TestGenius - Retail

Retail Associate Selection System

Test Development and Validation

November 21, 2019

193 Blue Ravine Road, Suite 270
Folsom, CA 95630
800-999-0438
www.biddle.com

Copyright © 2019 Biddle Consulting Group, Inc. • 800-999-0438

Table of Contents

Copyright © 2019 Biddle Consulting Group, Inc. • 800-999-0438

Executive Overview

Biddle Consulting Group (BCG) has been in the assessment space for over 40 years. Our administrative skills test (OPAC®) is one of the market leaders, with thousands of clients and millions of applicants tested. Our CritiCall® is the leading test in the 911/Emergency Calltaker space and is used by over 20% of the 911 call centers in the U.S. and the majority of State Highway Patrol agencies. Our firm developed the leading Nurse Assessment in the U.S., with the highest predictive validity in the history of nurse assessment. Our president, Dr. Daniel Biddle has been involved as an expert consultant/witness in over 100 state/federal court cases in the areas of testing, validation, and statistical analysis. Dr. Biddle's book, *Adverse Impact & Test Validation*, has sold thousands of copies and has been used as desk reference by federal enforcement officials for over a decade. Dr. Biddle has been hired by plaintiff and defense attorneys, competing test development firms, employers, and the federal government to work on classified and/or confidential testing matters.

Why is this background important, especially in the retail assessment space? It's to show that our firm develops assessments with a *bias*—a "measure twice, cut only once bias." Personnel selection is risky business. Employers that administer tests that exhibit adverse impact risk millions in federal class action lawsuits, as well as their name being plastered across news wires as being found guilty of "employment discrimination." Some retail employers attempt to dodge test liability by not testing at all, or by using trendy, watered-down tests that have little selection utility. Many such tests currently flood the retail testing space. Many take next to no time to administer, sometimes use games, or use catchy bio-data items that oftentimes predict only for certain group members.

It is through this background we realized that the retail industry needed a "real" solution—an assessment that measured a mixed combination of skills and abilities needed for the majority of retail positions in the U.S: trainability skills, people skills, and dependability. After two years of research and development, the final product is now ready and constitutes the most thorough, yet shortest (about 30 minutes to complete, or only 15-20 minutes if only 2 of 3 scales are used) retail assessments on the market.

What sets our Retail Assessment apart from others? Here are some of the ways in which retail assessments have "failed" in field of EEO litigation over the years:

1. **They lack a real, demonstrable connection to the job.** Let's face it: at the end of the day, the "test validity" defense offered by employers in litigation settings will be won or lost by a judge and/or jury. It will not be decided by the esoteric philosophies or theories of the testifying expert witnesses. If a judge can see a

Copyright © 2019 Biddle Consulting Group, Inc. • 800-999-0438

real, tangible connection between the test and the job, the case is won. If not, it's game over for the employer.

2. **Validity by inference.** One of the latest trends of some employers and testing experts is to "impute" the "general" validity of a test into a specific testing situation. This technique, known as "Validity Generalization" typically fails in court, leaving employers with millions in liability.[1] Leading EEO enforcement agencies are not fans of this technique, neither are we. Since 1991 "Situational Specificity" has been the law of the land (1991 Civil Rights Act (42 U.S.C.).

3. **The use of trendy, unproven selection techniques or content.** It is true that sometimes a single bio-data item can out-predict an entire test battery. Answering "yes" to "I built a model airplane that flew before I was 12" out-predicted pilot performance compared to the entire ASVAB test used for selecting military candidates. But how many women were out building model airplanes at age 12? What about some under-privileged people? An over-reliance on "magic" and "short" predictors can sometimes provide very "choppy" statistical results, predicting job performance only for certain groups. This is certainly not as "fair" as allowing all applicants to equally compete on tests that measure skills and abilities that have a wide and obvious relationship to the retail jobs to which they are applying. This is a much more well-rounded assessment strategy compared to splitting the entire diverse population into those who have had the opportunity for certain experiences. Our test is a more well-rounded assessment solution, where qualifications are weighed against the real-life skills necessary for the job.

4. **Testing "Black Boxes."** We have even reviewed testing "black boxes," where applicant scores are computed in "real time" against evolving test-criterion relationships. Applicant scores change from day to day based on "learning algorithms." Good luck explaining that to a judge, or an applicant for that matter.

5. **The "Stamp of Approval" technique.** Sometimes employers install assessments after only 1-2 job experts have reviewed the content and given approval. From a validity standpoint, typically more job experts are needed, with the majority approving the content as "job related."

6. **"Personality tests are everything" method.** Some employers (at times, running away from adverse impact liability) have tried installing short, fakeable personality tests as the "one size fits all" solution. Yes, personality tests typically have less adverse impact, but they are fakable, the research shows they are faked, and should only be a limited part of an assessment mix.

[1] https://www.opac.com/articles/validity-generalization.pdf

Copyright © 2019 Biddle Consulting Group, Inc. • 800-999-0438

Background

The TestGenuis – Retail Assessment was developed by Biddle Consulting Group, Inc. (BCG) to provide retailers with a valid and highly defensible test/selection system useful for hiring retail associates in a wide range of retail settings (e.g., clothing, home goods, grocery, hardware, electronics). This Retail Assessment was developed to fill a "void" in the retail assessments market as many of the available assessments appear to prioritize candidate experience (i.e., appealing test content, gamification, minimal test-taker effort) and short length, quick administration over and above defensibility and validity. This can leave retailers scrambling for answers should action be taken against them in the event of a Title VII challenge. While efficiency in any selection process is an important goal, especially one in which hiring volumes are high, by severely limiting the length of assessments or making the content too abstract, they become less reflective of the type of work being performed by retail associates, less relevant, less reliable, and often less valid/defensible.

The TestGenius Retail Assessment was designed to simultaneously address the need for greater efficiency in the retail associate hiring process (necessitated by the sheer volume of hiring within this sector), while also maximizing validity and legal defensibility which *should* always be a top priority when using assessments for high volume hiring. It is ideal for use earlier in the selection "funnel" because it reduces the number of unqualified applicants that move forward in the selection process, allowing retailers to focus their selection efforts on the qualified applicants which saves time and money.

The Retail Assessment is considered a "selection system" because it is not a single "test" but a system that consists of three items sets/scales that provide a more holistic and thorough evaluation of a retail candidate's skills, abilities, and personal attributes that are most likely to select qualified retail employees. Selecting the right retail associate is critical. Organizations that select the right person the first time realize decreases in involuntary turnover, increased productivity, and higher employee engagement among other benefits. When the right person is put in place, organizations ensure they are positioned to provide the best experience for their customers.

To keep test length and administration time manageable without sacrificing test reliability and validity, only test questions that closely match (high face-validity) the types of tasks and scenarios that typical retail associates encounter on the job were included on the Retail Assessment.

The Assessment was designed for unproctored online administration on any device, including mobile phones. Mobile device administration has been shown to increase access to more diverse candidate populations (Winfred, Doverspike, Muñoz, Taylor, & Carr, 2014).

Copyright © 2019 Biddle Consulting Group, Inc. • 800-999-0438

The following report provides a detailed explanation of the development and initial validation of the Retail Assessment and its components. Please note that the validation steps and processes described below were conducted in order to help insure the Retail Assessment would be "validatable" for a wide range of retail associate positions. Employers are advised to conduct the validation steps provided in the "Validation Strategies" section of this report if/when their use of the Retail Assessment results in statistically significant adverse impact.

Targeted Positions

The two groups of target positions for the Retail Assessment include **Retail Salespersons** (O*NET 41-2031.00) (such as "Retail Associates" at common retail outlets, such as Lowes, Walmart, CVS, REI and others) and **Customer Service Representatives** (O*NET 43-4051.00). While self-evident, these titles typically perform duties such as (per O*NET):

Retail Salesperson:

- Greet customers and ascertain what each customer wants or needs.
- Describe merchandise and explain use, operation, and care of merchandise to customers.
- Recommend, select, and help locate or obtain merchandise based on customer needs and desires.
- Compute sales prices, total purchases, and receive and process cash or credit payment.
- Answer questions regarding the store and its merchandise.

Customer Service Representatives:

- Confer with customers by telephone or in person to provide information about products or services, take or enter orders, cancel accounts, or obtain details of complaints.
- Check to ensure that appropriate changes were made to resolve customers' problems.
- Keep records of customer interactions or transactions, recording details of inquiries, complaints, or comments, as well as actions taken.
- Resolve customers' service or billing complaints by performing activities such as exchanging merchandise, refunding money, or adjusting bills.
- Complete contract forms, prepare change of address records, or issue service discontinuance orders, using computers.

Copyright © 2019 Biddle Consulting Group, Inc. • 800-999-0438

For example, a typical **Associate** Job Description reads:

> A position with numerous responsibilities, a sales associate primarily provides customer service. Additional job duties include stocking shelves, maintaining a clean work environment, assisting in sales, and performing cashier responsibilities, in some cases. Friendly workers with personable attitudes and motivated personalities typically make ideal Target sales associates. The employer also looks for energetic, knowledgeable, and positive sales associates. During training, new sales associates learn proper store protocol, merchandise inventory, loss prevention tactics, and customer interaction skills.

TestGenius - Retail Assessment Description

The Retail Assessment consists of three main content areas: 1) Trainability, 2) People Skills, and 3) Dependability.

Trainability Assessment

New hires who are better able to learn the knowledge and skill required to perform the job, perform better during training and on the job itself (Schmidt & Hunter, 2004). This requires the capacity to read, interpret, comprehend, retain, and then apply new information. The Trainability Assessment was designed to assess these capacities using realistic question that closely resemble stimuli (e.g., training materials, product codes, instructions) applicants will encounter during training and while performing their job duties. The Assessment contains 15 multiple-choice questions. Each question contains four answer options, with one answer keyed as the correct answer. The questions assess an applicant's job readiness for retail associate positions in the following core areas:

1. **Attention-to-detail** (i.e., inspection): Example items include comparing stock numbers (key against a list), isolating incorrect parts or product differences, or similar.
2. **Interpreting and Applying Information:** Interpreting and making basic conclusions from training materials, safety cards, written on-the-job instructions, and policy and procedure.
3. **Numeracy skills:** While most calculating is done electronically, basic stocking/shelving/facing duties are not, as well as other similar duties that may be required when the electronic calculations are unavailable (e.g., making change, processing a coupon, discount).

Copyright © 2019 Biddle Consulting Group, Inc. • 800-999-0438

People Skills Assessment

This Assessment consists of video-based, situational judgment (VSJT) items designed to measure interpersonal competence in retail-related situations. The final version of this test includes 15 video scenarios (10-30 seconds in length) that present interpersonal situations that occur in retail environments (assisting customers, resolving customer service issues, and working with coworkers). Each video is followed by a set of written response options from which the applicant is asked to select the "most effective" and "least effective" way of handling the situation. The questions assess an applicant's job readiness in the following core areas:

1. **Problem Avoidance/Solving:** Includes multiple discrete soft skills such as, applying appropriate and professional social skills while interacting with others, adaptability and personal flexibility in various situations, and making concessions and building consensus to achieve work related goals.
2. **Effective Communication:** The ability to interact and work effectively with coworkers, supervisors, and customers, including those with varying socio-economic/ethnic, or other backgrounds.
3. **Customer Centric Focus:** Includes interacting with customers in a helpful, polite, friendly and positive manner. Showing a positive and willing attitude when addressing customer questions or issues, taking initiative and follow-through and attention to appropriate details.

Dependability Assessment

Of the "Big 5" personality factors, conscientiousness has typically shown the strongest relationship with job performance (including performance during training) across many occupational settings (Schmidt & Hunter, 2004). In retail customer service workers, conscientiousness has been shown to incrementally add validity to the prediction of job performance over and above that offered by cognitive ability alone (Avis, Kudisch, & Fortunato, 2002). It has also been shown to moderate the relationship between job knowledge and the delivery of high quality customer service, such that better customer service is delivered by more conscientious workers even when their customer service related job knowledge levels are equally high (Motowidlo, Brownlee, & Schmit, 2008).

BCG developed a 29-item personality scale that assesses an individual's conscientiousness which focuses primarily on the achievement and dependability factors. Individuals higher in achievement orientation set high personal standards, strive to succeed, and direct their behavior toward goal accomplishment. Dependability describes those who are more reliable, thorough, trustworthy, and likely to follow through (John, Naumann, & Soto, 2008). The items are scored using a four-point Likert-type scale, from "strongly agree" to "strongly disagree" with no neutral option).

Copyright © 2019 Biddle Consulting Group, Inc. • 800-999-0438

Retail Assessment Development

Because retail associate positions require a mix of cognitive, interpersonal, and behavioral competencies, the retail industry would benefit from an assessment solution that measures these traits with a context-rich, high-fidelity and multi-faceted assessment tool, that is easily administered online via mobile device. BCG consultants led the test development effort which began in 2018 and was concluded in 2019.

Trainability Assessment

Item development for the Trainability component of the Assessment was focused on ensuring a close match between the questions and the content of retail associate positions. To do this, BCG item writers visited ten retailers located in the greater Sacramento, CA area and informally conducted job observations of associates performing their job duties. Interviews were also carried out with current and former retail associates to gain a clearer perspective about the types of work typically carried out by associates in a retail setting, as well as the types of customer and coworker interactions that often take place. These efforts allowed item writers to target their item development focus only to those topics and scenarios that are the most common to retail settings.

A total of 110 multiple-choice test questions were developed to assess Trainability. The items were developed at a Flesch-Kincaid reading level of 8th grade to reduce cognitive loading, and ensure that vocabulary and word usage (questions and instructions) were at an appropriate level for entry level hiring. The items were developed to assess basic attention to detail such as comparing product numbers or codes, applying job related information such as instructions or written information, basic numeracy skills including making change or figuring out the correct discount percentage, and basic decision making.

To gather data on the quality and psychometric properties of the test questions, they were administered online to a sample of 117 participants who indicated they had previous retail associate work experience. The participants were instructed to respond to the questions as if they were a job applicant vying for a job. BCG staff evaluated the responses and eliminated 17 anomalous responses from the data set before analyzing the results. The final sample consisted of 100 individuals.

The question response data was analyzed using Classical Test Theory statistics (i.e., p-values, item-total correlations, alpha) to identify questions that performed adequately in terms of difficulty and discrimination. The test questions that performed well during this initial pilot study were then adapted for administration on a mobile device, some longer items that would have required scrolling on a mobile screen were eliminated as well. A set of 15 questions that were capable of being delivered on a mobile device without zooming, pinching or scrolling, or that could be administered without

Copyright © 2019 Biddle Consulting Group, Inc. • 800-999-0438

variation between devices (PC, tablet, mobile), were selected by the test developers for inclusion in the Trainability Assessment.

A follow up study was performed in order to further refine the test content. A sample of 290 participants gathered through a targeted online survey system was obtained. In order to maximize the reliability, one item was removed for a final set of 14 Trainability test items.

People Skills Assessment

Situational judgment tests (SJTs) have received much attention in the personnel selection literature (e.g., Christian, et. al., 2010). Some of the key benefits attributed to SJTs include increased testing fidelity and lowered adverse impact in the selection process (Weekley & Jones, 1997). Specifically, when comparing written- and video-based SJTs, video-based SJTs have been identified as the preferable method for lowering adverse impact because the delivery is context-based and the academic load is minimized (Chan & Schmitt, 1997). The advantages offered by video-based SJTs are a good match for retail associate testing due to several factors including, improved candidate perceptions, the increased diversity of job applicants, and higher applicant flows.

Item development for the People Skills Assessment was focused on capturing important interpersonal interactions that occur in a retail setting between associates and customers, and between associates and other employees. These "critical incidents" were based upon actions that would differentiate performance between retail associates (i.e., situations where a positive behavioral choice would likely result in positive work outcomes and a poor behavioral choice would likely result in negative work outcomes). All scenarios were designed to measure an applicant's ability to appropriately respond to situations that retail associates may encounter starting the first day of the job.

BCG item writers developed an initial batch of 35 retail situational judgment scenarios and test questions and response options for each scenario. The questions consisted of four response options for which candidates are asked to choose the "most effective" and "least effective" response. This initial group of scenarios/questions was reviewed internally by BCG Consultants and modified to improve the clarity of the scenarios and question relevance and response option plausibility. Twenty of the initial scenarios/questions that were best suited for video-based delivery in terms of fidelity and clarity were selected for VSJT filming.

Item writers created vignettes around each of the 20 scenarios including the context, dialogue, equipment needed, and the minimum number of people involved in the scenario. They also confirmed that one possible "best" response, one possible "worst" response, and two plausible distractors that were neither the best nor the worst

Copyright © 2019 Biddle Consulting Group, Inc. • 800-999-0438

response were present for each vignette, and made modifications as necessary. Scripts were then written for each of the vignettes that had been approved for filming. Once the scripts were finalized, they were reviewed for their appropriateness and accuracy, such as determining whether the dialogue accurately mimicked the type of interactions that would commonly occur in retail settings, and whether the props and proposed setting closely mirrored what is commonly found in retail workplaces. Finally, the vignettes with the greatest potential to distinguish between levels of job performance were identified, reducing the number of vignettes from 20 to 15.

A follow up study was performed in order to further refine the test content. A sample of 290 participants gathered through a targeted online survey system was obtained. In order to maximize the reliability, one item was removed for a final set of 14 People Skills test items.

The 14 scripted vignettes were filmed in January 2019 at a local retailer in Sacramento, California. The filming of the scripted vignettes was overseen by BCG test development personnel. A combination of individuals (i.e., retail staff, BCG staff, and actors) were used to portray retail associates, retail supervisors, and customers during the filming.

Validation Workshops for Trainability and People Skills Assessments

BCG Consultants facilitated validation workshops for the Trainability and People Skills assessments at the BCG office in Folsom, CA during April 2019. Experienced retail supervisors were recruited to participate because they directly observe the work performed by retail associates as well as the personal attributes, skills, and abilities that most differentiate effective job performers from those who are less effective. During the workshops the Trainability and People Skills assessments were administered via mobile device to the retail job experts who were asked to complete the assessments as if they were candidates competing for a retail associate position.

Qualitative feedback about the assessments including the ease of navigating the content and answering the questions on a mobile device was collected. Additionally, the job experts provided feedback about the relevance of the test content, clarity of the questions and vignettes, and plausibility of the question answer options for both assessments. Each job expert was provided with their scores on the two assessments to provide a context from which to base their validation survey ratings and estimates.

The demographic information for the validation workshop job experts is presented in Tables 1 through 5.

Table 1. Gender of Job Experts

Gender	N
Male	12

Copyright © 2019 Biddle Consulting Group, Inc. • 800-999-0438

Female	8

Table 2. Race/Ethnicity of Job Experts

Race/Ethnicity	N
White	14
Black/African American	2
Hispanic/Latino	1
Asian/Pacific Islander	2
Native American/Alaska Native	1
Prefer not to answer	0

Table 3. Age of Job Experts

Age	N
Less than 20 years of age	0
20-29 years of age	5
30-39 years of age	5
40-49 years of age	1
50-59 years of age	5
60 or more years of age	4
Prefer not to answer	0

Table 4. Years of Experience Supervising of Job Experts

Years of Experience	N
Less than 1 year	1
1 year	4
2 years	4
3 years	1
4 years	2
5 years	1
Between 6-10 years	2
More than 10 years	5

Table 5. Type of Retail Experience Supervision

Retail Type	N
Building Supplies/Hardware	1
Electronics	4
Clothing	6
Grocery	4
Multi (e.g., Walmart, Target)	2
Other: Automotive, Seasonal Retail Mgmt., Swimming Pool	5

Copyright © 2019 Biddle Consulting Group, Inc. • 800-999-0438

Trainability Assessment Validation Study Results

The validation job experts provided ratings regarding relevance and cutoff scores at both the minimally and highly competent cut points, used to determine the hiring recommendation ranges for the Trainability assessment. To assess the relevance of the Trainability questions, job experts were asked to indicate "Yes" or "No" to the following validation item:

Table 6. Relevance of Trainability Questions

Validation Item	% Indicating Yes
Do the trainability questions measure attributes that are important to the job performance of entry level retail associates?	95% (19 of 20)

People Skills Assessment Validation Study Results

The validation job experts provided ratings regarding relevance and cutoff scores at both the minimally and highly competent cut points, used to determine the hiring recommendation ranges for the People Skills assessment. To assess the relevance of the People Skills questions, job experts were asked to indicate "Yes" or "No" to the following validation item:

Table 7. Relevance of People Skills Questions

Validation Item	% Indicating Yes
Do the video based questions measure attributes that are important to the job performance of entry level retail associates?	100% (20 of 20)

People Skills Assessment Keying and Scoring

Research by Motowidlo and Beier (2010) suggests that, when job experts with job experience are actively involved in the keying process, test scores are more related to job performance. Using a content validity strategy which is supported by the aforementioned research, we developed a rational keying system for the test based on the consensus opinions of the panel of 20 highly-experienced retail supervisors.

These 20 experts reviewed all 15 video vignettes and rated each of the four response alternatives as either "Most Appropriate," "Second Most Appropriate," "Third Most Appropriate," or "Least Appropriate." After analyzing the expert responses, a multi-point keying rubric was developed that awarded the most points to applicants who

Copyright © 2019 Biddle Consulting Group, Inc. • 800-999-0438

agree with "high-rater consensus" alternatives, fewer points to "moderate agreement" alternatives, and even fewer points to "majority disagreement" alternatives. This design also penalized applicants who select a "high consensus-best choice" as the "least effective" choice, or vice versa. Specifically, the following scoring logic was used.

- Correct responses with which there was > 90% agreement by the job experts during the keying process were awarded three points.
- Correct responses with which there was < 90% but > 70% agreement by the job experts during the keying process were awarded two points.
- Correct responses with which there was < 70% but > 50% agreement by the job experts during the keying process were awarded one point.
- Incorrect responses with which there was > 90% agreement by the job experts during the keying process were penalized three points.
- Incorrect responses with which there was < 90% but > 70% agreement by the job experts during the keying process were penalized two points.
- Incorrect responses with which there was < 70% but >50% agreement by the job experts during the keying process were penalized one point.

Using this scoring logic places a premium on the response alternatives that have a higher level of job expert consensus than others. This awards more points to those applicants whose item responses are more aligned with the keying experts, and penalizes those applicants whose item responses are in disagreement with the keying job experts. Based on the results of this consensus keying process, the People Skills Assessment has a total maximum score of 39 points possible. Note that some response options are not currently scored due to lack of consensus by the job experts on whether they were the "most appropriate" or "least appropriate" responses

The People Skills test measures the ability to interact and work effectively with others in different work-related settings. The skills and abilities measured by People Skills tests were identified as:

1. **Problem Avoidance/Solving:** Includes multiple discrete soft skills such as, applying appropriate and professional social skills while interacting with others, adaptability and personal flexibility in various situations, and making concessions and building consensus to achieve work related goals.
2. **Effective Communication:** The ability to interact and work effectively with coworkers, supervisors, and customers, including those with varying socio-economic/ethnic, or other backgrounds.
3. **Customer Centric Focus:** Includes interacting with customers in a helpful, polite, friendly and positive manner. Showing a positive and willing attitude when addressing customer questions or issues, taking initiative and follow-through and attention to appropriate details.

Copyright © 2019 Biddle Consulting Group, Inc. • 800-999-0438

Each item was then evaluated by subject matter experts in order to assess the percentage of each competency being measured. The items were assessed on a scale of 0% to 100% for each competency irrespective of the other competencies. Once the ratings were collected and outliers were removed, the average of the weightings became the subdomain weight for each item.

Dependability Assessment Development

Meta-analyses have consistently shown that measures of conscientiousness while being predictive of important job performance criteria, do not result in score differences amongst protected classes (e.g., ethnicity, gender, age). Therefore, these measures do not typically contribute to adverse impact, and may actually reduce adverse impact when used with cognitive ability tests (Oswald & Leaetta, 2010).

In 2018 BCG Consultants developed a bank of personality test items designed to assess facets of conscientiousness including achievement orientation and dependability. Each item was provided with a four-point Likert-type scale, which ranged from "strongly agree" to "strongly disagree" with no neutral option. The neutral option was omitted to ensure a scorable response was provided to each item. A set of 35 items, including 20 which were reverse scored, were selected for the construct validation study (see note on construct validity in the Validation Strategies section of this report).

A follow up study was performed in order to further refine the test content. A sample of 290 participants gathered through a targeted online survey system was obtained. In order to maximize the reliability, one item was removed for a final set of 24 Dependability test items.

Dependability Assessment Validity Study

The 35 BCG Dependability items and the 48 item NEO-PI3 (commercially available, factor analytically derived valid measure of conscientiousness) were combined into a single form interspersed with a 16 item "carelessness" scale.[2] In May 2018, this combined form was administered online to a sample of 202 participants who were between the ages of 18 and 74, (average age = 34) and fluent in English. The respondents were instructed to respond to the items as if they were job applicants vying for a job.

[2] The "carelessness" scale contained 16 dichotomous True/False items. Half were reverse coded. For example, one item stated, "I prefer the color red over blue." Later in the form the reverse of this item was presented, "I prefer the color blue over red." Test takers with inconsistent responses to these pairings would be eliminated from the data file.

Copyright © 2019 Biddle Consulting Group, Inc. • 800-999-0438

The demographic information for the respondents in the *convergent validity study* is presented in Tables 8 and 9.

Table 8. Gender of the Respondents

Gender	N
Male	131
Female	71

Table 9. Race/Ethnicity of Respondents

Race/Ethnicity	N
White	137
Black/African American	6
Hispanic/Latino	4
Asian/Pacific Islander	53
Native American/Alaska Native	0
Prefer not to answer	1

The data was screened prior to analysis to identify any participants who may have provided anomalous and/or careless responses to the items. The data from all 202 respondents was retained as none appeared to be careless or anomalous.

The item-total correlations for each of the 35 items on the BCG Dependability scale was evaluated using the total score on the NEO-PI3 (α = .932) as the criterion. Six of the Biddle Dependability scale items were found not to be significantly positively correlated with NEO-PI3 total scores and thus were deleted from the BCG Dependability scale. The internal consistency of the 29 remaining BCG Dependability scale items was evaluated using coefficient alpha and found to be α = .86. The U.S. Department of Labor's (DOL) benchmarks for interpreting reliability coefficients (DOL, 1999) categorize reliability coefficients of 0.90 and above as "excellent," 0.80 to 0.89 as "good," 0.70 to 0.79 as "adequate," and those below 0.70 as "may have limited applicability."

The convergent validity of the BCG Dependability scale (i.e., BCG's measure of conscientiousness) was determined by evaluating the strength of linear relationship between total scores on the BCG Dependability scale (29 items) and total scores on the NEO-PI3 (i.e., known valid measure of conscientiousness). The Pearson Correlation between these two measures of conscientiousness demonstrated moderate to high convergent validity ($r = .776$, $p < .001$).

Copyright © 2019 Biddle Consulting Group, Inc. • 800-999-0438

Reliability and Score Banding Study

A follow up study was performed in September 2019 in order to further refine the test content, established statistical bands, and obtain an updated internal consistency reliability estimate for the test. A sample of 311 participants gathered through a targeted online survey system was obtained. Once outliers were removed, the total number of participants was reduced to 290. The parameters for the participation were that the individuals must reside in the United States and be between the ages of 18 and 30.

The demographic information for the reliability and score banding study job experts is presented in Tables 10 through 13.

Table 10. Gender of Job Experts

Gender	N
Male	120
Female	167
Prefer not to answer	3

Table 11. Race/Ethnicity of Job Experts

Race/Ethnicity	N
White	180
Black/African American	36
Hispanic/Latino	31
Asian/Pacific Islander	32
Native American/Alaska Native	3
Prefer not to answer	8

Table 12. Age of Job Experts

Age	N
Less than 20 years of age	4
20-25 years of age	255
26-30 years of age	31

Table 13. Average Age of Job Experts

Average Age
23.96

Test Module Reliability

Based upon an analysis of the 290 scores from the study, internal consistency reliability coefficient was calculated for each of the three tests. In order to maximize the

Copyright © 2019 Biddle Consulting Group, Inc. • 800-999-0438

reliability on each test, items that did not contribute to the reliability of the test were removed. The final number of items for each test is shown in Table 14.

Table 14. Final Test Item Numbers

Test Name	Previous # of Items	Final # of Items
Trainability	15	14
People Skills	15	14
Dependability	29	24

The internal consistency of the remaining items was evaluated using coefficient alpha. The U.S. Department of Labor's (DOL) benchmarks for interpreting reliability coefficients (DOL, 1999) categorize reliability coefficients of 0.90 and above as "excellent," 0.80 to 0.89 as "good," 0.70 to 0.79 as "adequate," and those below 0.70 as "may have limited applicability.". The reliability for each of the tests modules is shown in Table 15.

Table 15. Test Form Reliability

Test Name	Alpha Reliability
Trainability	.67
People Skills	.78
Dependability	.87

Test Module Score Bands

The scores for each module are estimates of a test taker's actual proficiency level in the abilities being tested. All tests lack some degree of measurement precision called error. One estimate of measurement of error is the Standard Error of Measurement (SEM). The SEM provides a reliability estimate at each point in a score distribution rather than providing a single reliability estimate for the entire distribution. To maximize the accuracy of decisions based on the TestGenius Retail assessments, it is recommended that selection decisions should be based on score bands as opposed to raw test scores. Because of measurement imprecision, advanced psychometric methods have been developed to identify score bands that describe test score ranges that are statistically similar in terms of error. For the TestGenius Retail test forms, score bands were developed using the SEMs.

Analysis of the test taker data revealed three distinct bands for each of the test modules. These bands are defined as Strongly Recommended, Recommended, and Not Recommended. The score bands are defined in Table 16 – 18 for each of the test modules.

Table 16. Trainability Score Band

Copyright © 2019 Biddle Consulting Group, Inc. • 800-999-0438

% Score Range	Band
68% - 100%	Strongly Recommended
40% - 67%	Recommended
0% - 39%	Not Recommended

Table 17. People Skills Score Band

% Score Range	Band
68% - 100%	Strongly Recommended
48% - 67%	Recommended
0% - 47%	Not Recommended

Table 18. Dependability Score Band

% Score Range	Band
83% - 100%	Strongly Recommended
75% - 82%	Recommended
0% - 74%	Not Recommended

Test Form Score Bands

The TestGenius Retail module also provides an overall recommendation based upon the obtained score band for each test taker. In order to achieve an overall Strongly Recommended, the test taker must be Strongly Recommended in all 3 tests. To fall in the Recommended band, the test taker must NOT fall into the Not Recommended band for any test and also NOT meet the qualifications for Strongly Recommended. Any test taker with a Not Recommended score in any test will receive a Not Recommended overall score.

The overall band distribution is shown in Table 19.

Table 19. Dependability Score Band

# of Test Takers	Band	% of Individuals
45	Strongly Recommended	16%
90	Recommended	53%
155	Not Recommended	31%

Copyright © 2019 Biddle Consulting Group, Inc. • 800-999-0438

Validation Strategies

When is Validation Evidence for the Retail Assessment Required?

Employers that use the Retail Assessment in a way (e.g., cutoff, banding, ranking, etc.) that exhibits statistically significant *adverse impact* are required under the Uniform Guidelines to develop validation evidence for the Retail Assessment at the location(s) and position(s) for which it is used. Adverse impact occurs when the selection rate difference between two groups resulting in a p-value of less than .05 using the Fisher Exact Test (FET) with the Lancaster (1961) mid-P (LMP) adjustment (Biddle & Morris, 2011). See www.disparateimpact.com.

Retail Assessment Validation Requirements and Strategies

Practically speaking, a "valid" selection procedure is one that measures the actual requirements of the job in a fair and reliable way. A valid selection procedure is one that "hits the mark," and does it consistently, with the mark being one or more essential requirements for a given position that are targeted by the selection procedure. A valid selection procedure effectively measures the net qualifications that are really needed for the job, and not much more or less.

In the legal realm, a selection procedure is valid if it can be proven by an employer in litigation that it is ". . . *job related for the position in question and consistent with business necessity*" (to address the requirements of the 1991 Civil Rights Act, Section 703[k][1][A][i]). This standard is usually met (or not) by arguing how the selection procedure first addresses the Uniform Guidelines[1] (1978), followed by professional standards (i.e., the Standards and Principles), then by parallel or lower courts that have applied the standard in various settings.

Under the Uniform Guidelines, three types of validity evidence are allowed: content, criterion-related, and construct validity. Each will be briefly discussed below, followed by application to the Retail Assessment.

Content validity is demonstrated by data showing that the content of a selection procedure is representative of important aspects of performance on the job (see section 5B and section 14C). A content validity study is conducted by *linking* the essential parts of a job analysis (the job duties and/or knowledges, skills, and abilities) to the selection procedure. Thus, content validity is formed by creating a *nexus* between the job and the selection procedure. It relies on a process that requires Job Experts (incumbents or immediate supervisors) to provide judgments (usually by providing ratings on surveys) regarding *if* and *how well* the selection procedure represents and measures the important parts of the job.

Copyright © 2019 Biddle Consulting Group, Inc. • 800-999-0438

Criterion-related validity is statistical. This type of validity is achieved when a selection procedure is statistically correlated with important aspects of job performance at a level that is "statistically significant" (with a probability value less than .05). One interesting benefit of this type of validity is that the employer is not pressed to define exactly what the selection procedure is measuring. While is it always a very good idea to know and describe to applicants the skills or abilities that are measured by the selection procedure, it is not a requirement to do so because the selection procedure is scientifically related to job performance. By contrast, content validity has specific requirements for the employer to show and describe exactly what skills and abilities are being measured by the selection procedure and how they related to the job (see 15C4 – 5 of the Uniform Guidelines).

Criterion-related validity can be achieved by correlating selection procedure scores to several different types of job performance measures, including both subjective and objective measures. The most typical subjective performance measures include supervisor ratings and/or peer ratings of work products (quality and/or quantity) or job performance, and performance review scores.[2] Objective measures can include quantifiable work output measures (*e.g.,* number of widgets produced per hour), quality-related measures (*e.g.,* number of widgets returned because of defects), absenteeism, turnover, disciplinary actions, safety incidents, and other aspects of performance that are gathered and recorded in a uniform and consistent manner.

Construct validity is not applied frequently in the field of personnel selection, and typically requires a local or transported criterion-related validity study as a base foundation. This is especially true in high adverse impact situations, whereas the validity evidence may be less (and even not required for legal defensibility) in low to zero adverse impact situations.

Which type of validity evidence should be used for the Retail Assessment? Because the Trainability Assessment measures cognitive/academic abilities in a job-related context, a content validity strategy should be used. This strategy is also ideal for the People Skills Assessment, because it attempts to replicate and/or measure certain aspects of the retail position that involve interpersonal interactions. Because the Dependability Assessment measures conscientiousness (a latent trait), either criterion-related or a construct validation technique should be used. However, due to the low subgroup differences typically exhibited by assessments that measure this trait, there is less of a validity concern with this part of the Retail Assessment. With these recommendations, a strategy for content validating the Trainability and People Skills Assessment (based on Section 14C4-5 of the Uniform Guidelines) is below:

Copyright © 2019 Biddle Consulting Group, Inc. • 800-999-0438

Trainability and People Skills Assessment Content Validation Strategy

One of the first steps for content validating a test is defining the skill or ability in terms of observable aspects of work behavior (Section 14C4 of the Uniform Guidelines). For the Retail Assessment, this includes the following six skills/abilities:

Trainability Assessment

1. **Attention-to-detail** (i.e., inspection): Example items include comparing stock numbers (key against a list), isolating incorrect parts or product differences, or similar.
2. **Interpreting and Applying Information:** Interpreting and making basic conclusions from training materials, safety cards, written on-the-job instructions, and policy and procedure.
3. **Numeracy skills:** While most calculating is done electronically, basic stocking/shelving/facing duties are not, as well as other similar duties that may be required when the electronic calculations are unavailable (e.g., making change, processing a coupon, discount).

People Skills Assessment

1. **Problem Avoidance/Solving:** Includes multiple discrete soft skills such as, applying appropriate and professional social skills while interacting with others, adaptability and personal flexibility in various situations, and making concessions and building consensus to achieve work related goals.
2. **Effective Communication:** The ability to interact and work effectively with coworkers, supervisors, and customers, including those with varying socio-economic/ethnic, or other backgrounds.
3. **Customer Centric Focus:** Includes interacting with customers in a helpful, polite, friendly and positive manner. Showing a positive and willing attitude when addressing customer questions or issues, taking initiative and follow-through and attention to appropriate details.

Next, 7-10 qualified job experts (incumbents and supervisors) for the target position[3] should be surveyed using the following questions for <u>each item</u> on the Trainability and People Skills Assessments:

1. Which skill/ability is measured? (All six above listed, multiple selections allowed).

[3] More job experts should be used for target positions that include over 100 incumbents and/or have multiple locations.

Copyright © 2019 Biddle Consulting Group, Inc. • 800-999-0438

2. Does the item measure the skill/ability in a way that represents the target position? (1-3 Rating Scale: 1-not representative, 2-somewhat representative, 3-representative; at least 50% of job experts should assign a rating of "2" or "3").
3. Is the primary skill/ability measured by the item used in the performance of a critical or important work behavior(s)? (Yes/no, with at least 50% of the job experts answering "Yes").
4. Is the skill/ability measured by the item a necessary prerequisite to performance of critical or important work behavior(s)? (1-3 Rating Scale: 1-the skill/ability measured by this item is not necessary for critical or important work behaviors, 2-the skill/ability measured by this item is somewhat necessary for critical or important work behaviors, 3-the skill/ability measured by this item is necessary for critical or important work behaviors; at least 50% of job experts should assign a rating of "2" or "3").
5. Does the skill/ability measured by this item closely approximate an observable work behavior? (1-3 Rating Scale: 1-no, 2-somewhat, 3-yes; at least 50% of job experts should assign a rating of "2" or "3").

Given the above recommendations and strategies, we offer caution about using a criterion-related validation strategy, especially for litigation or pre-litigation (or audit/enforcement settings). There are several technical challenges that arise when considering using a criterion-related validity strategy for assessments like the Retail Assessment. These include range restriction (because of the high turnover typical in these types of positions), criterion unreliability, and a high base rate of applicants who meet the qualifications necessary for performing the job at a minimum level. For these reasons, we advise employers to coordinate with the BCG team for further exploration of the viability of a criterion study.

Score Reporting

Attachment A contains an example candidate score report for the Retail Assessment. The score report provides an overall hiring recommendation and also indicates a candidate's standing separately on each of the three components of the Assessment.

Copyright © 2019 Biddle Consulting Group, Inc. • 800-999-0438

References

Avis, J. M., Kudisch, J.D., Fortunato, V.J. (2002) Examining the incremental validity and adverse impact of cognitive ability and conscientiousness on job performance. *Journal of Business and Psychology*,17(1), 87-105.

Biddle, D. A., & Morris, S. B. (2011). Using Lancaster's mid-p correction to the Fisher exact test for adverse impact analyses. *Journal of Applied Psychology, 96*, 956-965.

Chan, D.,Schmitt, N. (1997). Video-based versus paper-and-pencil method of assessment in situational judgment tests: Subgroup differences in test performance and face validity perceptions. *Journal of Applied Psychology, 82*, 143–159.

Christian, M. S., Edwards, B. E., Bradley, J. C. (2010). Situational judgment tests: Constructs assessed and a meta-analysis of their criterion-related validities. *Personnel Psychology, 63*, 83–117.

John, O. P., Naumann, L. P., Soto, C. J. (2008). Paradigm shift to the integrative big five trait taxonomy: History, measurement, and conceptual issues. In John, O. P., Robins, R. W., & Pervin, L. A. (Eds.), Handbook of personality: Theory and research, 3rd edition (pp. 114-158). New York, NY: The Guilford Press.

Motowidlo, S. J., Beier, M. E. (2010). Differentiating specific job knowledge from implicit trait policies in procedural knowledge measured by a situational judgment test. *Journal of Applied Psychology, 95,* 321-333.

Motowidlo, S.J., Brownlee, A.L., Schmit, M.J. (2008). Effects of personality characteristics on knowledge, skill, and performance in servicing retail customers. *International Journal of Selection and Assessment*, 16(3), 272-280.

Oswald, F. L., Leaetta, M. H. (2010). Personality and its assessment in organizations: Theoretical and empirical developments. In S. Zedeck (Ed.), *American Psychological Association handbook of industrial and organizational psychology* (pp. 153-184). Washington, DC: American Psychological Association.

Schmitt, N., Clause, C. S., & Pulakos, E. D. (1996). Subgroup differences associated with different measures of some common job-relevant constructs. In C. L. Cooper & I. T. Robertson (Eds.), International review of industrial and organizational psychology, (Vol. 11, pp. 115–139). New York: Wiley.

Schmidt, F.L., Hunter, J. (2004). General Mental Ability in the World of Work: Occupational Attainment and Job Performance. *Journal of Personality and Social Psychology*, 86(1), 162-173.

Copyright © 2019 Biddle Consulting Group, Inc. • 800-999-0438

Uniform Guidelines – Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, and Department of Justice (August 25, 1978), Adoption of Four Agencies of Uniform Guidelines on Employee Selection Procedures, 43 Federal Register, 38,290-38,315, referred to in the text as; Equal Employment Opportunity Commission, Office of Personnel Management, Department of Treasury (1979), Adoption of Questions and Answers to Clarify and Provide a Common Interpretation of the Uniform Guidelines on Employee Selection Procedures, 44 Federal Register 11,996-12,009.

US Department of Labor: Employment and training administration (2000), *Testing and assessment: an employer's guide to good practices*. Washington DC: Department of Labor Employment and Training Administration.

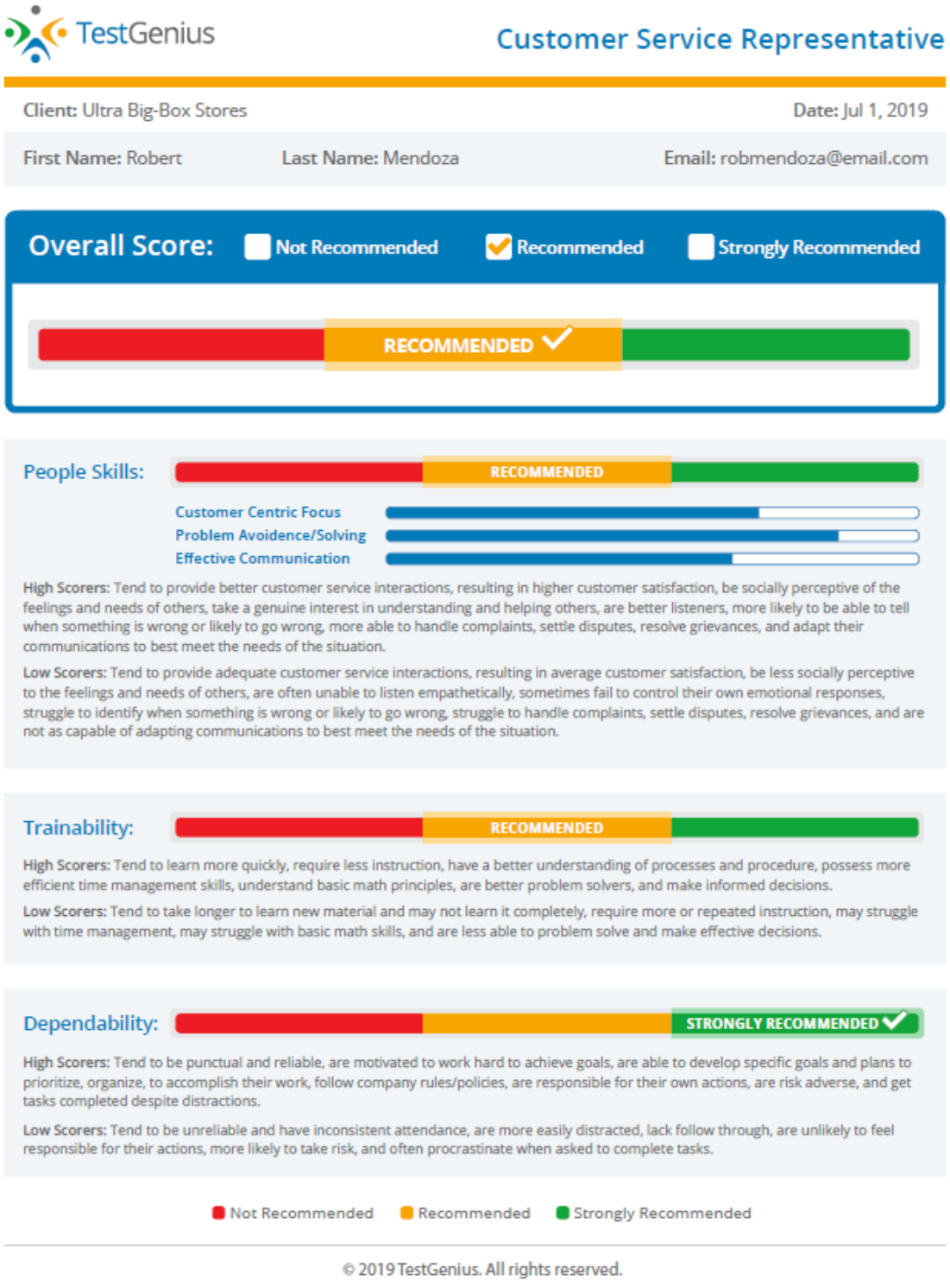
Weekley, J.A., Jones, C. (1997). Video-based situational testing. *Personnel Psychology, 50*, 25–49.

Winfred, Jr. A., Doverspike, D., Muñoz, G.J., Taylor, J.E., Carr, A.E. (2014).The use of mobile devices in high-stakes remotely delivered assessments and testing. *International Journal of Selection and Assessment*, 22(2), 113-123.

Copyright © 2019 Biddle Consulting Group, Inc. • 800-999-0438

Attachment A
Candidate Score Report Example

Copyright © 2019 Biddle Consulting Group, Inc. • 800-999-0438



[1] While the Uniform Guidelines do not formally constitute a set of legal requirements, they have consistently been awarded "great deference" starting as early as the Griggs v. Duke Power Company (401 US 424, 1971) case. They have also been unilaterally adopted verbatim as a legal standard in several cases—e.g., Brown v. Chicago (WL 354922, N.D. III, 1998).

Copyright © 2019 Biddle Consulting Group, Inc. • 800-999-0438

[2] It is important to note that the Uniform Guidelines require that criterion measures consist of *actual job performance*, not ratings of the overall knowledge, skill, or abilities of the incumbents (see Section 15B).

Copyright © 2019 Biddle Consulting Group, Inc. • 800-999-0438