Developing, Validating, and Analyzing Written Tests

**BCGi: Adverse Impact & Test Validation Book Series**

---

**Overview of Biddle Consulting Group, Inc.**

- **Affirmative Action Plan (AAP) Consulting and Fulfillment**
  - Thousands of AAPs developed each year
  - Audit and compliance assistance
  - AutoAAP™ Enterprise software

- **HR Assessments**
  - AutoGOJA™ online job analysis system
  - TVAP™ test validation & analysis program
  - CritiCall™ pre-employment testing for 911 operators
  - OPAC™ pre-employment testing for admin professionals
  - Video Situational Assessments (General and Nursing)

- **EEO Litigation Consulting / Expert Witness Services**
  - Over 200+ cases in EEO/AA (both plaintiff and defense)
  - Focus on disparate impact/validation cases

- **Compensation Analysis**
  - Proactive and litigation/enforcement pay equity studies
  - COMPare™ compensation analysis software

- **Publications/Books**
  - EEO Insight™: Leading EEO Compliance Journal
  - Adverse Impact (3rd ed.) / Compensation (1st ed.)

- **BCGi Institute for Workforce Development**
  - 4,000+ members
  - Free webinars, EEO resources/tools
  - Members only webinars, training and much more…

---

**Biddle Consulting Group Institute for Workforce Development (BCGi)**

- **BCGi Standard Membership (free)**
  - Online community
  - Monthly webinars on EEO compliance topics
  - EEO Insight Journal (e-copy)

- **BCGi Platinum Membership**
  - Fully interactive online community
  - Includes validation/compensation analysis books
  - EEO Tools including validation surveys and AI calculator
  - EEO Insight Journal (e-copy and hardcopy)
  - Members only webinars, training and much more…

---

www.BCGinstitute.org
Developing, Validating, and Analyzing Written Tests: Presentation Outline

• Overview
• Five Typical Steps for Developing Written Tests
• Seven Steps for Developing a Content Valid Job Knowledge Written Test
• Steps for Developing a Customized Personality Test Using a Concurrent Criterion-related Validity Strategy

Developing, Validating, and Analyzing Written Tests: Overview & Background

Contact Information

Stacy L. Bell, M.S.
Executive Vice President/FPSI
stacy@fpsi.com

Fire & Police Selection, Inc.
193 Blue Ravine, Ste. 270
Folsom, CA 95630
1-916-294-4242 x. 245
1-888-990-3473
www.fpsi.com
www.NationalFireSelect.com
About the Speaker

- Fire & Police Selection, Inc. (FPSI) is a sister-company of Biddle Consulting Group.
  - Over 80 Physical Ability Test installations at fire departments across the U.S.
  - Over 800 personnel, fire, and police clients across the U.S. and Canada
  - Entry-level written, physical, interview assessments, manipulative skills tests, cognitive ability tests, and personality tests available for selection of public safety personnel
  - Post-Academy tests
  - Fire Promotional tests and assessment exercises
  - Online applicant testing
  - Custom job analysis and validation services

Why is this Topic Important to HR/EEO Professionals?

- Why do I need to know about this topic?
  - Written tests are likely to exhibit adverse impact, which means they are among the selection tools most likely to be litigated
  - Strong written tests can be highly effective tools for reducing the applicant population and selecting qualified workers

- What are the key essentials I need to know about this topic?
  - Written tests that measure skills/abilities, job knowledge, and personality all have different development and validation steps
  - Customized, job-specific tests will generally have higher validity than "one-sized fits all" tests
  - Personality tests that are given with cognitive tests can enhance the validity of your testing process and reduce adverse impact
  - If you're audited and PPTs have AI, FEAs will go through the validity of your test
  - The employer's responsible—not the test publisher

- What are the consequences surrounding these issues?
  - Job-specific tests will maximize job performance of your workforce
  - The "start up" cost of a validation case is $30k to $80k

How are Written Tests Developed, Validated, and Analyzed?

- Written tests vary in type and purpose:
  - Job Knowledge
  - Ability
  - Personality
  - Entry-level
  - Promotion
  - Certifications

- Written tests need to be validated if they exhibit adverse impact:
  - Content or criterion-related validity
  - Numerous steps are involved…
Developing Skill/Ability Written Tests: Process Overview

- Step 1: Determine the KSAPCs to be Measured by the Test
- Step 2: Develop a Test Plan for Measuring the Selected KSAPCs
- Step 3: Develop the Test Content
- Step 4: Validate the Test
- Step 5: Score and Analyze the Test

Step 1: Determine the KSAPCs to be Measured by the Test

- Use a Selection Plan if possible
- Criteria for selecting which KSAPCs to measure on the written test:
  1. “Needed day one” on the job
  2. Important or critical (necessary) for the performance of the job
  3. Linked to one or more critical (necessary) job duties
  4. For job knowledge only: Should have high ratings on “Level Needed for Success”
     - Measure the job domain at a level that requires the applicant to have the information in memory (consequence of error)
- The KSAPCs included on the written test should meet these criteria generally and specifically (e.g., math skills)

Step 2: Develop a Test Plan for Measuring the Selected KSAPCs

- Components of a test plan:
  - What is the purpose of the test (baseline vs. mastery)
  - How will the test be scored? (multiple hurdle vs. compensatory)
  - Is the test a speeded test or a power test?
  - What is the reading level for the test?
  - How will the test be administered? (e.g., paper/pencil, computer-based, etc.)
  - Will a preparation or study guide be provided?
  - Will there be test preparation sessions offered?
Step 2: Develop a Test Plan for Measuring the Selected KSAPCs

- Considerations when choosing the number of test items to include on the written test:
  - Adequate sampling
    - Some KSAPCs require more items than others (e.g., mathematical reasoning vs. reading ability)
  - Proportional sampling
    - The test should be internally weighted
  - Sufficient number of test items to yield appropriate test reliability
    - The number of test items per critical KSAPC and in the test overall contributes most to test reliability.
    - Use of job experts to identify internal weights

Step 2: Develop a Test Plan for Measuring the Selected KSAPCs

- Choosing the type of test items:
  - Should the items be easy, difficult, or complex?
  - Bloom’s Taxonomy for Item Writing
    - Six levels of mastery that can be used to ensure that test items measure the intended KSAPCs at various levels
    - Knowledge
    - Comprehension
    - Application
    - Analysis
    - Synthesis
    - Evaluation
  - Choosing the item format
    - Multiple-choice
    - True/false
    - Open-ended or essay

Step 3: Develop the Test Content

1. Select a panel of 4-10 Job Experts
2. Review the Selection Plan, test plan, and validation surveys that will be used
3. Have Job Experts sign a confidentiality agreement
4. Train the Job Experts on item writing
5. Job Experts to practice writing and have peers critique the items
6. Write test items following the Selection Plan and test plan
7. Item writers to exchange items and critique
8. Delete all poor test items, keeping only the best
9. Create a final version of these draft items which will be reviewed by the validation panel

- If experienced test developers are used rather than Job Experts, begin at Step 6
Step 4: Validate the Test

- There are a number of factors to consider when validating the test:
  - The quality of the test item
  - The job-relatedness of the test item
  - The proficiency level of the test item
  - Does the item measure current information that must be memorized whereby there are serious consequences to job performance (job knowledge)

Step 4: Validate the Test

- Validation criteria for test items:
  - Every item on the test should be evaluated using the criteria discussed earlier.
  - There is no firm minimum criterion to apply to the key validation factors offered in the Guidelines or in the professional standards.
  - Two seminal court cases provide guidance on key validation criteria:
    - Contreras v. City of Los Angeles (1981)
      - Used a three-phase process to develop and validate an exam for an Auditor position. A “5 out of 7” rule (71%) was used to screen items for inclusion on the final test. This process was approved by the Ninth Circuit.
    - U.S. v. South Carolina (1978)
      - A panel of 10 Job Experts evaluated items on a National Teacher Exam. These review panels determined that between 63% and 98% of the items on the various tests were content valid and relevant for use. The U.S. Supreme Court endorsed this process as “sufficiently valid.”

Step 5: Score and Analyze the Test

- Item-level analysis
  - Point biserials
  - Item difficulty
  - Differential Item Functioning (DIF)

- Test-level analysis
  - Descriptive
    - Mean and standard deviation
  - Psychometric analyses for non-mastery-based tests
    - Test reliability
      - Standard Error of Measurement (SEM)
      - Conditional Standard Error of Measurement (CSEM)
  - Psychometric analyses for mastery-based tests
    - Decision Consistency Reliability (DCR)
    - Kappa Coefficients
Seven Steps for Developing a Content Valid Job Knowledge Written Test

• Step 1: Conduct a Job Analysis
• Step 2: Develop a Selection Plan
• Step 3: Identify Test Plan Goals
• Step 4: Develop the Test Content
• Step 5: Validate the Test
• Step 6: Compile the Test
• Step 7: Post-Administration Analyses

Step 1: Conduct a Job Analysis

• Job duty ratings
  – Frequency
  – Importance
• KSAPC ratings
  – Frequency
  – Importance
  – Differentiating “Best Worker” (if ranking)
  – When Needed
    o For job knowledge tests
      – Mastery level
        • Low level, familiarity, working knowledge, or mastery
• Duty/KSAPC linkages

Step 2: Develop a Selection Plan

• Review the KSAPCs from the job analysis
• Include essential KSAPCs that are linked to important job duties
• Job knowledge tests
  – Knowledge area must be important
  – Knowledge area must be needed the first day of the job
  – Knowledge area must be required at some level of mastery
  – Knowledge area must be measurable through a written test
Step 3: Identify Test Plan Goals

- Review the job knowledge areas that will be measured
- Identify the appropriate sources and/or textbooks
- Decide whether or not study sessions or preparatory materials will be provided
- Identify the number of items that will be included on the test
  - Typically job knowledge tests consisting of similar knowledge domains yield reliabilities in the high .80s to the low .90s when they include 80 items or more

Firefighter Certification Test Development Survey

Job Expert Name: [Name]
Date: [Date]

Instructions: Assume that you have $100 to “buy” the perfect firefighter for your department (based only on job knowledge qualifications - assume other important areas such as physical abilities and interpersonal skills have already been tested). How much money would you spend in the following areas to ensure that you have bought the most qualified firefighter for your department? Be sure that your allocations equal exactly $100.

<table>
<thead>
<tr>
<th>Knowledge Source to Choose From</th>
<th>Amount of Dollars You Would Spend to “Buy” the Perfect Firefighter</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pumping Apparatus Driver/Operator Handbook (1st ed.)</td>
<td></td>
</tr>
<tr>
<td>Principles of Vehicle Extrication (2nd ed.)</td>
<td></td>
</tr>
<tr>
<td>Fire Apparatus Company Officer (3rd ed.)</td>
<td></td>
</tr>
<tr>
<td>City and Emergency Services Intensive (6th ed.)</td>
<td></td>
</tr>
<tr>
<td>Animal Apparatus Driver/Operator Handbook (1st ed.)</td>
<td></td>
</tr>
<tr>
<td>Essentials of Firefighting (4th ed.)</td>
<td></td>
</tr>
<tr>
<td>Advanced Firefighting (3rd ed.)</td>
<td></td>
</tr>
<tr>
<td>Air Rescue and Heli Support (9th ed.)</td>
<td></td>
</tr>
<tr>
<td>The Source Book for Fire Company Training Evolutions (5th ed.)</td>
<td></td>
</tr>
<tr>
<td>Fire Inspection and Code Enforcement (6th ed.)</td>
<td></td>
</tr>
<tr>
<td>Hazardous Materials (7th ed.)</td>
<td></td>
</tr>
<tr>
<td>Ability to Compute Hydraulic Equations</td>
<td></td>
</tr>
<tr>
<td>TOTAL (must equal $100)</td>
<td></td>
</tr>
</tbody>
</table>

Process-by-Context Matrix: Police Sergeant

<table>
<thead>
<tr>
<th>Source</th>
<th>Definition</th>
<th>Principle</th>
<th>Application</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Essentials of Modern Police Work</td>
<td>4</td>
<td>10</td>
<td>20</td>
<td>54</td>
</tr>
<tr>
<td>2. Community Policing</td>
<td>5</td>
<td>7</td>
<td>13</td>
<td>25</td>
</tr>
<tr>
<td>3. Rules of Evidence</td>
<td>3</td>
<td>10</td>
<td>17</td>
<td>50</td>
</tr>
<tr>
<td>4. Department Rules &amp; Regulations</td>
<td>1</td>
<td>3</td>
<td>6</td>
<td>10</td>
</tr>
<tr>
<td>5. State Criminal Code</td>
<td>4</td>
<td>5</td>
<td>9</td>
<td>18</td>
</tr>
<tr>
<td>7. City Ordinances</td>
<td>2</td>
<td>2</td>
<td>6</td>
<td>10</td>
</tr>
<tr>
<td>8. Performance Appraisal Guidelines</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>9. Labor Agreement with the City</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Total</td>
<td>21</td>
<td>45</td>
<td>84</td>
<td>150</td>
</tr>
</tbody>
</table>
Step 4: Develop the Test Content

- Select a diverse panel of four to ten Job Experts with a minimum of one year of experience to review the test plan
- If Job Experts are going to write test items, provide proper training
- Review Job Experts’ items to ensure consistency (expect about 20-30% deletion)

Step 5: Validate the Test

- Job Experts to use the TVAP Survey to assign various ratings to the test items
- Job Experts to identify an appropriate time limit
  - Common rule-of-thumb is one minute per question plus an additional 30 minutes (e.g., a 150-item test would yield a three hour time limit)
  - A reasonable time limit would allow for at least 95% of the applicants to complete the test within the time limit

Step 6: Compile the Test

- Review Job Experts’ ratings and discard items that do not meet the criteria (see TVAP User Manual)
- Analyze “Angoff” ratings and discard raters whose ratings are statistically different from other raters by evaluating rater reliability and high/low rater bias
- Calculate the difficulty level of the test
Step 7: Post-Administration Analyses

- Item-level analysis
  - Point biserials
  - Item difficulty
  - Differential Item Functioning (DIF)
- Test-level analysis
  - Descriptive
    - Mean and standard deviation
  - Psychometric analyses
    - Test reliability
- Adjust the unmodified “Angoff” by using the SEM or CSEM where applicable

Steps for Developing a Customized Personality Test Using a Concurrent Criterion-related Validity Strategy

- Overview
  - These steps are appropriate for organizations that have at least 250 job incumbents in a specific job group
  - Including a personality measure in a hiring process is a great way to measure critical “soft skills” often not included in the testing process
  - Personality tests are often used to mitigate adverse impact in the hiring process as they tend to have little, if any, adverse impact
  - The ideal hiring process would consist of balanced measures to include both cognitive components and personality/work behavior components proportionately

Developing Personality Tests: Process Overview

- Step 1: Research the Personality Traits that Underlie Job Performance
- Step 2: Develop a Bank of Test Items that Measure the Targeted Traits
- Step 3: Develop a Lie Scale to Detect and Screen Out Faking and/or Random Responders
- Step 4: Develop a Job Performance Rating Survey (JPRS)
- Step 5: Convene the Supervisory Staff to Complete the JPRS on at least 250 Target Incumbents
- Step 6: Administer the Personality Items to the Job Incumbents
- Step 7: Choose a Strategic Test Building Strategy
Developing Personality Tests: Process
Overview (con’t)

• Step 8: Setup a Database for Conducting Correlational Analyses using a “Split Half, Calibration/Validation” Study
• Step 9: Create the “Optimal Test” Using the Calibration Sample Only
• Step 10: Evaluate the Validity of the Test by Correlating it to the JPRS Dimensions on the “Hold Out” Sample
• Step 11: Conduct a Fairness Study to Evaluate Whether the Test is an Adequate Predictor of Job Success for Subgroups
• Step 12: Assemble the Test and Determine the Appropriate Use in the Hiring Process
• Step 13: Complete a Criterion-related Validation Report to Address Section 15B of the Uniform Guidelines

Step 1: Research the Personality Traits that Underlie Job Performance

• Conduct literature searches to identify typical personality traits associated with specific positions
• Focus on the “Big Five” personality traits which describe five broad personality dimensions applicable to most positions:
  – Openness
  – Conscientiousness
  – Extroversion
  – Agreeableness
  – Neuroticism
• Results may indicate that one or more of these traits may be relevant to the targeted position(s) and in various directions (e.g., high levels for one position and low levels for another)

Step 2: Develop a Bank of Test Items that Measure the Targeted Traits

• Develop short, behaviorally-based statements
• Use a four-point scale (e.g., strongly agree to strongly disagree)
• Develop discrete items to control for “faking good”
• Generate low-level and high-level items for each targeted trait
• Have an expert review the items and discard duplicates (ideally about 100-200 items, with at least 20 measuring each targeted trait)
Step 3: Develop a Lie Scale to Detect and Screen Out Faking and/or Random Responders

- Nearly every criterion validation project is plagued with random or dishonest responders
- Develop 10-15 “lie-scale” items to be used for detecting and eliminating such responders
  - Example: One item may say “I prefer dogs to cats” and another item may say “I prefer cats to dogs”
- Evaluate these response patterns during the data build process and non-matching responders can be excluded

Step 4: Develop a Job Performance Rating Survey (JPRS)

- Quite possibly the most important step in the process
- Work with Job Experts to identify 10-20 critical job behaviors for the target position
- Work behaviors must be observable, rather than traits or abilities
- Use a rating scale for each dimension that forces a distribution whereby incumbents are rated by their supervisors (e.g., a scale that uses ten equal categories, with about 10% of the employees in each)
- This process will leverage the range of ratings obtained in the study, which allows the test items with greater levels of variance with which to correlate

Step 5: Convene the Supervisory Staff to Complete the JPRS on at least 250 Target Incumbents

- Obtain buy-in from the employee Unions ensuring value of the process and confidentiality
- Explain instructions to the supervisors
- Supervisors shall only provide ratings to those incumbents they have directly supervised for at least one month
- Ideally, each incumbent would be rated by at least two supervisors
Step 6: Administer the Personality Items to the Job Incumbents

- Administer the test to the incumbents
- Stress and stress again that test results will not be seen by supervisors nor any way used for evaluating their own personal job performance or pay status

Step 7: Choose a Strategic Test Building Strategy

- Two options
  - Conduct a factor analysis on the entire sample of test takers to build test scales to correlate to the JPRS
    - Either use factor scoring or manual item-factor weighting to evaluate the correlation of test scales
    - Weights are established by identifying items that are above a certain item-factor correlation threshold
  - Sort items into a priori trait categories (e.g., by grouping the dimensions together) and evaluate the validity against the entire sample using a one-tailed test for significance (because they are based on a directional hypothesis for the correlation)

Step 8: Setup a Database for Conducting Correlational Analyses using a “Split Half, Calibration/Validation” Study Design

- Third, and preferred, strategy
- Identify and remove responders based upon lie-scale results
- Join the two data files into a single database that contains each person’s (raw, unscored) response on the test items and their ratings given by supervisors
- Randomly divide the entire sample into two equal data sets
- One data set will be used to build the test (called the “Calibration” sample), and the other will be used to validate the test (called the “Validation” sample)
Step 9: Create the “Optimal Test” Using the Calibration Sample Only

- Run correlations between each test item and each of the JPRS dimensions on the Calibration sample
- After identifying the items and dimensions that exhibit the strongest correlations, assign point values to the items that somewhat mirror their correlational strength
- If items are negatively correlated with job performance, reverse score them for keying purposes (also double weighting the strongest items)
- Calculate the reliability of the new test form and ensure that the overall test (or each separate, scored subscale) has internal reliability that exceeds .70
  - Using a Calibration sample artificially inflates reliability, which will probably yield an overall reliability of .30-.40

Step 10: Evaluate the Validity of the Test by Correlating it to the JPRS Dimensions on the “Hold Out” Validation Sample

- Use the Validation sample to correlate the test to the JPRS dimensions
- Flag correlations that exceed statistical significance levels
  - Keep only those correlations that are significant
  - Refrain from going back-and-forth between the Calibration and Validation sample and “data mining”
  - This process should be data-driven

Step 11: Conduct a Fairness Study to Evaluate Whether the Test is an Adequate Predictor of Job Success for Various Subgroups

- Section 14B8 of the Guidelines requires conducting a fairness study whenever it is technically feasible to do so
- Typically, if a study includes at least 30 minorities or women, a fairness study should be conducted which requires using Moderated Multiple Regression (MMR) to evaluate the overall fairness of the test or DIF analyses for evaluating fairness at the test item level
- Guides for conducting both of these analyses are available for free through BCG/FPSI
Step 12: Assemble the Test and Determine the Appropriate Use in the Hiring Process

- Consider the test's validity, reliability, and adverse impact when deciding how the test will be used (e.g., pass/fail, banded, etc.)
- Given the moderate levels of validity (high teens to mid-twenties) and reliability (ranging between .70s-.80s) along with low levels of adverse impact, banding is usually the best option
  - Using the conditional Standard Error of Difference when setting score bands is especially helpful
- Be sure to include a warning on the test introductory page to discourage applicant faking

Step 13: Complete a Criterion-related Validation Report to Address Section 15B of the Uniform Guidelines

- Benefits of completing a report that address the federal Uniform Guidelines
  - Provides a ready, off-the-shelf document that can be used in litigation situations
  - Documents the key (complex) steps used to build the test
  - Provides the test with “shelf life”

Summary

- Whether developing and validating an entry-level aptitude written test, a job knowledge test for promotions, or a personality test, it is imperative that the developer addresses the rigorous and robust requirements of various federal enforcement agencies
- Doing so benefits the employer, through insulation against potential litigation, and the applicant, by offering a fair, and valid hiring process