Adaptive Testing:
Adapt and Overcome the Shortfalls of Traditional Proficiency Assessments

Robert "Mac" McLaughlin, Dr. Stephen Gunter
Camber Corporation
Orlando, FL
rmclaughlin@camber.com, jgunter@camber.com

Jeff Pearson
Veterans Benefits Administration
Orlando, FL
jeffrey.pearson@va.gov

ABSTRACT

Most military tests direct outcomes to a simple pass or fail; a Go/No Go model in which the individual must meet specific standards of performance. Another testing approach can assess critical job competencies and current proficiency levels across a continuum, which ranges from entry level to mastery. When implemented for occupational specialties and skill areas for which it is suited, such a model offers significant long-term advantages in maximizing training budgets, and developing skill mastery throughout a warfighter's career.

Specific proficiency levels may be designated as minimum standards for different pay-grades, or for testing similar jobs sharing common competencies at different required levels of proficiency. Unlike a binary pass/fail approach, this identifies specific areas and the degree of remedial training required to meet standards. It also identifies gradual skill decay and offers targeted remediation before it reaches the point of certification failure.

Over the long term, this can reduce training expenses by identifying specific training requirements. Standard remedial training approaches are often very broad, with participants sitting through hours of training on standards, which they may actually meet, waiting for the specific training content in which they are deficient. It may also allow for faster advancement to skill mastery, as targeted remediation means additional training hours are available for skill advancement beyond minimum certification standards. Identifying areas of skill decay provides for a just-in-time training approach that can reduce future test failures and the impact of individuals taken away from their primary job for corrective training.

Adaptive test engines, which use a branching logic to adjust test question difficulty at multiple points during the test, offer an effective means to achieve this outcome. This paper presents an overview of the benefits of a diagnostic testing model, steps required for its implementation, and experiences designing such a test.

ABOUT THE AUTHORS

Robert "Mac" McLaughlin is a former Army Intelligence officer with experience in training development and Human Performance Assessment methodology, particularly oriented to live training and human-in-the-loop process evaluation for military and federal clients. For the past two years, Mac has worked with Camber Corporation as a test lead for diagnostic assessment and skills certification tests for the Veterans Benefits Administration (VBA). He has an MBA in International Business and a BA in English.

Stephen Gunter possesses a Doctorate in Industrial/Organizational Psychology with specialized experience and knowledge in psychometrics and the development and validation of employment tests, certification tests, situational judgment tests, and situational interviews. Stephen has experience leading large-scale, team-based projects; and teaching graduate level courses on the theory, measurement, and application of employment and certification tests. He is responsible for the development of the diagnostic test model, certification test blueprints, and the accuracy and psychometric evaluation of VBA tests.

Jeff Pearson is a retired Coast Guard officer and currently serves as a Human Performance Technologist and instructional design professional for the VBA. Jeff possesses a Master of Science degree in Instructional and Performance Technology and has over 25 years of experience developing high-performing individuals, teams and organizations. Jeff has collaborated with individuals and teams during the development of several innovative models, including a new Competency-based Training System (CBTS) that leverages diagnostic testing to determine performance deficiencies and prescribe remedial training specific to the needs of each VBA employee.

Adaptive Testing:
Adapt and Overcome Shortfalls of Traditional Proficiency Assessments

Robert "Mac" McLaughlin, Dr. Stephen Gunter
Camber Corporation
Orlando, FL
rmclaughlin@camber.com, jgunter@camber.com

Jeff Pearson
Veterans Benefits Administration
Orlando, FL
jeffrey.pearson@va.gov

INTRODUCTION

Most military tests direct outcomes to a simple pass or fail; a Go/No Go model in which the individual must meet specific standards of performance. Standard remedial training approaches are often very broad, with participants sitting through hours of training on standards which they may already meet, waiting for the specific training content in which they are deficient. This is an inefficient use of training dollars and time and binary outcome testing models offer limited data to direct targeted, individualized remediation. Another testing approach can assess critical job competencies and current proficiency levels across a continuum, ranging from entry level to mastery. When implemented for occupational specialties and skill areas for which it is suited, such a model offers significant long-term advantages in maximizing training budgets and developing skill mastery throughout an individual's career.

Adaptive test engines, which use a branching logic to adjust item difficulty at multiple points during the test, offer an effective means to achieve this outcome (Crotts, Zenisky, Sireci, & Li, 2013; Lord, 1980; Wainer, 1993). Specific proficiency levels are designated as minimum standards for different pay-grades, or for testing similar jobs that share common competencies at different required levels of proficiency. Unlike a binary pass/fail system, the adaptive test engine identifies specific areas and the degree of remedial training required to meet standards. It may also allow for faster advancement to skill mastery, since targeted remediation means that available training hours shift toward skill advancement beyond minimum certification standards. Used in an ongoing program, it identifies gradual skill decay and offers targeted remediation before it reaches the point of certification failure. Identifying areas of skill decay provides for a just-in-time training approach to reduce future test failures and the impact of individuals pulled off the line for corrective training.

The Veterans Benefits Administration recently developed, and is currently testing, a Competency-Based Training System (CBTS) for the Vocational Rehabilitation & Employment Service in an effort to determine the performance needs of each Vocational Rehabilitation Counselor (VRC), remediate identified deficiencies with individually targeted training, and establish a career roadmap and digital credentialing process. The engine that drives VBA's CBTS is a series of small (fewer than 25 items), adaptive diagnostic assessments that assign the training required for VRCs to have the necessary Knowledge, Skills, and Abilities (KSAs) to perform their occupational functions at an appropriate level of proficiency. A unique assessment delivery interface provides immediate feedback upon assessment completion, including an overall score and the required remedial training. This data is transferred to VBA's Talent Management System, which automatically adds the required training to the individual employee's learning plan.

Achieving an accurate assessment of proficiency and providing truly targeted remediation require assessment items to be tightly linked to KSAs, terminal and enabling learning objectives, and competencies. Purposeful linking of these components has produced a matrix that (1) synchronizes all elements of the CBTS, ensuring test and training validity, and (2) provides for a systemic approach to lifecycle maintenance.

The adaptive diagnostic assessment approach the authors present in this paper provides a more efficient alternative to the traditional assessment and training assignment methods many organizations currently employ. The transitional journey to an adaptive diagnostic assessment model is certainly "a road less traveled" and successful implementation requires a good change management strategy. The return on investment can be significant. At a minimum, organizations can expect to achieve a reduction in overall training expenditures (eliminates redundant and ineffective training). Full implementation will likely yield the additional benefits of (1) increased proficiency, (2) increased productivity, (3) better visibility of an individual's unique training requirements, and (4) increased professional development opportunities.

LINEAR VS. ADAPTIVE TEST MODELS

To understand the development of the adaptive, diagnostic assessment, one should have an overview of linear and adaptive test models and the inherent strengths and weaknesses of each. Conventional linear tests, in either a paper and pencil or electronic test format, present the same test items to all examinees. There are important strengths and weaknesses of this approach. The strengths of the linear testing approach include: (1) ease of construction and administration and (2) can be used for testing programs that require the test to be constructed at a specific level of difficulty (e.g., certification test). However, this leads to an important weakness in linear tests. Since the items are fixed, so too is the specific difficulty level of the test. Thus, if a group of highly proficient individuals take a test composed of only moderately difficult items, then the test will be relatively easy for everyone. Conversely, if a group of novice individuals take a test composed of moderately difficult items, then the test will be relatively difficult for everyone. Consequently, the main weakness of a linear test is that it may not be able to estimate the proficiency of all individuals (from novice to expert), unless the test is sufficiently large or the test measures a relatively small content domain.

When analyzing the challenges posed by the diagnostic assessment project, critical requirements were to develop a test that reliably covered all areas of KSA content, assessed the examinee's current level of proficiency in major job competencies, and direct the examinee to specific, relevant training courses to remediate skill deficiencies. The test would serve as a preparatory and development tool for examinees as they prepared for a high-stakes, skills certification test that is taken after approximately two years in the position. It was also to serve as a tool for evaluating post-certification skill decay; offering targeted remediation for individuals who might not regularly exercise all required job skills due to specialized assignments. Thus, another consideration was that examinees might take the diagnostic test more than once, and the items and test forms should offer sufficient variability to minimize item exposure.

Development of Adaptive Testing

The need to more accurately evaluate and place examinees across the full proficiency scale gave rise to adaptive testing models. Adaptive tests have several advantages over linear tests. First, adaptive tests can more precisely estimate an individual's current level of proficiency in a KSA or competency (Lord, 1980). Second, adaptive tests can often be much shorter than linear tests because they more strategically administer items that more closely represent an individual's current proficiency level (Wainer, 1993). Third, adaptive tests can control item exposure better because only items that best represent an individual's proficiency are administered. Adaptive tests, by necessity, are administered electronically. They are categorized as item-level adaptive or multistage adaptive.

In an adaptive test, every item has an established level of difficulty (e.g., easy to very difficult). The best method of determining the item's difficulty is to pilot items with a sampling of the intended audience and calculate difficulty based on individual item performance results. Without the luxury of being able to pilot items, the difficulty can be derived from an analysis of cognitive tasks; ranking the relative complexity of mental functions required for the task, adjusted by the difficulty of the content area or KSA to which the item is written. Fine-tuning of item difficulty is subsequently adjusted based on actual item performance during test administrations.

The difficulty of the items that are administered during the test changes, based on the examinee's performance. The difficulty of test items may increase or decrease as the examinee answers items correctly or incorrectly. The accuracy of the evaluation increases as the test narrows the examinee's demonstrated range of knowledge to a more precise point on a proficiency scale (Lord, 1980). The result need not be a simple pass or fail, but can offer a more detailed understanding of how well the examinee understands specific content areas at this stage in their educational development.

Item-Level Adaptive Tests

Item-level adaptive tests, also known as computerized-adaptive tests (CATs), have been widely researched. The selection of each new item is based on the results of the preceding items that have been administered. A routing decision is made when the item is submitted. The dynamics of item selection mean that once an item response has been submitted, the examinee is unable to go back to items as may be possible on a linear test.

The test typically begins with an item of average difficulty because, in many cases, the proficiency level of the examinee is unknown. Because the examinee's proficiency is often unknown many assume that the examinee is "typical" and, therefore, has average proficiency. If the examinee gets the first item right, the CAT selects a subsequent item ranked slightly harder. If the examinee gets the item wrong, the CAT selects a slightly easier item. While the difficulty of the items that are presented may fluctuate up or down as the test continues, it is expected that, on the basis of the response patterns to the items, the difficulty level of the items that are presented will eventually stabilize at a specific range on the proficiency scale that best represents the proficiency of the examinee.

One issue with CATs is achieving balanced test content (Luecht, 2005). Some tests measure a wide range of KSAs and achieving adequate measurement across the wide-ranging content domain can be challenging. Does the adaptive test engine provide for adequate coverage of all content areas from the test specification plan? Content-related validity supports and clarifies test interpretations and is a fundamental requirement for defending educational tests for their intended purposes. Some content areas are inherently easier or more complex than others within the same test. Examinees performing particularly well or particularly poorly may be routed along a proficiency track that fails to adequately evaluate their knowledge of all content areas if item difficulty alone is the sole factor determining item selection (Luecht, De Champlain, & Nungester, 1998). There are alternative approaches that take into account the need to balance content and provide adaptation. One such approach is multistage testing.

Multistage Adaptive Tests

Multistage adaptive tests (MSTs) take a different approach, presenting sets of items (e.g., three to six items) and adapting after each group. These small groups of items are built to a specific level of proficiency and are the building blocks of a MST. The nature of the items within a stage may be content-specific, content-balanced or stimulus-dependent according to different design strategies and desired effects. We present these three design strategies in the next section.

The examinee begins a MST with a set of items at an intermediate proficiency range. Performance on this initial routing stage provides a first, rough estimate of the examinee's ability and routes to a second set of items which are either easier or more difficult (see Figure 1). MSTs have fewer adaption points than item-level adaptive tests, and the routing is based on the cumulative performance of the set of items rather than of each item. This method typically requires less computing power than item-level tests, and offers other solutions to the data management challenges of balancing test content.

Common MST Stage Designs

Figure 1 presents a common MST design. Each grey box represents a set of items of a common proficiency level, supporting content coverage according to the test specification plan. These sets are called modules, and the total score for each module is used to route someone to the next module of easier, same difficulty, or harder items. The point at which the test makes a routing decision is called a stage. The example in figure 1 can be described as a 1-3-5-5 MST configuration. This notes that there are a total of four stages to the test and how many proficiency levels are available at each of the stages.

There are many different, possible configurations. Generally, the most important characteristic of any configuration is the number of items in the entire test. The number of proficiency levels that the test designer requires will determine the minimum number of stages. Using Figure 1 as an example, because there are five proficiency levels, or intervals, the minimum number of stages is three. Three stages allow for an examinee to be given items within proficiency level 1 or 5. This will help ensure a more accurate estimation of one's proficiency.

Figure 1 depicts five different proficiency levels, with one as the lowest and five as the highest level of relative item difficulty. This example shows the routing path of a high-performing individual, depicted by the blue arrows. The individual routes to a more difficult interval based on stage one results. He or she remains at the same level based on stage two results, but then routes to the highest proficiency level based on the results of stage three. The individual's final evaluation is high on the proficiency scale, noted here by the green box.

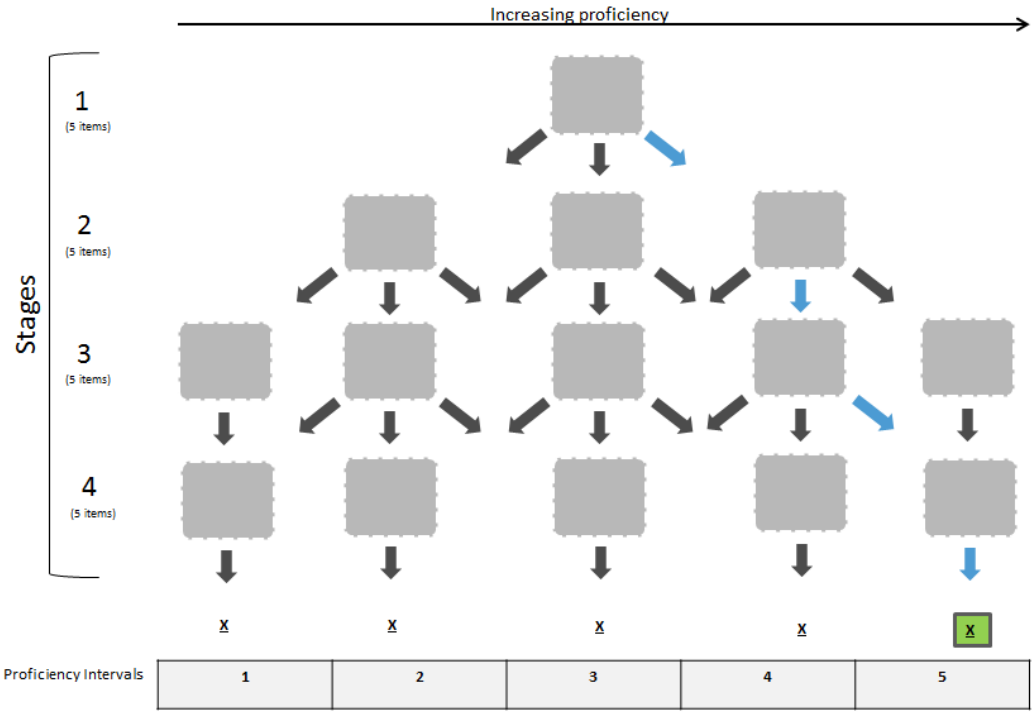

Figure 1. Sample MST Routing Diagram

Note that MST routing does not allow an examinee to skip a proficiency level, such as routing directly from a proficiency level of three to a proficiency level of five. Proficiency at level four must first be demonstrated, providing a more reliable evaluation that what might otherwise be a combination of limited knowledge and some lucky guesswork. A characteristic of MST that is a blend of a linear test and an item-level adaptive test is that an examinee can go back, review, and change answers to items within a module. However, once an examinee is routed to another module they cannot go back and review those items. For example, as shown in Figure 1, if an examinee is completing the five items in stage 2, then he or she cannot go back are review the five items in stage 1.

Content-specific modules contain sets of items within a particular content area (e.g., vocabulary) and seek to balance overall test content by presenting content-specific modules in each adaptive level of a test. The content-specific modules are selected and presented according to the current proficiency level assignment. The test adaptive engine does not route to a new proficiency level until all items within a content-specific module have been answered. The decision to route to a lower, same, or higher proficiency level can be based on the total score for a module. For example, the decision could be that if an examinee gets 2 or fewer items right out of 5 then they are routed to a lower proficiency level module. If an examinee gets 4 or 5 items right, then they are routed to a higher proficiency level module.

Depending on the complexity of the adaptive engine, it is possible to build a test that adapts content-specific sets independently of one another (Keng, Ho, Chen, & Dodd, 2008). In this format, an individual's performance results may increase difficulty level in one content area while lowering it in another. This provides for a very precise report of proficiency and performance in specific KSAs, not just a general rating across all of them as in a pass/fail test format. The data management of such a model could become challenging, but the test would accurately adapt to measure independent proficiency levels within each content area of the test.

Content-balanced stages present a group of items selected from across the test content areas (e.g., vocabulary, verbal analogies, and reading comprehension). Test adaption occurs at the completion of the content-balanced group, instead of after the completion of a series of smaller, content-specific groups as described previously. This provides for a simpler programming model, which then routes to a new content-balanced stage of items all at the same proficiency level. Managing possible contextual conflicts may be easier in a content-balanced model than content-specific, although the accuracy of evaluation within individual content areas may not be as accurate.

Stimulus-dependent stages consist of a set of items that all relate to the same scenario, problem or set of facts ("testlets"; Wainer & Kiely, 1987). They may be used to more thoroughly explore some complex concepts or situations. Individual items within the stimulus-dependent stage may be tagged to specific content areas as appropriate. Test designers may use these different structures to balance test content or to create different cognitive challenges, however it is not common to see different stage structures within the same test.

MST Benefits and Challenges

Multistage adaptive tests offer several advantages as well as some disadvantages. Test adaption is based on cumulative performance on a set of items, easing some of the programming and data management burdens. The capability of designing and building test modules before administration allows greater control over test assembly and test form quality control. This can be very helpful in tests with many content areas or with complicated cross-classifications of content. Developers may check the items grouped in each module in detail to make sure that all test specification requirements are met with an appropriate representation of different cognitive tasks within each content area. This all helps to increase the probability that examinees at the high or low end of the proficiency scale receive appropriate items across all content areas in the test specification plan.

There may be little, if any, discernable difference to the examinee from an MST or a CAT, depending on the design of the MST. Since MST's adapt after a module is completed, it is possible that the test software will allow an examinee to go back and review answers for that module before submitting them. Once submitted and the test has routed to a new module, the items may not be revisited or changed. Stimulus-dependent and content-specific modules would have a different feel to the examinee than independent CAT items, as they would be clearly grouped toward a common scenario, content area or set of facts.

The design and quality assurance (QA) review of test modules helps to reduce the likelihood of contextual conflicts, in which clues or the answer of a test item may have previously been given away. With the items in each specific module predetermined, test designers may examine the different combinations of all items within the modules to reduce dependencies among the items and other context effects, which might influence test results. It does not remove this risk completely and searching for these conflicts is a regular part of item bank maintenance as new items are developed and old items revised.

There are a few limitations posed by MSTs. Construction of MSTs typically requires more work on the part of item writers and test designers than item-level CATs. A considerable number of items is required for a MST to fill multiple stages and proficiency levels. If we assume that each module in Figure 2 has five items, then the entire MST that an examinee completes is 20 items. This means that the entire configuration will require 70 items. In order to minimize the likelihood that examinees will see the same items in and across each module and to enhance test security, MST designers can create multiple versions of each module, which contain common and unique items that meet the proficiency and content requirements. This will require a significant number of items to be constructed and included in the item bank. If, for example, a MST designer wanted to construct at least four alternate modules at each proficiency level and stage, then the item bank would require 280 items.

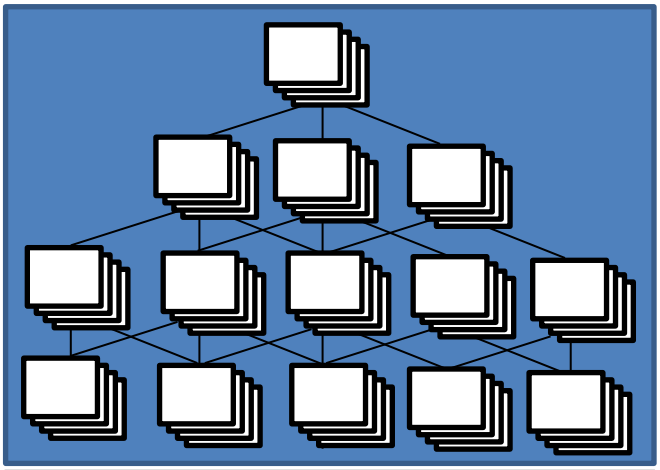

Figure 2. Diagnostic Assessment MST Model

The modules may be assembled prior to test administration or developed "on-the-fly" by a computer algorithm and randomly assigned to examinees. In the instances that there are repeat examinees, the computer program can randomly assign a module at a specific proficiency level that the examinees has not been administered previously.

It is not uncommon for the content of a test to change across time. Regulations, references, or knowledge required for the job may change and, therefore, requires that the relevant items be reviewed and revised. The benefit of having multiple assembled modules is that only the module(s) that contain the item(s) that must be reviewed and revised will need to be taken offline. The modules that do not contain that affected item(s) can still be administered to examinees. Item revision within pre-assembled and "on-the-fly" assembled modules is addressed by removing the module(s) with the revised items from the admin queue until it is considered in relation to other items within and across the modules (i.e., contextual conflicts, clues, or answers).

DEVELOPING THE DIAGNOSTIC ASSESSMENT

Taking current work and research on adaptive testing into account, our team determined that a multistage adaptive test model was the best solution. A design team, composed of testing experts from Camber and policy decision makers from VBA, collaborated together to determine the diagnostic assessment specifications and policies. The design team studied the Massachusetts Adult Proficiency Test (MAPT) for Reading (Sireci, Baldwin, Martone, Zenisky, Kaira, Lam, et al., 2008) as a basic model of a proven and validated multistage adaptive test. A number of decisions remained, however, in establishing the basic groundwork upon which the test would be built.

The diagnostic assessment was determined to support an already-existing skills certification test that determines whether incumbents have attained a minimum level (i.e., "journey-level") of technical knowledge and job qualification. To achieve "journey-level" expertise VRCs complete basic "new counselor" training and spend an additional 18-24 months in their job position acquiring on-the-job experience. The employees are expected to have a high level of technical proficiency within the position, able to operate with minimal supervision, and make few errors for all routine challenges which might occur on the job. They were not expected to be expert in all content areas or be familiar with all uncommon or highly specialized challenges they might encounter in rare circumstances, but the examinees should know, for example, which references to consult to work through such challenges.

A major policy decision that was made by the design team was that the diagnostic assessment must contain any KSAs that are included in the skills certification test. Additional KSAs could be contained in the diagnostic assessment, however. The determination of the full range of KSAs that would be included in the diagnostic assessment is as follows. The KSAs, or content areas, were identified and refined through a multistage process. Job tasks were collected by observation and interviews of incumbents, then a range of broad competencies and specific KSAs were developed to support the job tasks with the help of incumbents and supervisors. Independent focus groups verified and refined the KSAs and linked them to the tasks. Field surveys were then conducted, in which incumbents and supervisors rated each identified task and KSA according to the frequency it was performed on the job, its relative difficulty to acquire or perform, and the criticality (or severity) of the consequences if someone did not perform a task at a minimally acceptable level or have a minimally acceptable level of the KSA. A composite of these three scores was used to rank the overall importance of the tasks and KSAs to the job, serving as a reference for the test design team to select the specific KSAs to be tested.

Basic Test Structure

The breadth of information required for the diagnostic assessment test was considerable. One of the main decisions was to split content into a number of smaller tests (i.e., "testlets"). Each testlet focused on a specific competency, the content of which was defined by a number of KSA descriptions from the test specification plan. The goal was to create testlets that would each take examinees approximately 45 minutes or less to complete. The testlets could be taken by the examinees at their leisure in any order they wished. Once all diagnostic testlets were complete, a full evaluation could be presented on all competencies toward which they were working for certification, showing their proficiency levels across each KSA with recommended remedial training specific to those shortfalls.

The design team elected to use a structure of content-balanced stages within each test. The 30 content areas, or KSAs, fell across 5 major competencies in the overarching diagnostic test specification plan. The team divided these into seven testlets, each covering between three to six KSAs. Those KSAs that ranked higher on the composite difficulty-frequency-criticality survey were grouped in testlets with fewer total KSAs, so that these content areas were more thoroughly tested with a higher number of associated questions.

Based on historical data from the skills certification test, the team estimated a maximum test length of 25 questions could be completed in the desired 45-minute window. Several adaption points would be required in every testlet to increase the evaluation accuracy. The team determined that the testlet models should have a minimum of three, with no more than five, adaption points depending on the specific number of items and content areas within each. Keeping a goal of no more than 25 items per test in mind, the tests were mapped to present between four to six stages.

The design team proposed a proficiency spectrum that ranged from entry-level personnel to the journey-level audience of the certification test. The scale also reached into skill mastery beyond journey-level knowledge (Figure 3). Since the diagnostic test was intended to serve as a prep test and training remediation for the certification test, the test audience was considered to be largely entry-level and journey-level proficiency, with the test items focused most heavily in the intermediate range. The design team determined that the diagnostic assessment would map proficiency across five broad categorizations of entry, low-intermediate, intermediate, high-intermediate and journey level. These levels were mapped across a point scale to use as a metric in categorizing the relative difficulty of each item for building adaptive stages. Adding additional depth to the item bank at the skill mastery level, will allow the test to live on after basic certification as a "sustain and build" tool for honing advanced skills and identifying areas of subsequent skill decay.

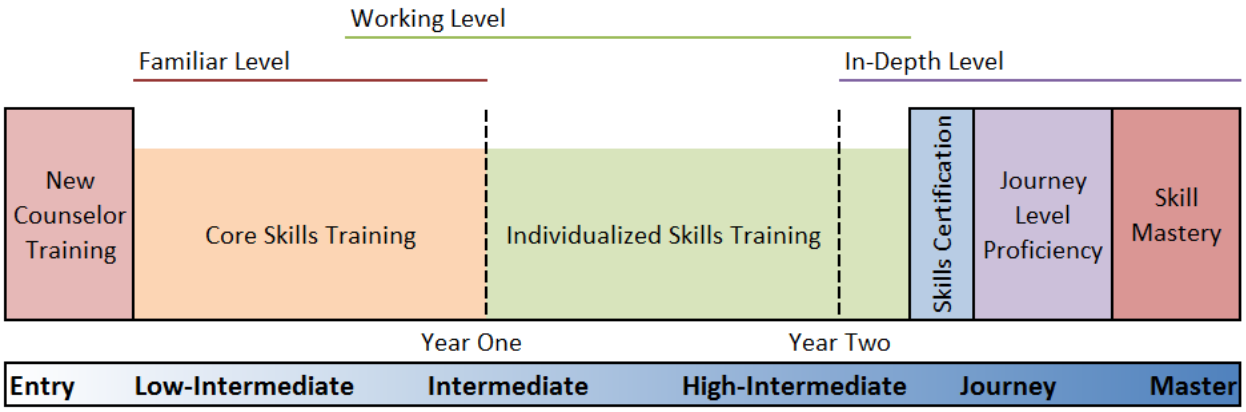

Figure 3. Competency level spectrum mapped to examinee professional development track

Test Adaptive Engine

Rather than utilizing a pre-assembled module design, the design team worked with programmers to build an adaptive engine and automatic test assembler (ATA) that would build each new module on-the-fly; randomly drawing from item "buckets" fitting specific test design and proficiency criteria. Test items were assigned a series of different categorization markers. Each item was linked to a specific KSA on the test specification plan, critical content meta-tags, remedial training specific to the item, and one of several "facets" detailing cognitive functions or sub-areas of knowledge specific to each KSA. Items were further tagged with a difficulty rating from 1 to 100, describing where it should fall on the proficiency scale. Initial difficulty ratings were based on the relative difficulty of cognitive tasks and key content, to be more accurately refined based on ongoing assessment engineering research and on each item's performance results in test administrations.

When defining the specifications of each diagnostic testlet, the designers selected the specific KSAs that would be covered, the number of adaptive stages, and the number of items selected within each stage. At the beginning of each testlet's administration, the engine first builds the initial routing module. It randomizes the order in which the KSAs will appear in that module. It then randomly selects one item which is both assigned to that first KSA and falls within a difficulty rating range according to the level of the current stage. The engine then selects a random item assigned to the next KSA meeting difficulty level guidelines, and so on, until the been selected. By setting the number of items per stage as a multiple of the number of KSAs within the test, the designer was assured of content-balance, as each KSA would receive equal coverage in the random items pulled for the stage. Once an item was selected, it was taken out of the pool for the remainder of that administration.

For the diagnostic testlets, the team decided that a correct score of 71% or better for all items within a given stage would route the examinee up one level. A correct score of 39% or less would route the examinee down one level, while scores in a given stage ranging from 40% to 70% would keep the individual at the same level for the next stage. This score was used solely for routing to the next stage within the testlet. The engine kept track of scoring for each item completed, as the feedback report would identify the examinee's relative position on the proficiency spectrum as well as performance within specific content areas.

Figure 4 displays a routing diagram for a four-stage diagnostic testlet. The adaptive engine builds the initial routing module from items whose assigned difficulty rating falls within a range defined for the intermediate level. Based on the examinee's performance on those items, the engine routes the next stage to low-intermediate, high-intermediate or remains at the same level. The order of KSAs is randomized again and new items are selected to meet KSA and difficulty level parameters. By the third stage, examinees performing at the very high or low ends of the spectrum might unlock the entry or journey level of item proficiency. Each testlet was designed with a minimum of four stages, with the fourth or later stages considered to fine tune and confirm previous routing paths.

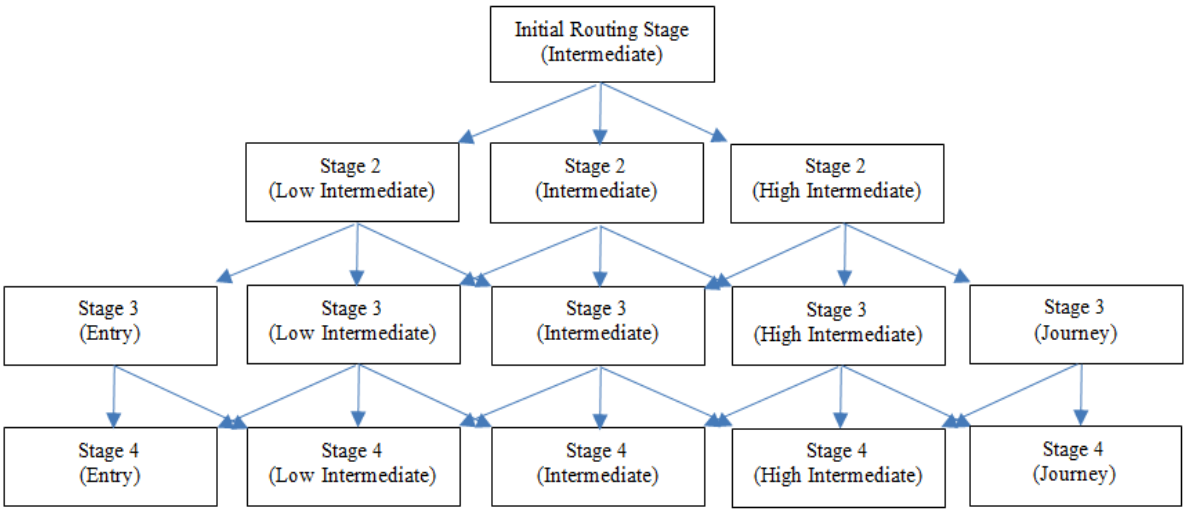

Figure 4. 1-3-5-5 Four-stage testlet routing diagram

With a large enough pool of test items, on-the-fly stage development offers more variations of test forms, addressing item exposure concerns for individuals taking some diagnostic testlets multiple times. The decision not to use pre-constructed modules meant less ability to work around any context related problems among the test items. On-the-fly stage development, even within the current selection guidelines, does not offer the same level of security against contextual conflicts. It was considered an acceptable risk for the initial version of the diagnostic assessment. Future versions of the adaptive engine could potentially address this concern by the use of the facet categorization or meta-tagged content as item discriminators.

Item development required a heavy concentration of items across the intermediate levels of proficiency since the largest number of routing permutations remained in the middle of the spectrum. Smaller pools of items are needed at the extreme ends of the scale since even very high or low performing examinees first route through intermediate level stages. Coverage of all content areas is still a requirement, challenging item writers to develop test items in all KSAs that would meet proficiency level requirements. As noted previously, content representation is not the same as content quality and pre-screening items for content validity and quality, is required. In the current version of the diagnostic assessment, items rejected by content reviewers are sent back to writing workshops.

Feedback Report

Individual feedback reports (Figure 5) are saved for each examinee and available for recall and review. The overall report for each testlet displays the examinee's relative position on the proficiency scale of that specific job competency. The report shows specific performance results within each KSA and hyperlinked tabs to display results specific to KSA facets (sub-sections of KSAs) or relevant error codes that would have been assigned by the VBA's quality review team.

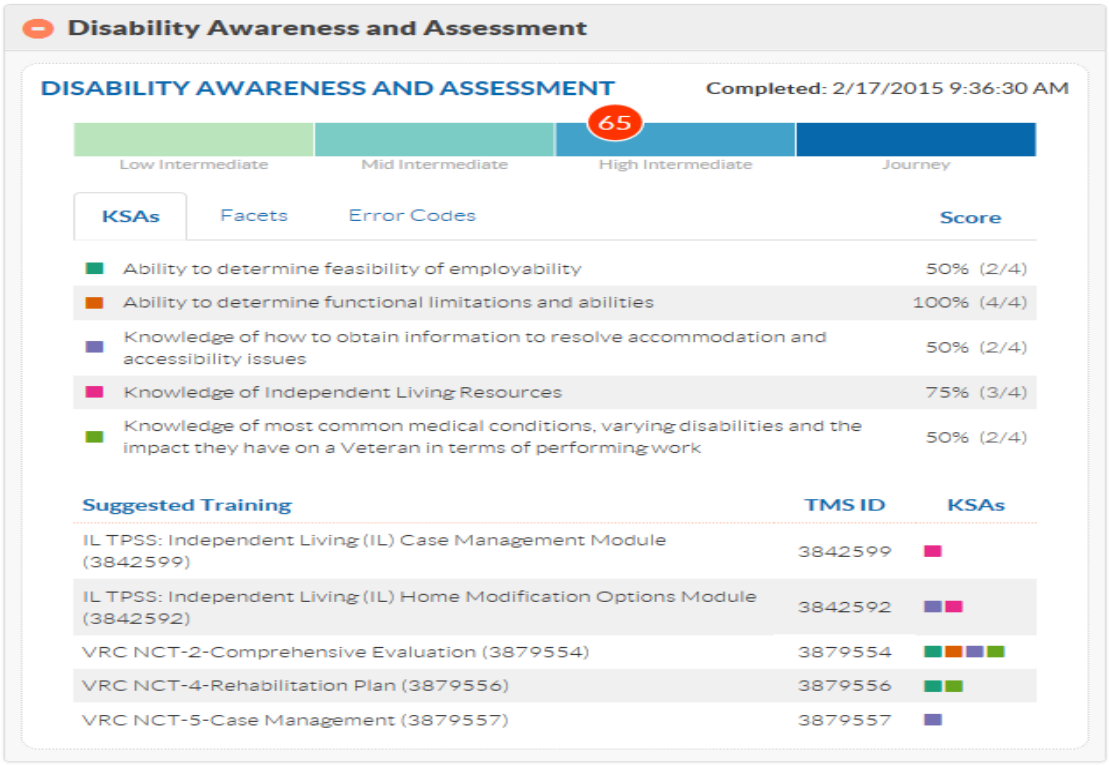

Figure 5. Feedback report

The report also displays recommended training courses, based on the overall test results and specific items missed by the examinee. Color codes relate training courses to the content area(s) they cover. Originally conceived of as a prep test and remedial training tool leading examinees to a high-stakes skills certification test, the client envisioned examinees taking some testlets every 9 to 12 months to gauge development progress in competency areas requiring additional training.

Implementation

The diagnostic assessment testlets have undergone internal testing and preliminary testing by training team members and a field advisory committee. In addition, planning is underway to conduct a pilot test in a select number of regional offices. If well received, the proposed plan is for wider adoption, with participating regional office personnel using the diagnostic assessment testlets to guide personal training and development programs instead of the traditional, annual broad-based national training curriculum. Two additional diagnostic assessments are now in development for the VBA; one addressing the skills and training of Pension Management Center Veteran Service Representatives (PMC VSRs) and another which is focused on leadership and coaching skills for team supervisors.

Badging and Job Profiles

A consideration which was not part of the initial test concept, but which took shape during test development, is the concept of examinee skill "badging" and the development of job-specific proficiency profiles. Higher education institutions and professional organizations have increasingly been using digital badges as means to motivate,

demonstrate and validate learning and development. While individuals may enjoy an array of digital badges associated with their professional online profile or signature blocks, ultimately the value is less the image of the badge as it is a link to a verified electronic record of performance or accreditation.

VBA and the test design team are working on a strategy of electronic badging for various performance results on the diagnostic assessment test. Early discussions focused on a simple "journey-level" achievement badge versus a set of ranked badges related to proficiency level within each competency. Further discussion led the team back to a core detail that has been known throughout the administration of the skills certification program; that some competencies and KSAs require an in-depth knowledge, while others really only require a basic working familiarity to perform job functions successfully.

A set of tiered proficiency badges is being considered for each testlet under a working conceptualization of a bronze, silver, gold "Olympic" model. All require a demonstration of competency; bronze is not a default achievement for simply taking the test, and gold level would indicate a perfect or near-perfect score. These could be combined with a detailed categorization by subject matter experts, to define a set of job profiles in relation to badge levels and competency areas. Closely related jobs, or the same job at different pay-grades, may take the same selection of testlets, but have different minimum badge level requirements reflecting the depth of knowledge required in the respective KSAs. Specialized job team-members may test on a selection of common-knowledge testlets, as well as one or two testlets focused on competencies and KSAs unique to their role.

As the diagnostic assessment item bank matures, difficulty levels are confirmed through item history and the concept validated through testing, this concept could supplant the current skills certification test model. It could also better evaluate the required job knowledge for different specialized jobs and pay-grades by different minimum badge-level requirements in specific competency areas. This concept has been met with considerable excitement and is expected to be an area of research and testing in FY 2016.

SUMMARY

Efficient training is a challenge faced by all organizations. A linear pass/fail test with broad-based training remediation is a simple approach, but results in considerable inefficiencies and wasted time and resources. Advancements in technology and testing methodology offer a means to evaluate individual job proficiencies and develop personalized remediation strategies that isolate the specific content areas and degree of training required to meet established organizational standards. Adaptive tests are not a sole solution to increase organizational training efficiency, but they can provide detailed, personalized performance data critical for training managers with limited budgets and training hours, who are wrestling to address the challenges of job knowledge standardization, skill decay and professional development.

The competency-based training system, leveraging the power of a multistage adaptive assessment, has the ability to remediate performance deficiencies quickly with individualized, targeted training. It provides data to accurately assess the quality of existing training and determine when new training is required. It uses an automatic test assembler to deliver test items randomly within specific difficulty and content guidelines, minimizing the risk of test compromise. Additionally, adaptive diagnostic assessments provide flexible delivery alternatives for the target audience. Employees can complete all of the testlets associated with a specific competency, in any order, and as their individual work schedules allow. Employees may retake specific testlets at a later date to evaluate their personal development efforts after remedial training courses and to monitor skill decay over time.

A multistage adaptive model is not necessarily the best assessment tool for every organization. Organizations must select an assessment model based on several factors including (1) organizational test delivery preference, (2) test security requirements, (3) number of employees, (4) available development time, (5) potential return on investment, (6) scalability, (7) ease of maintenance, (8) current state of information technology infrastructure, and (9) funding (for design, implementation and maintenance).

Adaptive diagnostic assessments are a powerful tool that can vastly improve workforce capability when correctly employed. The models presented in this paper are a few of the viable options that exist and are worthy of consideration.

REFERENCES

Crotts, K. M., Zenisky, A. L., Sireci, S. G., & Li, X. (2013). Estimating measurement precision in reduced-length multi-stage adaptive testing. *Journal of Computerized Adaptive Testing, 1*(4), 67-87.

Keng, L., Ho, T. H., Chen, T. A. A., & Dodd, B. G. (2008). A comparison of item and testlet selection procedures in computerized adaptive testing. Paper presented at the meeting of the *American Educational Research Association & National Council on Measurement in Education*.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Luecht, R. M. (2005). Some useful cost-benefit criteria for evaluating computer-based test delivery models and systems. *Journal of Applied Testing Technology, 7*(2).

Luecht, R. M., De Champlain, A., Nungester, R. J. (1997). Maintaining content validity in computerized adaptive testing. *Advances in Health Sciences Education, 3*, 29-41.

Sireci, S. Baldwin, P., Martone, A., Zenisky, A. L., Kaira, L., Lam, W., et al. (2008) *Massachusetts Adult Proficiency Tests Technical Manual, Ver. 2*, Amherst, MA: University of Massachusetts Amherst.

Wainer, H. (1993). Some practical considerations when converting a linearly administered test to an adaptive format. *Educational Measurement: Issues and Practice, 12*(1), 15-20.

Wainer, H., & Kiely, G. L. (9187). Item clusters and computerized adaptive testing: A care for testlets. *Journal of Educational Measurement, 24*, 185-201.