experts who believed that the item is "Based on Current Information." An item that fails to meet the >65% endorsement criteria used for this question may not be based on current job knowledge (or, even if the item is *technically* based on current information, the Job Experts are indicating that, with their current level of understanding of the job, it may not be *practically* based on current information). A "Yes > 65%" criteria is used for this column.

Level Needed (Only for job knowledge items): This column provides the "Yes/No" ratio regarding the level of difficulty of the item. A "Yes > 65%" criteria is used for this column. Test items designed to measure job knowledge should be written at a level of difficulty that is similar to how the job knowledge will be actually applied on the job. This survey question is designed to address Section 14C(4) and Question & Answer #79 of the *Uniform Guidelines*.

Valid for Use: This column indicates whether or not the item can appropriately be used for testing purposes. The notation in this column will either be a "Yes" or "No." In order to be valid for use, at least 86% of the survey respondents must have reported the item as being "Fair", at least 65% of the respondents must have linked the item to a KSA with an "Importance" rating greater or equal to 3.0, and at least 65% of the respondents must have responded "Yes" to "Needed on the first day on the job" rating.

Valid for Job Knowledge Test: This column indicates whether or not the item can be used for a job knowledge test. The notation in this column will either be a "Yes" or "No." In order to be valid for use for measuring job knowledge, the item must first be "Valid for

Use" (see above). Also, the average for the "Memorized" rating must be greater or equal to 1 and the average for the "Memorized" rating must be greater or equal to 1.

Type of Use: This column refers to whether or not scores from this item can be ranked or only used as a Pass/Fail criterion. In order for an item to be used for ranking, at least 50% of the Job Experts must link the item to a KSA that has a "Best Worker" rating of at least 3.0. If this condition is not met, then the item can only be used as a Pass/Fail criterion. This addresses Section 14C[9] of the Uniform Guidelines, which specifies the requirements for using test scores above a minimum level to rank-order job applicants.

Validity Criteria Standards

A greater-than-65% criteria is used for determining the appropriateness of the use of each test item for the majority of the survey questions for at least three reasons. First, this criteria level has been previously endorsed in two high-profile court cases that involved written tests where Job Expert judgments on job-relatedness were evaluated. Second, the criteria represent a "clear majority." Third, 65% is a "natural break" that works well for Job Expert panels of various sizes. For example, in a three to five-member Job Expert panel, the 65% criteria allows for one dissenting Job Expert; in a six to eight-member panel, two can dissent; in a nine to eleven-member panel, three can dissent; and in a 12-member panel, four can dissent.

Please note that *conservative standards* have been used for the "Accept" and "Reject" values produced by the program. The type of test, nature of the position, and the extent to which other tests are used should be some of the factors considered when using the summary data from this sheet.

Glossary

Angoff Ratings (pronounced ANG-off)—Ratings that are provided by Job Expert on the percentage of minimally-qualified applicants they expect to answer the test item correctly. These ratings are averaged into a score called the "unmodified Angoff score" (also referred to as a "Critical Score").

Critical Score—The score level of the test that was set by averaging the Angoff ratings that are provided by Job Experts on the percentage of minimally-qualified applicants they expect to answer the test items correctly.

Cutoff Score—The final pass/fail score set for the test (set by reducing the Critical Score by 1, 2, or 3 CSEMs).

Job Analysis—A document created by surveying Job Experts that includes *job duties* (with relevant ratings such as frequency, importance, and performance differentiating), *KSAs* (with ratings such as frequency, importance, performance differentiating, and duty linkages), and *other relevant information* about the job (such as supervisory characteristics, licensing and certification requirements, etc.).

Job Duties—Statements of "tasks" or "work behaviors" that describe discrete aspects of work performance. Job duties typically start with an action word (e.g., drive, collate, complete, analyze, etc.) and include relevant "work products" or outcomes.

Job Expert—A job incumbent who has been selected to provide input on the job analysis or test validation process. Job Experts should have at least one year on-the-job experience and not be on probationary, temporary, or "light/modified duty" status. Supervisors and trainers can also serve as Job Experts, provided that they know how to perform the target job.

KSAs—Knowledge, skills, and abilities. Job knowledge refers to bodies of information applied directly to the performance of a work function; skills refer to an observable competence to perform a learned psychomotor act (e.g., keyboarding is a skill because it can be observed and requires a learned process to perform); abilities refer to a present competence to perform an observable behavior or a behavior which results in an observable product (see the *Uniform Guidelines*, Definitions).

Outlier (pronounced OUT-ly-er)—A statistical term used to define a rating, score, or some other measure that is outside the normal range of other similar ratings or scores. Several techniques are available for identifying outliers.

References

American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Angoff, W.H. (1971). Scales, norms, and equivalent scores. In Thorndike, R.L., *Educational Measurement*, pp. 508-600. Washington, DC: American Council on Education.

Biddle, D. (2011). *Adverse impact and test validation: A practitioner's guide to valid and defensible employment testing* (3rd ed.). Burlington, VT: Gower.

Contreras v. City of Los Angeles, 656 F.2d 1267 (9th Cir. 1981).

Dorans, N.J. & Holland, P.W. DIF detection and description: Mantel Haenszel and standardization. In Holland P.W, Wainer, H. (Eds.). *Differential Item Functioning*. Hillsdale, NJ: Lawrence Erlbaum Associates; 1993, 35-66.

Lancaster (1961). Significance tests in discrete distributions. *Journal of the American Statistical Association, 56*, 223-234.

Nunnally, J.C. & Bernstein, I.R. (1994). *Psychometric theory* (3rd ed.). New York: McGraw Hill.

Peng, C.J. & Subkoviak, M. (1980). A note on Huynh's normal approximation procedure for estimating criterion-referenced reliability. *Journal of Educational Measurement*, *17* (4), 359-368.

SIOP (Society for Industrial and Organizational Psychology, Inc.) (1987, 2003), Principles for the Validation and Use of Personnel Selection Procedures (3rd and 4th eds). College Park, MD: SIOP.

Subkoviak, M. (1988). A practitioner's guide to computation and interpretation of reliability indices for mastery tests. *Journal of Educational Measurement*, *25* (1), 47-55.

Tabachnick, B. G., & Fidell, L. S. (1996). *Using multivariate statistics* (3rd ed.). New York: Harper Collins.

Thorndike, R. L. (1951). Reliability. In E. F. Lindquist (Ed.). *Educational measurement* (pp. 560-620). Washington, DC: American Council on Education.

U.S. Department of Labor: employment and training administration (2000). *Testing and assessment: An employer's guide to good practices.*

U.S. v. South Carolina, 434 US 1026 (1978).

Zieky, M. (1993). Practical questions in the use of DIF statistics in item development. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 337–364). Hillsdale, NJ: Lawrence Erlbaum.

Attachment A - Administering the Test Item Survey

Introduction

This document provides instructions on how to use the Online Test Validation Survey. To access the Test Validation Survey, go to:

https://www.onlinetestvalidation.com/Survey/Webforms/Login/Login.aspx

To log in, enter your email address and the password that was sent to you via email.

Administering the Online Test Item Survey

Upon logging in, the Job Experts will be taken to the "Home Page". This page will show which test validation surveys they have been assigned to complete. To begin the test validation process, the Job Experts will click the "Start Rating" link if they have not accessed the survey before, or the "Edit Rating" link if they have previously completed the ratings for a particular survey and are making revisions.

Job Experts who are entering the survey for the first time will then be presented with the demographics survey. Upon completion, the Job Experts will be taken to the validation portal where they will be able to see the number of KSA and Test Item ratings they need to provide. (Please note that if the Administrator has imported KSA ratings into the system before the Job Experts begin the validation survey process, the Job Experts will not be asked to rate the KSAs again.) This portal will also display the number of ratings they have completed as well as the number pending. The Job Experts can also revise their demographic information by clicking the "Update Demographic" button at the bottom of the page.

Once the demographic portion of the survey is completed, and if the Administrator has not imported KSA ratings into the system, the Job Experts will be asked to complete the KSA section of the survey. If the KSA Importance and Best Worker ratings were imported by the Administrator in the administrative section, the Job Experts will only be asked to begin the survey. Once the KSA section is complete, the Job Experts will then proceed to fill out the item validation survey responses.

At any time, the Job Experts may log out by clicking the "Stop" button at the bottom of the page, and then clicking the "Log Out" link on the upper right portion of the screen. Any changes to the page before hitting "Stop" will not be saved. Upon logging back in, the system will show the number of ratings that are still pending. By clicking on the "Start Rating" link, the survey will continue from where it was stopped.

Security Hint! Upon completion of the validation survey, the Job Experts may log back in at any time to edit their responses. This means that if the test items were imported into the system, anyone using a legitimate username and password can view those test items. Therefore, once the validation survey process for a particular test has been completed, it is strongly recommended that the administrator change Job Expert passwords so that results cannot be changed and the test items can no longer be viewed.

Attachment B - Test Item Writing Guidelines

Test Item Writing Guidelines

Before writing any items, you should:

- Read all of the related materials about the job (job description, job analysis, test plan, etc.).

Understand both the job and the source materials (e.g., textbooks, standard operating procedures, training documents, etc.) on which the test will be based.

Use only the most recent publication/edition of the source materials when developing the test items.

1. If appropriate, make sure the potential test takers are told on which publication(s) (including the edition date) the test items are based.
2. Do not use a source from which to develop test questions simply because it is convenient. You should be using sources that are the most appropriate for the information being tested.
3. Examine and review alternative sources to see if those sources might be a better fit.
4. Do not use ambiguous and/or contradictory sources.

- Design your test based upon a written test plan.

1. Develop a written test plan. The test plan should be your road map in designing your test. *(Additional information about the test plan is offered later in this guide).*
2. Develop a content-by-process matrix to determine what kinds of items to write. *(This will also be covered later in this guide)*

- Select the number of alternatives you wish to include for each item before starting to write the test.

1. Research has shown that there is relatively little advantage to having more than three alternatives for any one multiple-choice question. However, using four is also very acceptable. Using more than four alternatives is generally impractical since it is extremely difficult to develop that many plausible, incorrect alternatives on a consistent basis.
2. Attempt to use the same number of alternatives for every question during each section of the test since this will help with calculating the reliability of the test after it has been administered.
    a. Note, however, that it is helpful, but not absolutely necessary, that all items on the test contain the same number of alternatives
3. Attempt to only use test items that contain a minimum of three alternative choices for each test question.

a. Research shows that two-alternative test items (such as “True” or “False” items) are much easier to answer correctly than test items that contain three or four alternatives. This is because if a test taker does not know the correct answer for a two-alternative test item, they have a 50% chance at guessing the correct answer (whereas they only have a 25% chance of guessing the correct answer to a well-written four-alternative test item).

b. Research also shows that the reliability of tests containing many two-alternative test items will likely be lower than tests that contain test items with three or four alternative choices for each item

Item Writing

The Basics

- Use correct spelling.
    1. Check the spelling of every word in the test document, including the instructions and introductory materials.
    2. Check the spelling again just before the final version of the test is offered or printed. Sometimes there are errors that are made during the final editing process that should be double-checked before a test is printed.
    3. Have someone other than yourself check the completed test for spelling, punctuation, or logic errors. Do not rely solely on your computer word-processing program's "spell check" function - it has serious limitations such as not catching the word "form" being spelled as "from.”
- Use correct grammar.
    1. Do not rely on your computer’s grammar checking function. Have your grammar checked by a qualified proofreader before releasing the test to be printed.
    2. Use all punctuation correctly, including commas, hyphens, question marks, colons, and semi-colons. Questions that contain incorrect punctuation can be misleading, confusing, or ambiguous.
    3. Do not use contractions in your test items. For example, use the phrase "do not" instead of the contraction "don't.”
    4. Have another person (other than yourself) check the completed test for grammatical errors before it is sent for printing. Do not allow any test to be printed or administered without it being double-checked for proper grammar and usage.
- Avoid negatively worded items, if possible. (In other words, do not say, “Which of the following does NOT belong?”)
    1. Studies show that these types of questions are unnecessarily difficult for some test takers, which can potentially reduce the reliability of the test.
    2. An exception would be when an employee on the job would be confronted with several choices and they need to be able to identify the choice that is not appropriate. For example, a question about fire safety might ask, “Which of the

following should NOT be used to help put out an electrical fire." The alternatives might include "sand," "baking soda", or "water." The alternative that should not be used for fighting electrical fires is water because water conducts electricity. In this instance, asking about what is not appropriate is acceptable since that is similar to what might actually occur on the job.

Check for typographical errors and make any necessary corrections.

1. During your final check, make certain that each question contains one of each lettered alternatives – that is, one "A," one "B," one "C," and one "D" (if using a four-alternative test item).

Note: If you are using Microsoft Word, it is a good idea to turn OFF the automatic numbering/lettering function when you are writing or editing multiple-choice tests since that function will sometimes overwrite entries you have made. To do this in Word 2003, go to the **Tools** menu, click **AutoCorrect Options**, and then click on the **AutoFormat As You Type** tab. Under **Apply as you type**, select or clear the **Automatic bulleted lists** or **Automatic numbered lists** check box.

- Capitalize all limiting or directive words.[3]
    1. For example, capitalize the word NOT and other absolute negative words.
    2. Also, capitalize key limiting and directive words that are vital for guiding the test taker to choosing the correct alternative such as REQUIRED, PROHIBITED, ALLOWED, ALL, MUST, EXCEPT, MOST, MIMIMUM, MAXIMUM, etc.
- Avoid using slang, ambiguous, and obsolete (or archaic) words.
- Break up long sentences into shorter sentences.
    1. Avoid run-on sentences that are difficult to follow.
    2. Keep the verb and subject word in the question close together to avoid ambiguity. For example, do not say: "In California, a person has committed the offense of simple battery, a gross misdemeanor, if he/she has…" Instead you could say: "A person in California has committed the gross misdemeanor offense of simple battery if he/she has…"
- Make sure all alternatives are parallel and similar in form. For example, make sure they are all written in the same tense and, if possible, have similar length and complexity.
- Use acronyms correctly. If the source material uses an acronym to describe something, the test-writer should use both the acronym and the full title. For example, an item should read: "The Central Intelligence Agency (CIA) is a part of the United States government."

[3] Some test writers prefer to underline directive words instead of making them ALL capitals. That is acceptable also. However, no matter what approach you use, you should use that same approach throughout the entire test.

1. Of course, this would not apply if you are appropriately attempting to measure a person's knowledge of what the acronym means.

Setting an item in an applied situation, when appropriate, can enhance the effectiveness of that item. In other words, it might help make a test be more job-related by wording the test items so that the test taker can readily see the relationship between the test item and the job.

1. Research indicates that test takers generally view tests that appear to be job related as being more fair.

THE STEM OF THE ITEM

The part of a test item that leads up to (but does not include) the choice of answers is called the stem of the item.

- Use incomplete statements as stems whenever possible *instead of* fill-in-the-blanks. "The capital of Pennsylvania is_____" is preferred over "_____ is the capital of Pennsylvania."

  1. If using blanks, try to put the blank as close as possible to the end of the last sentence in the stem.
  2. We recommend that you avoid using double alternatives. For example, it is better not to write an item that states: "The two longest rivers in the United States are _____ and _____." These types of items can be overly complex and difficult to score, and it is possible that the test taker could accurately guess at the correct response based on knowing only one of the two pieces of required information."

- Include all qualifying information in the stem. In other words, there should NOT be new information provided in the alternatives that should have been provided in the stem.
- Avoid using "ALL of the above," "NONE of the above," or "ALL of the above… EXCEPT" in the item stem. Research indicates that these types of items are often inappropriate for most jobs.

  Note: It is not a fatal flaw to have a few items of this type in a test. However, the overall number of these types of items should be limited whenever possible.
- Eliminate irrelevant words and ideas from the stem and alternatives. Be clear and concise.
- Use active voice rather than passive voice. Please note that most computer word-processing programs have a grammar-checking program that identifies passive voice, but these are not always reliable or accurate.

  For example: The staff is required to watch a safety video every year. (active voice) A safety video will be watched by the staff every year. (passive voice)
- Write the test-item's stem and alternatives so as to not provide grammatical clues to the correct answer.

For example, when an incomplete statement is used as a stem, use "a/an" if some of the alternatives in that item begin with vowels and some begin with consonants. (You do not want to give a clue to the correct answer by which you use.)

- State only one idea or central problem in each item, if possible.
- Include as much of the item in the stem as possible.
- If you are asking questions concerning job knowledge, you should use objective ideas that can be referenced back to appropriate source material. Unless you are trying to determine what a person's opinion is (which is generally not acceptable in a testing situation), do not ask for personal opinions or refer to the test-takers' emotions (e.g., wording such as "Indicate what you feel is the correct answer" would not be acceptable).
- Try to avoid using absolute terms such as "ALWAYS" and "NEVER" when asking questions, unless using those terms is related to the job being tested.
- Keep items as simple as possible. If the stem must be read more than once to be understood, you should rewrite (simplify, clarify) the item.
- To enhance the test taker's perception of fairness, we recommend that a similar number of men/males and women/females be mentioned when asking various test questions (when appropriate).

THE ALTERNATIVES OF AN ITEM

The choices for answers given to the test taker are called alternatives. Incorrect alternatives are called distractors. The correct choice for each test item is called the correct alternative.

- Have only ONE correct alternative for each item.

  1. Eliminate overlapping alternatives when appropriate. For example, do **not** use the following **unless such overlaps are commonly encountered on the job**:

     a) 5 to 10 days.

     b) 5 to 20 days.

     c) 15 to 25 days.

     d) 20 to 25 days.

  2. Use only plausible distractors. Do not try to be funny or cute. If you become stuck, work on another item and come back to the original item later on. Many item writers think of effective distractors for previous items when working on other items.

- Place periods at the end of each alternative when the item stem ends in an incomplete statement. Do not place periods at the end of each alternative where the stem either contained a blank space for the correct alternative or ended with a question mark.
- Numbers.

  1. Generally, it is best to write out the numbers one to ten and to use the numeric form for numbers over ten (e.g., 11, 12, 27, 53). However, if you have an item where the alternatives are both less than ten and greater than ten, use the numeric form for all alternatives in that item.

2. Use ascending order for alternatives containing numbers. For example, the alternatives might be listed as: 1, 4, 8, 10; or 25, 30, 35, 40. In other words, the first alternative should contain the lowest number, the second alternative should contain the next higher number, and so on.

   Note: This would not be appropriate if the order gives away the correct response.

3. Be consistent when using units of measurement. Do not mix and match units within a single item, unless such mixing of units or metrics is encountered on the job. For example, if you use minutes in one alternative, it is typically best to use minutes for all of the alternatives in that item unless mixed units or metrics are encountered on the job.

4. Use exclusionary terms such as "What is the MAXIMUM number of days" to make numeric questions unambiguous.

- Do not use both categories and sub-categories as your alternatives. If you do, then test takers who know only the headings of sections in the reference source will be able to answer the item on this basis alone. For example, you would not want to use the alternatives of "vertebrate," "mammal," and "human" as choices in the same item (since "humans" are both "vertebrates" and "mammals").
- Be as clear and concise as possible. Eliminate any possible misinterpretations.
    1. Do not merely reword one of the distractors to create a new distractor.
    2. Do not use synonyms of the other distractors as an additional alternative.
    3. If you cannot come up with a sufficient number of distractors, then reword the item or eliminate that item.
- Each alternative should be independent from the other alternatives.
    1. Avoid using "ALL of the above, "NONE of the above," "A and C," or other such alternatives. These types of questions have been shown to be inappropriately complex for some test takers. (In other words, these types of questions are sometimes harder than the job itself, which is not acceptable).
    2. Also, some test-wise applicants can determine a pattern during testing where either "ALL of the above" or "NONE of the above" is always (or virtually always) the correct (or incorrect) response.

Do not make the correct alternative significantly different in form from the other alternatives.

1. Alternatives within the same item should be of similar length.
2. Alternatives within the same item should be parallel (e.g., same tense, form, structure, etc.).
3. All alternatives should be grammatically consistent with the stem.
4. Avoid using words that sound alike or look similar as alternatives.

GENERAL RULES

- Do not allow the stem and/or alternatives of one item help a test taker answer any other item within the same test.

1. If possible, avoid using identical alternatives for different test items within the same test since this may give the test taker a clue as to how to answer another item.
2. If you cannot avoid using identical alternatives for more than one item, list the alternatives in the same order for both items.
3. Items where test takers are asked to match a list of items to a list of alternatives is generally problematic since, as the test taker eliminates some responses, their chances of guessing the correct remaining alternative increases. (This is especially problematic if the number of items is the same as the number of potential alternatives).

- Make certain the complete stem and all alternatives associated with that stem are on the same page when the test is printed.
- Use gender-neutral and race/ethnic-neutral terms and pronouns in your items, unless this information is vital to the item or if that would enhance the job-relatedness of the item.
    1. If the item has several actors, try using gender-neutral names, such as Pat or Chris.
    2. You can sometimes use titles to eliminate specifying gender. For example, the terms "officer," "supervisor," and “manager” do not denote any particular gender.
    3. Be consistent. If you make Chris Johnson a female in one item, make sure Chris Johnson is a female in all of the items in that test.

Avoid using abbreviations during a test whenever possible.

1. Use only those abbreviations that are commonly understood by the test taker. If in doubt, do not use the abbreviated version.
2. Do not mix words and abbreviations. For example, if you use the abbreviation "ft" in one part of the item, do not use the word "feet" in another.
3. If you must use abbreviations, you should consider writing out what the abbreviation means the first time it is mentioned in the test (e.g., “How many miles (mi.) is it from New York to Boston?”

   Note: Of course, if you are asking the test taker to identify the abbreviation in a job-related fashion, or if the abbreviation is part of the knowledge that all qualified applicants should know prior to being hired or promoted, then using the abbreviation is acceptable.

- Use the time and date structure that the test takers will most easily understand.
    1. For example, many police and fire departments use military time (e.g., 1800 hours instead of 6:00 p.m.). However, entry-level test takers are unlikely to understand what military time means, therefore, use the traditional form of telling time (i.e., A.M. and P.M.) that is generally understood by the vast majority of test takers (unless the alternative time scheme must be known by test takers at the time of taking the test and their first day on the job or training.).

Randomize the order of alternatives and location of the correct answer.

1. It is best if the final version of the test has a similar number of As, Bs, Cs, and Ds as correct alternatives for a four-alternative test.
2. The pattern for the correct responses should be random. Make certain there is NO discernible pattern for the correct responses on the test.

OVERVIEW

- **Item Difficulty.** It is okay to make *some* items easy and some *items* difficult. However, you should try not to make too many test items either *too* easy or *too* difficult.
    1. Make the level of item difficulty appropriate to the job being tested.
- **Job-Related.** Make sure all of the test items are job-related (i.e., related to important or critical job duties). Do not become immersed in the reference source and forget to rely on the job analysis to determine what is important for the job.
    1. Determine that each item matches something in the job content according to the job description.
    2. Each item should tap into a relevant aspect of the job. Just because information is contained in a reference source does not automatically mean it should be tested. It is up to you, the item writer, to help insure that the information being tested is relevant.
    3. Make sure each item is applicable to the employer and/or location where you will be testing. For example, it would be inappropriate to ask items about railroad crossings on a promotional police officer test if there are no railroad tracks in the municipality for which you are testing.
- **Complexity**.
    1. Match the test items to the complexity of the job. Maintain the appropriate vocabulary and reading level.
    2. Check the reading level with a computerized word-processing program to make certain it is at or below the reading level required on the job (or during training). List the desired reading level in the test plan and the test reading level in the test development guide.
- **Reference Source.** When asking questions about specific knowledge areas, do not go outside of the reference source for the alternatives or rationale for answering the alternatives. For example, do not use a recent Supreme Court decision to justify your choice of a correct alternative if that Supreme Court decision is not mentioned in the reference source the test takers might have been instructed to study (unless the test takers were informed in advance of the test that they would also be responsible for knowledge about that Supreme Court decision).
- **Avoid providing clues to the correct answer**.
    1. Do not make the distractors significantly different from the correct answer in form.
    2. Make distractors equally plausible to the uninformed candidate.

a) Try to avoid using distractors that are direct opposites of the correct answer, since they are generally easy to guess as being incorrect.

b) Avoid using the same/similar words (or synonyms) in both the stem and the correct alternative. This provides a clue to the correct answer.

c) Avoid wording the correct alternative in more (or less) detail than the distractors.

d) Try to avoid using absolute terms in distractors. Words such as "ALL," "ALWAYS," "NONE," "NEVER," and "ONLY" are often associated with incorrect responses to test questions.

e) Avoid creating a sub-set of alternatives that is all-inclusive. If two of the alternatives cover all possibilities, the other distractors are easily eliminated from consideration as being correct by the test-wise candidate.

f) Avoid writing two or more distractors that mean the same thing. The test-wise candidate can easily eliminate these alternatives.

- **Have only <u>one correct alternative</u> for each test item.** Search for possible conflicts with other resource materials or other sections of the reference material from which you are working.

  1. You *can* ask a test taker to select the BEST alternative from a list of several alternatives as long as: (1) you clearly indicate you are asking for the BEST response (as opposed to asking for a "correct" response); (2) employees must typically choose the BEST alternative from several alternatives on the job; and (3) one or more of the other alternatives are not the BEST choice.

- **Process-by-content matrix.** Three types of items can be included in this matrix (see the "Item Examples" section later in this paper for examples of each type):

  1. Knowledge of terms/definitions,

  2. Knowledge of principles and concepts; and/or,

  3. Application of principles and concepts.

- **Know when to admit defeat.** Do not be afraid to throw out or set aside items that cannot be saved. It is better to discard or set aside a bad item than to include it in the test or spend too much time working on it.

Situational Questions

Test takers are generally more likely to accept the testing process as being fair and valid when there is a transparent relationship between the content and/or context of the test and the content and/or context of the job. For this reason, we encourage item writers to include questions that include contextual aspects of the job, whenever possible. For example, if you are hiring a building contractor, you *could* simply ask:

- **(45 – 15)/2 =**
    - **A. 10**
    - **B. 15**
    - **C. 30**
    - **D. 45**

Or, in order to word a question in a way that reflects how information is actually used on the job, you might ask:

- **You must use a crane to lift a pipe that is forty-five (45) feet long. In order to pick the pipe up and keep it horizontal, you are going to lift it at the center of the pipe using a fifteen (15)-foot spreader. How far should the two parts of the spreader be placed from EACH end of the pipe?**
    - **A. 10 feet**
    - **B. 15 feet**
    - **C. 30 feet**
    - **D. 45 feet**

By framing the question in a job-related context or situation, you can require the test taker to recall and apply the facts he or she may know to a specific, job-related situation. This approach allows you to determine whether the test taker can apply job-related knowledge to plan how to safely and successfully lift heavy objects using a crane on the job.

Example of Item Format When Developing Tests

Item and Answer Documentation Example:

According to Harper's Criminal Investigation Workbook, testimony given by an accomplice or participant in a crime which tends to convict others is called _____ evidence.

A. state

B. police

C. defense

D. dissent

Correct Alternative: A

Source: Harper's Criminal Investigation Workbook (3rd Edition), p. 47.

Notes:

- *Develop three (3) or four (4) alternatives per item.*
- *Indicate the correct alternative so those reviewing the test items can make an informed decision about whether the "correct alternative" you have chosen is correct in all types of situations.*
- *Indicate the source from which the correct alternative was found, including page number or section (e.g., "Merit System Manual, page 33") to make it easier for those reviewing the test to verify the answer.*
- *Optional hint if many sources are used when developing the test: Place the name of the source within the text of the item stem in capital letters. (e.g., "According to the MERIT SYSTEM MANUAL,...")*
- *Avoid negatively-worded items. ("Which of the following does NOT belong?")*
- *Capitalize limiting, directional, or negative words.*

Randomize the order of distractors and the correct alternatives for the various test questions.

Examples of Various Types of Test Items

I. IF THE STEM IS A QUESTION

Example: What color are polar bears in the Arctic?

A. Black

B. Yellow

C. White

D. Brown

Notes:

- *The stem must be a grammatically complete sentence.*
- *Place a question mark at the end of the question.*
- *The question should be the last sentence in the stem.*
- *Capitalize the first letter of each alternative for this type of question, even if the alternative is not a proper noun.*

Do not put a period after each alternative for this type of question.

II. IF THE CORRECT ALTERNATIVE COMPLETES THE STEM:

Example: A person who burns his/her own car and reports it stolen in order to obtain the insurance money is guilty of

A. insurance fraud.

B. theft.

C. arson for profit.

D. burglary.

Notes:

- *Use no punctuation at the end of the stem. (Do not use a blank line or colon.)*
- *The sentence leading to the alternatives should be the last sentence in the stem.*
- *Place a period at the end of each alternative since the alternatives complete the sentence.*

Capitalize the first letter of each alternative only if it is grammatically correct to do so.

III. IF THE ALTERNATIVE IS EMBEDDED IN THE STEM:

A robbery is an example of a _____ offense.

A. traffic

B. criminal

C. civil

D. zoning

Notes:

- *Do not put a period or other punctuation at the end of these alternatives. If punctuation is required, place the punctuation in the stem.*
- *Capitalize the first letter of the alternative only if it is grammatically correct to do so.*

Test-Wide Hint: Make all lines indicating where the response fits the same length, regardless of the length of the alternatives.

Examples of Knowledge Types of Test Items

I. KNOWLEDGE OF DEFINITIONS/TERMS

The legal proceeding whereby one party to an action may be informed as to the facts known to other parties or witnesses is a

A. discovery.

B. garnishment.

C. indictment.

D. tort.

Note:

This item taps a test taker's knowledge of the definition of the word/term "discovery."

II. KNOWLEDGE OF PRINCIPLES AND CONCEPTS

The changing phases of the moon are caused by

A. the tilt of the earth's axis.

B. the rotation of the moon on its axis.

C. the tidal patterns of the earth's oceans.

D. the orbit of the moon around the earth.

Note:

The concept/principle being tapped here is the understanding that the cause of the changes in the phases of the moon is the orbit of the moon around the earth. These types of items include those that tap the test taker's knowledge of facts.

III. APPLICATION OF PRINCIPLES AND CONCEPTS

On hot sunny days, parked cars with the hottest interiors are those that are _____ in color.

A. white

B. red

C. black

D. green

Note:

This item taps into the application to parked cars of the principle (concept/fact) that black surfaces absorb heat (and light) more efficiently than other colors and thus will be relatively hotter.

Test Plan Example

The following is an example of a Test Plan outline for a test for the promotion of police officers to the position of police sergeant.

The length of the Police Sergeant exam will be 100 items. The number of items drawn from each source document will be based on the following:

A. The importance and frequency of behaviors associated with the knowledge contained in that source document.

B. The priority assigned to each source document by the advisory committee of job experts.

C. The nature of the content and the relative length of the source document. (Some documents are very short. For this reason, the number of items drawn from a document is partially determined by how feasible it is to write items from that source. In other words, some content areas are less amenable to testing than others.)

The reading level for this exam will be that of someone who has completed at least four years of high school (i.e., grade level 12).

Following the determination of the length of the test and the number of items to be derived from each source, a test plan is developed. The use of a process-by-content matrix ensures adequate sampling of job knowledge content areas and problem-solving processes. The tested areas in the sample process-by-content matrix involve the following

A. knowledge of terminology

B. understanding of principles

C. application of knowledge to new situations

While knowledge of terminology is important, the understanding and application of principles are considered to be of primary importance. This is reflected in the recommendation that a majority of items involve application of knowledge and understanding of principles. Furthermore, not all documents are equally well suited for each of the three problem-solving processes. Therefore, the manner in which each is sampled takes this into consideration.

PROPOSED PROCESS-BY-CONTENT MATRIX

For the Position of Police Sergent

	SOURCE	Knowledge of Terms and Definitions	Knowledge of Principles and Concepts	Application of Principles and Concepts	TOTAL
1	Community Policing	0	7	13	20
2	Rules of Evidence	2	6	12	20
3	Department Rules & Regulations	0	4	6	10
4	State Criminal Code	2	7	11	20
5	State Vehicle Code	2	4	14	20
6	City Ordinances	2	2	6	10
	Total	**8**	**30**	**62**	**100**

Note: The document above represents the overall goals the test writers set for themselves to follow. Turn to the next page to see the final Process-by-Content Matrix that shows what really occurred during the test's development. Notice how it is slightly different from the test plan.

FINAL PROCESS-BY-CONTENT MATRIX

For the Position of Police Sergent

	SOURCE	Knowledge of Terms and Definitions	Knowledge of Principles and Concepts	Application of Principles and Concepts	TOTAL
1	Community Policing	2	6	12	20
2	Rules of Evidence	2	6	12	20
3	Department Rules & Regulations	0	6	8	14
4	State Criminal Code	2	3	15	20
5	State Vehicle Code	2	4	14	20
6	City Ordinances	2	0	4	6
	Total	**10**	**25**	**65**	**100**

Note: This is an example of the Process-by-Content Matrix for a completely developed test. During the test's development, it was determined that not all source documents were equally well suited for each of the three problem-solving processes. While the item writers attempted to conform to the test plan, some sources did not yield the required number of items. So, the number of items actually developed from each source was different from the preliminary test plan in some respects. This can occur when the source materials do not supply sufficient information to write appropriate test items. Alternatively, when the item writers examined the source materials more closely, they identified additional areas in the source materials that should be covered during the test that were not identified when they originally wrote the proposed Process-by-Content Matrix. From this, you can see that test writing is a dynamic process where the reality of writing test items is frequently dictated by the quality of the source materials. If changes to the plan are made, you should make certain that you are measuring only knowledge, skills, or abilities that are associated with important or critical work behaviors.

Attachment C - Upward Rating Bias on Angoff Panels

In some situations, Job Expert panels responsible for setting the Critical Score level for a test set the bar too high. For example, we have experienced situations where only 50% of credentialed applicants in a given area of expertise taking a pre-employment test (measuring the same competency areas in which they are credentialed) would pass a recommended cutoff score set by the rater panel. In highly-regulated fields where credentialing programs are rigorous, a situation where a pre-employment test fails 50% of the credentialed applicants could have two possible explanations: (1) the credentialing program is setting the bar much too low (and unqualified candidates are being credentialed), or (2) the rating panel that established the Critical Score for the pre-employment test set the bar too high. While there is a range of other plausible explanations between these two extremes, it has been our experience that the latter explanation is more often the case.

This "upward bias" tendency does not (of course) rule out the opposite, where a rater panel underestimates the ideal minimum competency level. However, it has been our experience that rating biases of the overestimation type are more common than those of the underestimation nature.

While there are several viable theories that may explain why this phenomenon may occur with rating panels, one particular theory seems to provide a practical explanation. The *conscious competence* theory (which is sometimes also called the "Four Stages of Learning" theory) was originally posited by psychologist Abraham Maslow in the 1940s. This theory provides an explanation of how people learn in four *progressive* stages:

1. **Unconscious Incompetence** (where you don't know that you don't know something);
2. **Conscious Incompetence** (you are now aware that you are incompetent at something);
3. **Conscious Competence** (you develop a skill in that area but have to think about it); and,
4. **Unconscious Competence** (you are good at the skill and it now comes naturally).

These four "learning stages"—ranging from unconsciously unskilled, consciously unskilled, consciously skilled, to unconsciously skilled—have been widely adopted in both theory and practice in educational, psychology, and organizational behavior fields since their inception. It is the fourth stage (unconsciously skilled) that may cause some of the upward bias sometimes observed in rating panels. This is because individuals who have had so much practice with a particular skill—to the point where it becomes "second nature" to them and can be performed easily without intense concentration—can sometimes *underestimate* their level of competency when they first started in the position (i.e., causing them to incorrectly remember how long it took them to master the skill), which may cause them to *underestimate* the difficulty of the item for applicants who do not have the same job experience.

Common examples of skills that can be attained at this "fourth level" include driving, sports activities, typing, manual dexterity tasks, listening, and communicating. For example, performing a "Y turn" is second nature to most people who have been driving for several years. In fact, many experienced drivers may not even recall ever having to acquire this skill, but in actuality many experienced drivers had to work hard at this skill repeatedly until mastered.

This issue can create an upward bias when applying minimum passing score recommendations. Some raters might now be able to teach others in the target skill, although after some time of being unconsciously competent, the person might actually have difficulty in explaining exactly how they perform a particular skill because the skill has become largely instinctual. This arguably gives rise to the need for long-standing unconscious competence to be checked periodically against new standards.

Below are four suggestions that can be followed to help alleviate this potential problem:

1. Select Job Experts who have between 1 and 5 years of experience to serve on the rating panel;
2. Have the Job Experts reveal and discuss their ratings on the first few test items so the outliers in either direction can be reined in by the workshop proctor;
3. Conduct rigorous discussions with the Job Experts regarding the "true" minimum qualification level relevant to the test; and,
4. In some situations, masking the answer key from the Job Experts can help reduce the potential for upward rating bias.