Implementing Assessment Engineering in the Uniform Certified Public Accountant (CPA) Examination

Matthew Burke, Ph.D.

Richard Devore, Ed.D.

Josh Stopek, CPA, MBA

Abstract

This paper describes efforts to bring principled assessment design to a large-scale, high-stakes licensure examination by employing the frameworks of Assessment Engineering (AE), the Revised Bloom's Taxonomy (RBT), and Cognitive Task Analysis (CTA). The Uniform CPA Examination is practice-oriented and focuses on the skills of accounting. In preparation for a Practice Analysis, the authors are revising skill definitions to support the development of construct maps. Construct maps are the means by which distinct levels of proficiency are represented along a continuum and linked to tasks used to provide evidence of proficiency. A modified CTA is employed to better understand the "moving parts" of existing complex performance exercises used in the examination. The RBT is used to provide a hierarchical structure to the cognitive skills underlying task performance. Construct maps provide the basis for a new type of test blueprint that incorporates psychometric concerns into the test development process. This research focuses on tackling a critical step in AE; writing skill statements to populate the construct map.

Keywords: Assessment Engineering, Construct Maps, Skill definitions

The Uniform Certified Public Accountant (CPA) Examination is one of three requirements for licensure as a CPA. The purpose of the examination is to provide assurance to State Boards of Accountancy (BOA) that candidates have obtained a satisfactory level of knowledge and skills to successfully protect the public interest from improper accounting practices (http://www.aicpa.org). The examination is made up of multiple choice questions (MCQs), task based simulations (TBSs), and constructed response items (CRs) that are each designed to assess particular aspects of the knowledge and skills deemed necessary for a candidate to successfully practice in the profession. It is therefore a primary concern that the particular knowledge and skills tapped by questions in the examination be clearly specified and articulated so that the exam adequately functions for its intended purpose.

Traditionally, the focus of test development and construction has been to ensure that content requirements are met. Less emphasis has been placed on identifying and describing the underlying cognitive skills that test items tap into. In the past, developing statements of the skills and describing them has been done through practice analysis (PA). The PA uses focus groups and a survey to gather information about the particular knowledge and skills deemed critical to the practice of accounting. The PA is done every few years and provides a valuable source of input from informed stakeholders. The most recent PA was completed in 2008. The current investigation presented in this paper is directed towards expanding and updating skills definitions for the CPA examination in a new way that allows for more input from measurement specialists and cognitive scientists. The focus of the discussion is to clearly explicate *why* and *how* we are investigating making changes to the Content Specification Outline (CSO) and the Skills Specification Outline (SSO) to update and make full use of our skill definitions so that they better meet the end uses of the different stakeholder groups.

Background

When the Uniform CPA examination was computerized, realistic simulations of accounting tasks were included as items. These simulations opened new opportunities to measure the skills of accountancy in ways that were not possible previously. Naturally, skill definitions were developed, items were created and administered, empirical data were analyzed, and the simulations were refined. As a result of the success of the item type, the simulations now take on a larger role in scoring the examination.

This paper describes the process used to revisit and refine the definitions of skills as stated in the current "Content and Skills Specifications for the Uniform CPA Examination" (http://www.aicpa.org). Full descriptions of both the original and current skills definitions are located on the AICPA website. (See references for the full link to the original and current versions of the SSO.) To begin, a brief description of the two previous versions of the skill definitions is provided so that the motivation for developing new skill definitions and the method used to provide the skill definitions are obvious. Afterwards, a description of the method used to refine the skills definitions is presented.

The skill definitions have evolved through rounds of development and revision by way of two Practice Analyses (PAs) and some focus groups with SMEs. The first PA was done in 2000, and subject matter experts (SMEs) focused on creating descriptions of skills that were amenable to testing in the new electronic delivery format. The initial statements of the skill definitions were produced as a part of the transition from a paper and pencil format to a computer-delivered exam. With the exam in a computerized delivery format, an opening was created to test higher order skills in addition to the knowledge and understanding of accounting that was addressed in the paper and pencil version with MCQs and some open-ended questions. The skill definitions

resulting from the initial PA were communication, research, analysis, judgment, and understanding. The definitions of these skills are presented below in Table 1.

Table 1. Skill Definitions from 2000 PA

Communication	The ability to effectively elicit and/or express information through written or oral means
Research	The ability to locate and extract relevant information from available resource material
Analysis	The ability to organize, process, and interpret data to provide options for decision making
Judgment	The ability to evaluate options for decision making and provide an appropriate conclusion
Understanding	The ability to recognize and comprehend the meaning and application of a particular matter

After determining these definitions, the Board of Examiners (BOE) needed to determine which emphasis should be placed on each skill in the course of examination. It was stated that skills assessments will "…be made using a variety of methods, such as simulations, or relational case studies, which will test candidates' knowledge and skills using work related situations" (Examination Content Specifications, 2002, p. 2). The challenge since then has been to continually investigate which items explicitly tap the particular skills and exactly how they do so. This process will enable constant improvement through careful, systematic, and data-driven review.

To review and revise these skill definitions, the purpose of the 2008 PA and other work (Luecht & Gierl, 2007) was to provide a greater level of detail for the skill definitions. The result of this effort was much more specificity in the revised skills definitions than in the originals. The revision of the skill definitions accompanied the modification of the exam to its new format,

CBT- evolution (CBTe). Currently, more emphasis is placed on the simulations than in the previous incarnation (CBT); this increased emphasis on simulations provides an opportunity to make more specific claims about which skills are tapped during the simulations.

Whereas the original version of the SSO implied that skills and item types were strongly related, the current skills definitions are clearly specified in terms of item types on the exams. The current SSO states that item types are indeed proxies for skill definitions. The full version of the current SSO is too lengthy to present in a single table. It is made up of three general categories of skills: (a) knowledge and understanding, (b) application of the body of knowledge, and (c) written communication. The brief definitions of the current skill definitions are provided in Table 2.

Table 2. Current Skill Definitions (Brief Version)

Knowledge and understanding	• Knowledge is acquired through experience or education and is the theoretical or practical understanding of a subject. • Knowledge is also represented through awareness or familiarity with information gained by experience of a fact or situation. • Understanding represents a higher level than simple knowledge and is the process of using concepts to deal adequately with given situations, facts, or circumstances. • Understanding is the ability to recognize and comprehend the meaning of a particular concept.
Application of the body of knowledge, including analysis, judgment, synthesis, evaluation, and research	• Higher-level cognitive skills require individuals to act or transform knowledge in some fashion.
Written communication	• Written communication is assessed through the use of responses to essay questions, which will be based on the content topics as outlined in the CSOs. • Candidates will have access to a word processor that includes a spell check feature.

The brief version of the SSO is presented here; when fully expressed, the current skill definitions are much more detailed than the original version. (See references for a link to both the original and current SSOs.) The current SSO is much more thorough in its description of the types of skills that candidates are believed to need to perform adequately in the accounting profession. The current version of the SSO is an improvement on the original, but the way the skill definitions affect operational work can still be improved.

In summary, the second generation of skill definitions is a reaction to and improvement

on the original skills definitions. Moreover, substantial work is happening to ensure that the third version is yet another improvement. The primary motivation for this work is to provide more meaningful measurement information to the many groups of stakeholders for the Uniform CPA exam. These stakeholders include—but are not limited to—the American Institute of Certified Public Accountants (AICPA), the National Association of State Boards of Accountancy (NASBA), the Board of Examiners (BOE), the Examination Review Board (ERB), the Psychometric Oversight Committee (POC), and the candidates themselves.

Each stakeholder group will probably have specific concerns regarding what is meant by "more meaningful measurement information," and each group of stakeholders will likely have differing concerns regarding the usefulness of updated skill definitions. For example, candidates who must retake a section of the exam would benefit from enhanced score reports that reflect the particular skills that they did or did not adequately demonstrate on the exam (presuming these skill definitions will be useful for preparing to take the exam again). NASBA decides how candidate scores and score reports will be released; providing more information about skills may benefit this group when releasing score reports. The AICPA is responsible for creating, administering, and scoring the exam and as such, can benefit greatly from improved skill definitions in many ways (e.g., item and test development, scoring, score reporting, and validity arguments). Now that the motivation for why we are changing the skill definitions is clear, the discussion will focus on how we plan to make the changes.

Process for Developing New Skill Definitions

We are attempting to combine components of several existing techniques/methodologies to arrive at new skill specifications that we hope will be more meaningful to ourselves and other stakeholders. These components include (a) the general frameworks we are relying on to provide a structure within which we can build useful skill definitions; (b) the "hands-on" component by which we actually arrive at the skill definitions (and how this differs from previous revisions); and (c) the characteristics of the skill definitions and how we represent them so they are more useful than the previous SSOs.

General Frameworks

This work is based on three major frameworks. The first framework, Assessment Engineering (AE), is a principled approach to the process of test development, administration, and scoring (Luecht, 2009). This framework is used to guide the specification of skills as well as the construction of items purported to measure those skills. The second general framework that we have adopted is the Revised Bloom's Taxonomy (RBT) (Anderson & Krathwohl, 2001). The RBT provides a clear delineation of the differences between and among general cognitive skills, and provides a model to help us arrive at the specific skill definitions we hope to employ in the future. The third framework is Cognitive Task Analysis (CTA) (Clark et al., 2008). CTA involves the interaction of SMEs, cognitive scientists, test developers, and psychometricians. CTA is a technique that uses interviews and detailed observation to determine the required steps/processes for performing a task. All of these frameworks will be described briefly to foster a greater understanding of how we are using them.

Assessment Engineering (AE). Assessment engineering is a principled approach to item and test development that focuses on design and development of the assessment as well as the analysis, scoring, and score reporting of assessment results (Gierl & Leighton, 2010; Shu, Burke, & Luecht, 2010). Traditional test development involves balancing content coverage needs with psychometric concerns (Shu, Burke, & Luecht, 2010). AE is different from traditional test development approaches in that psychometric concerns are more directly incorporated in the process of building a test through applying engineering principles (Luecht, 2008).

In AE, the goal is to define multiple classes of items that represent the expected capability of examinees who have attained a given level of proficiency for each construct tapped by the assessment device. AE redefines the traditional test blueprint. Content concerns are not ignored; rather, this approach allows for including rigorous, psychometrically based guidelines for item development to ensure that content coverage and psychometric concerns are both considered in item and test development. AE requires the specification of Construct Maps, the defining of Task Models, the building of Templates, and psychometric calibration and scaling. Each of these components of AE is described briefly below.

Assessment engineering begins with the process of construct modeling. Construct modeling involves the development of construct maps. Wilson (2005) and Gierl and Leighton (2010) provide much more thorough descriptions of the building of construct maps. In the context of certification/licensure tests, the "construct" is the ability being measured (e.g., auditing and attestation ability). Construct maps (CMs) are an explicit attempt to link observed behavior to latent theoretical explanations of that behavior. A CM represents ordered proficiency claims along a dimension represented as a continuum. These CMs describe which levels of ability along a dimension are meaningfully different and attach to them the skills (i.e., observable

behaviors) expected to be displayed by someone who has attained that level. The CMs help guide item development by providing an explanation of what behaviors are indicative of mastery of a given level of proficiency. Additionally, the CMs imply the scale along which the relationship between latent and observed characteristics is to be described.

Task models (TMs) are cognitively oriented statements that define a class of potential items that could be used to provide information about whether or not candidates have achieved a certain level of ability along the construct. TMs include descriptions of context complexity, cognitive load, and auxiliary information available to examinees. A single TM is located at a particular point along the proficiency scale, and the TMs are defined in such a way as to provide evidence about a certain point along the construct map. Each TM is defined very clearly to assess the observable behaviors that are informative about the latent characteristics of the examinee. Each TM can be used to define multiple templates.

Templates are specific instances of TMs. Templates define the form that a class of items should take (e.g., spreadsheet entry) as well as define the scoring evaluator for the items produced. When properly implemented, templates describe the guidelines to be followed for producing multiple items that share similar psychometric properties. Each item produced from a single template should be psychometrically similar to every other item produced from the template. Item characteristics (e.g., difficulty) are "inherited" from the family to which they belong. The feature of inherited item characteristics allows for items developed from the same template to be interchangeable.

In AE, psychometric models are used as quality assurance mechanisms. When proper controls are in place, every item developed from a template should have predictable, controlled parameters. During calibration of instances, if it is found that one item is functioning differently

from another, then further work is needed to refine the template (Otherwise, the item writer is not following instructions.) In this way, the controls envisioned by the methodology can be empirically verified and refined if necessary.

Revised Bloom's Taxonomy (RBT). The RBT (Anderson & Krathwohl, 2001) is an expression and description of the varying types of cognitive skills believed to exist and apply to a wide range of human endeavors. Based clearly on the original taxonomy created by Bloom et al. (1956), the RBT is an attempt to modernize the taxonomy in light of roughly 50 years of cognitive research. The most prominent feature of the taxonomy for our purposes is that it differentiates among several different general "types" of higher-order cognitive processes, or skills. The categories of cognitive skills in RBT are presented in Figure 1 below.

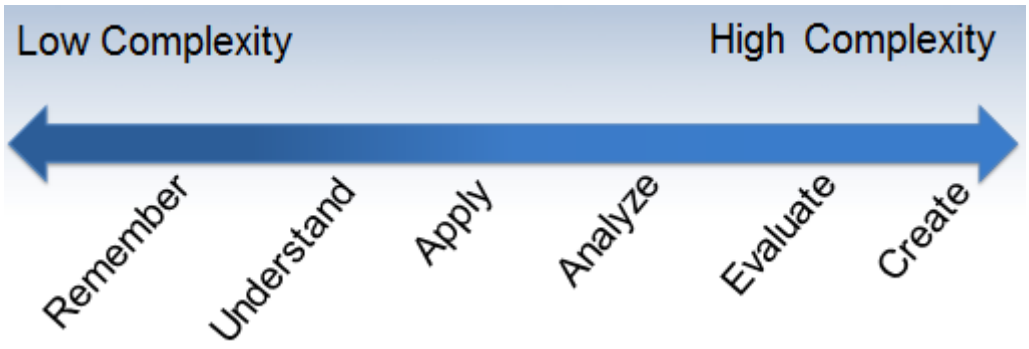

Figure 1. The Revised Bloom's Taxonomy

Figure 1 displays the six general categories in the RBT. As one goes from "Remember" to "Create," it is assumed that a corresponding increase occurs in the complexity of the cognitive processes. The RBT provided us with an established framework that we were able to use to orient our skill definitions. It was not our intent to use the taxonomy to constrain our skill definitions; rather it was viewed as a general framework we could use to help organize our

thinking about our skill definitions. Thus, we have not felt compelled to make our definitions fit into the taxonomy, but the taxonomy has proved useful in providing some measure of differentiation among the interrelated skills that we believe the TBSs rely on. Additionally, review of the TBSs has revealed that some of these skill categories are not represented in the exam. For example, it is not the intent of the exam to measure "Create". One of our ongoing concerns with using RBT has been that it is not perfectly clear if we want TBSs to measure all of these categories of cognitive processes. This issue will be revisited briefly in the summary.

Cognitive task analysis. Cognitive task analysis can be summarized as a variety of inter-related qualitative methods that attempt to provide insight into the "moving parts" underlying a task, and it is applicable in many different areas. Clark et al. (2008) groups the many versions of CTAs found in the literature into three general categories: (a) observation and interview; (b) process tracing; and (c) conceptual techniques. Observation and interview involves examining the technique or task of interest, typically by way of watching an expert perform the task and asking questions to try to elicit the steps that must be taken. Process tracing involves capturing an expert's approach to solving a task by way of a think-aloud protocol or a structured interview asking for a detailed description of steps while, or immediately after, performing the task. Conceptual techniques use interviews to try to build representations of the interrelated concepts making up a domain or task by trying to identify important concepts indicative of understanding.

We have adopted an approach made up of elements of each of the three general types of CTAs. Essentially, our attempt to begin revising the skill definitions entails engaging in a semi-guided interview whereby measurement experts and item developers talk to SMEs about what demands a particular task places on candidates. Simply put, we want to know what the expert is thinking and doing when faced with a task. We are trying to identify the "moving parts", or

critical components, of the task from the perspective of someone who clearly knows how to do it. SMEs, cognitive scientists, test developers, and psychometricians collaborate to identify whether the item is functioning as intended regarding the skills involved.

Taken together, the frameworks of AE, RBT, and CTA provide a way of framing our new definitions of the skills. AE provides an overall structure we can employ to make sure we are using items on the exam that fit with our conceptualizations of the skills that candidates are supposed to be knowledgeable about, and how those skills relate to the bigger picture of accounting ability. RBT is providing us a framework to organize our characterizations of the skills required to successfully employ the skills of accountancy (in terms of more or less complexity). CTA is the method employed to review the items. To revise our skill definitions, we engage in a CTA of each individual TBS on the CPA exam, and we use the RBT to roughly categorize the cognitive demand of the items. We then create a much more specific characterization of the item in terms of a skill definition that relates the skill to be employed and the content to which it is appropriate. These skill definitions will then be subjected to several analyses to determine effects on the multiple aspects of validity. The goal of this process is to try to find ways to add value to related test development processes for multiple stakeholders.

Hands-on Approach

We have begun by examining existing items on the test. (To date, we have focused on the TBSs.) The interview focuses on asking SMEs to perform such tasks while clearly explaining and/or demonstrating what a candidate must do to arrive at a correct solution, what information is relevant to answering the item, and what type of information works as an effective distractor. Through repeated questioning about the items from the standpoint of individuals who do not know the subject matter (e.g., cognitive scientists, psychometricians, and test developers), it has

been possible to arrive at a consensus as to which cognitive skill in the RBT is required to answer the item. The skill settled on in the taxonomy is only to help us to categorize items in terms of the general cognitive skills that underlie them.

An important determination still remains to be made. Generally, we want to be as specific as we can be about the demands an item places on examines, but do so using the appropriate terminology and specifics of the accounting skills represented in the items. After the general cognitive skill is specified, further statements are identified or generated collaboratively to try to provide specific, actionable descriptions in accounting terms of what the item is telling us about candidates' skills. So, once a measurement opportunity in a TBS has been classified as seeming to be under the general heading of "Analysis", further work is done to arrive at a description of what the item asks the candidate to do using the appropriate accounting terminology. For example, there are multiple TBSs on the exam that we feel can be accurately described by the specific skill phrase: "*Determine the impact of changes to related accounts on financial metrics*." The general cognitive skill "Analysis" accurately captures the essence of what a candidate must do, and the particular skill phrase hopefully makes this general cognitive skill description more tangible to multiple stakeholders. This is an iterative procedure that requires thorough review and should be done on a regular basis to continually ensure that items are functioning as intended. Figure 2 summarizes the current procedure.

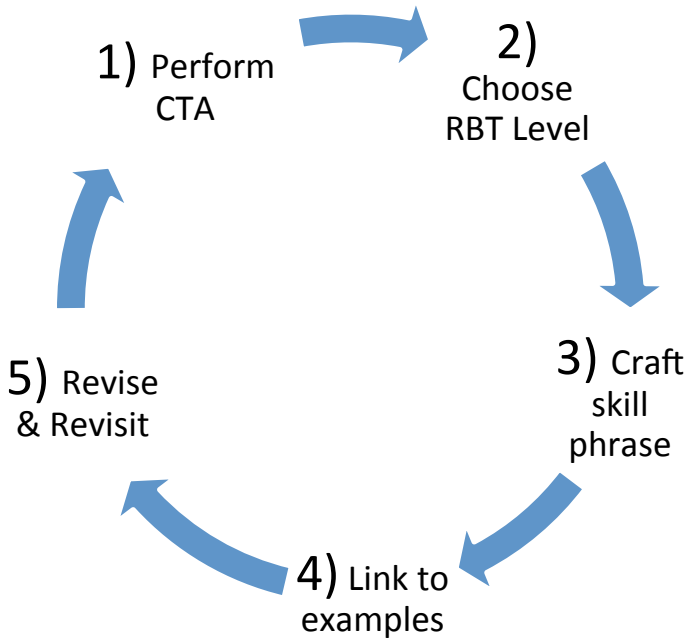

Figure 2. Method at-a-glance

The critical aspect of this work is to make the information we can glean about the skills useful and meaningful. To provide more meaningful information to the stakeholders, the skills definitions must be written to an ideal level of specificity. A challenge at each step of developing and refining the skill definitions has been settling on a level of detail that is neither too specific nor too general. A skill definition that is too specific (i.e., applies only to one very particular task or area of contextual knowledge) is not useful. Any inference made based on that definition does not generalize to other tasks or areas of knowledge. In contrast, a skill definition that is too general hinders inference because it applies to too many tasks or areas of knowledge. Ideally, the skill definitions should be phrased so that they apply to multiple TBSs in a way that is informative to multiple stakeholders. If this goal can be achieved, then the skills definitions can begin to be used to provide more meaningful measurement information.

In summary, the hands-on approach we are taking essentially requires examining existing

items and reasoning backwards to the presumed underlying skills. The skill phrases are couched in the profession's terminology, and are based on items currently included in the Uniform CPA Examination. (Note: Preparing for our upcoming PA entails gaining a sufficient understanding regarding the state of skills as they exist in our current item bank. Engaging in a principled review of items to determine the nature of the content and skills currently measured by the CPA Exam can directly inform the upcoming PA and subsequent test specifications. By "reverse engineering" the skill statements from the previous PA, the forthcoming test specifications and test program will benefit from improvements based on the perceived weaknesses of its predecessor.) These skill phrases are defined with the intent of making them amenable to action. Specifically, the skill phrases should make sense and be useful to stakeholders. For example, the definitions should be useful in writing items, in describing the ability of candidates, and providing direction to candidates hoping to improve their understanding of accounting tasks. This process entails significant time and effort, but it is time and effort well spent if it can ultimately be made useful.

Validity and Future Directions

The primary reason that we are attempting to incorporate an AE framework into the Uniform CPA Examination is to improve the quality of our test development process. By improving the test development process we are directly addressing the validity of the exam. Incorporating AE will add value to the process of test development in several ways; thus, it is our responsibility to show that this is indeed the case. Therefore, this section of the paper will introduce and briefly describe some of the critical aspects of validity evidence. The work we are doing (and still to be done) to address whether or not our effort is having its desired effect will also be provided.

Aspects of Validity

Assessment engineering directly addresses certain aspects of validity and is indirectly related to others. The purpose of this section of the paper is to discuss how the work being done to implement AE will be validated. The primary focus here will be to briefly outline the work that must be done so that the implementation of AE adds to the quality of the examination.

There are several aspects of validity evidence, each of which is an important element of the simple definition of validity: "Are the inferences we make on the basis of our assessments supported?" With a professional certification/licensure examination, a primary concern is content validity; however, other important aspects of validity should also be addressed. According to the Standards for Educational and Psychological Testing (AERA/APA/NCME, 2008), there are several important aspects of validity: (a) content; (b) evidence based on response processes; (c) evidence based on internal structure, (d) relationships to other variables, and (e) consequences.

Content validity. Content validity represents the degree to which a measure represents the facets of the construct it is designed to measure (i.e., whether the topics covered are relevant to the construct being measured). For example, exam blueprints and test specifications are used to define content domains. These content domains reflect the subject matter that will be assessed. To know

that all of the critical content relevant to an assessment has been successfully covered requires the input of content experts.

In regard to content validity, all of our work to implement AE and revise the skill definitions will become part of the next Practice Analysis (PA). It is our opinion that content validity incorporates both the *content knowledge* and the *skills* to be demonstrated as part of proficiency with the construct of accounting. For the Uniform CPA Examination, we constantly monitor changes to the profession and conduct a large scale PA roughly every 7 years. The PA involves obtaining the input of a representative sample of CPAs and other knowledgeable subject matter experts to review our assessment. The PA is typically done in the form of a survey that asks the participants to rate the relevance and importance of the content we present in the examination. For our upcoming PA, the survey materials will be created in compliance with generally accepted best practices, such as SME review of content specifications, but we are trying to do more. We are attempting to add value to the process by providing the SMEs with thorough descriptions of the nature and types of skills that are necessary to successfully reply to exam questions as distilled from the detailed CTA in which we are engaging. The survey material is created through a principled series of interactions of SMEs with multiple testing experts with an eye toward bringing this information into play in a meaningful way during the PA. By providing an inside look at the work to revise skill definitions, these subject matter experts can determine whether or not the work is beneficial regarding its treatment of the exam content. This input is critical in establishing that we are addressing the appropriate content for the practice of accounting.

Also, part of what we are trying to do is to restructure the Content Specifications Outline (CSO) and Skill Specifications Outline (SSO) so that they fit better within an AE framework and bring the utility of Task Models and Templates into play. Using AE's infrastructure to organize test specifications can provide practical benefits to test development practices, but the restructuring of

the CSOs and SSOs must be done in such a way that we do not alter or impair the functioning of the exam in terms of its coverage of the requisite content areas as we add components that emphasize skills assessment. Any changes made to the test specifications must be subject to content review.

Evidence based on response processes. The response processes of test takers carry a wealth of information regarding the relationship between the construct and how it is represented through test questions designed to elicit critical behaviors from candidates to demonstrate proficiency (or lack thereof). Considering the thought processes of candidates can aid measurement precision. Essentially, this aspect of validity is concerned with whether or not the way in which candidates think through the items is in line with the construct as it has been defined.

Typically, this type of evidence is obtained by observation or interview of candidates to gain insight on their thoughts while (or shortly after) replying to items. "Think aloud" procedures are a good example of this type of evidence (for a thorough review, see van Someren, Barnard, and Sandberg, 1994). The thought processes revealed in these observations and interviews can be quite informative about any discrepancies between test developers and test takers when it comes to thinking about critical aspects of assessing proficiency. The way the candidates think about the items should align with the way test developers think about the items. Additionally, combining TAP methods with other techniques, such as eye tracking or keystroke tracking studies can potentially enrich the quality of understanding of the thought processes involved in responding to items.

It is also possible to gain insight into the response processes of candidates without asking them to describe their thinking. By developing items within a framework that clearly explicates the "moving parts" that are critical to defining proficiency, it is possible to design items that can be ordered in difficulty in an *a priori* fashion (for example, see Daniel and Embretson, 2010). We can develop a series of items of increasing difficulty by manipulating the critical aspects of complexity

relevant to the current topic. If done correctly, these manipulations of difficulty should then be revealed in the observed difficulty of the items.

In regard to the response process aspect of validity, we currently have plans to engage in both a think-aloud procedure with candidates (along with a related follow up study), as well as perform an empirical study of item modeling (i.e., controlling difficulty). Think-aloud procedures are focused on investigating the thought processes of the examinees targeted by the assessment. Empirical studies of item modeling provide evidence concerning manipulations of items to control difficulty.

By addressing the individual thought processes of candidates, we gain evidence about the accuracy of the construct we have defined. The think-aloud procedure provides an opportunity to evaluate whether the construct we have defined is specified well enough to capture the responses of candidates accurately and interpret them appropriately. A study of this sort will be relatively easy to set up. The recruitment of participants can be done on a small scale in a similar fashion to the field testing commonly done here at the AICPA. Essentially, a very small group of candidates (e.g., 5 or 6) will be given some TBSs similar to those used operationally on the exam. These items will be administered to each candidate individually in the setting of informal observation/interview. As the candidates work through the measurement opportunities, they will be asked to describe their thought processes in general and will be prompted to give further detail about aspects of interest to the test developers, especially in regard to the skill definitions produced through the item review process. If possible we hope to follow up on the TAP with an alternative process tracing technique like an eye tracking study (Cooke, 1994) to provide converging lines of evidence that inform us about item functioning. Procedures of this sort provide insight into the approach candidates adopt as they work through the simulations, and can highlight critical aspects of candidates' understanding of the construct that may allow us to revise/refine our representation of the construct of accounting.

By addressing the controlled manipulations of item complexity we should also gain evidence about the accuracy of the construct we have defined. It will allow us to test predictions about variables affecting the difficulty of items. Any manipulations of the components of TBSs that drive difficulty should have a predictable effect that allows for the control of difficulty in an *a priori* fashion. So, if the construct has been defined appropriately, it should be possible to target items in terms of difficulty; these manipulations of difficulty should be born out in the candidates' responses. There are two basic ways to achieve this goal: (a) field test some TBSs, or (b) place the TBSs into the exam as pre-test items.

Related to doing a study of controlled manipulations of item complexity, we currently have data concerning some of the instances used in the previous computerized version of the exam (CBT). These instances were items that were created from an existing item by changing numerical values in the measurement opportunities. As such, the intent was to change only surface features of the items while still having them tap the same aspects of content. Investigation of these CBT instances has revealed that it is possible to create multiple versions of the same item that all function similarly from a psychometric viewpoint. Admittedly, this is only tangential evidence to support the claim of response process validity in the Uniform CPA examination. However, findings of this sort are crucial pieces of evidence that AE can function as advertised.

We also have produced new instances of some operational items and pretested them in operational settings to further test our ability to apply guidelines in item writing that effectively control the difficulty of new items. At time of publication, we are collecting data on these new instances.

If these sources of evidence (i.e., the think-aloud study and the empirical difficulty study) are favorable, then we gain some security in regard to the faith we place in the inferences we make

based on exam performance. Simply put, our construct definitions align with the response processes of our target population. Any misconceptions or incongruences may be illuminated as well.

Evidence based on internal structure. The internal structure of an exam is the mechanism/model by which it represents the construct of interest. It is therefore a necessary component of a validity argument that the presumed mechanism used to define the construct does indeed align with the empirical evidence collected concerning the functioning of the exam.

For example, the Uniform CPA Examination employs a 3PL model to represent examinee ability. As stated previously, the 3PL model is a unidimensional model that represents ability as a single continuous scale that allows for a more-or-less interpretation. To verify that evidence of a valid internal structure exists, it is necessary to show that empirical investigation of test data indicates that a unidimensional representation is appropriate.

Also, the choice of a 3PL as the psychometric model overlaps with the process of construct mapping. When detailing the interplay between content, skills, and tasks, the use of a hierarchical, unidimensional scale makes sense in light of the assumptions we make in the development and delivery of the Uniform CPA examination.

In regard to aspects of validity related to internal structure, work is being done to investigate the restructuring of item types to reduce the possibility of guessing at items, and thus investigate alternatives to the 3PL. For example; we have recently conducted field tests comparing multiple choice questions to open ended versions of questions tapping the same content and skills. Additionally, the sections of the Uniform CPA examination are represented as unidimensional constructs. The marginal reliabilities for each section of the exam indicate that we can state in complete confidence that we do a very good job of maintaining a unidimensional scale with content

balanced in accordance with the guidance of critical stakeholders. As we move forward with implementing AE, evidence of consistent or improved marginal reliabilities should show that we have maintained this record. Again, a critical constraint on the work we are doing to revise the skill definitions is that we cannot degrade the quality of inferences we wish to make on the basis of exam performance. The exam currently works very well for its intended purpose. Any changes to the exam or test development procedures must not reduce the quality of the exam for its intended purpose.

Relationship to other variables. The score produced by the CPA examination is likely to be strongly associated with other things that rely on the same content, skills, and tasks. Traditionally, validity in testing often was represented as the correlation between a test score and some appropriate criterion. The challenge of this type of validity is that it is not always easy to identify and/or quantify an appropriate criterion. In the case of the Uniform CPA exam, the criterion (i.e., the practice of accounting) cannot be observed until after successful completion of the exam because we are the gatekeepers to licensure.

In regard to relationships to other variables, work is being done to investigate the most useful factors in controlling the difficulty of the items we develop. By finding ancillary variables that help to predict difficulty of items, we are finding evidence that can be used to support claims of the appropriateness of our item types. This is also related to the studies regarding the response process aspect of validity.

Consequences. The consequences of the use of test scores are a critical aspect of validity. Testing, above all, is about fairness and a level playing field as the basis of comparison. Decisions about setting cut scores, for example, directly influence who will be granted licensure and who will

not. As the purpose of the CPA Exam is to protect the public interest by preventing unqualified individuals from entering the profession, it is crucial that a cut score be chosen that effectively separates those that can't from those that can. If set too low, unqualified candidates may enter the field and the public could suffer. Conversely, if the score is set too high, qualified candidates would be improperly denied a license and unable to serve the public in a profession in which they are capable. Decisions about licensure are most effective when they are based on an exam whose cut score has been set based on rigorous, principled application of cognitive and psychometric theory.

In respect to consequential validity, an indirect link exists between the work being done and its influence on the use of test scores. For example, it is our plan to investigate the utility of incorporating what we are learning through implementing AE into our next Standard Setting. The infrastructure of AE incorporates meaningful representations of the interplay of skills and content knowledge that provide a degree of control over item development processes resulting in predictable psychometric characteristics of items. Construct Maps, Task Models, and Templates are a tangible way to represent the information necessary for standard setters to understand and perform their jobs in setting cut scores. Carrying over the lessons we learn in the review of items and preparation of materials for a PA and doing so with a principled approach like AE is a crucial part of understanding what a minimally competent candidate should be able to do (and what items let you know they can do it). These lessons can be useful when making decisions about setting the cut score for passing.

To summarize, a significant amount of work is being planned to address content and construct validity evidence for the Uniform CPA examination. As the work progresses in the near future, any additional strategies/studies that must be implemented can be prioritized as they arise.

Simultaneous Consideration of Multiple Aspects of Validity

No single aspect of validity holds any priority over any other. Each aspect of validity is important to consider when test scores are put to high stakes use. Ideally, all aspects of validity should be attended to equally. For the Uniform CPA examination, we are investigating ways in which AE may allow us to simultaneously consider multiple aspects of validity.

We are currently developing interactive test specification documents that specifically link content to skills and ultimately to examples of operational simulations. We will be incorporating information about content and response process aspects of validity in a unified CSO/SSO. Item writing under an AE framework moves the emphasis from art to science and from hand-crafting individual items to controlling the characteristics of multiple items in a directed way.

Our first chance to ascertain whether this work is having a positive impact on test development will come when we can begin to put these skill definitions to use in a test of AE's claims of directed item development. In particular, we plan to incorporate the skill definitions directly in the authoring of items to directly address the skills as represented in the Construct Maps and Task Models. The CSO/SSO is the vehicle that we will employ to represent our beliefs about the interplay of knowledge and skill in relation to the task we use to assess proficiency. The CSO/SSO will enable us to embody the task models and templates in a useful format to give specific instruction to item writers so that we can control not only content coverage, but also do our best to ensure that we have control over the psychometric properties of items. This entails dealing with multiple aspects of validity evidence simultaneously. We can offer directed item development to item writers due to AE's emphasis on the interconnected nature of testing systems. Thus, overlapping aspects of validity are concurrently attended to.

Lessons Learned

This work gives us a clearer understanding of the composition of our test bank. Spending this much time dissecting items with the combined input of SMEs, measurement specialists, and cognitive scientists provides a better understanding of the items we have and the information they provide. This work has already provided some feedback for this purpose.

Another lesson learned is the importance of this work in providing us a greater understanding of the thought processes of SMEs in dealing with accounting tasks. Greater explication of the interrelated concepts that define a particular aspect of accounting ability gives us greater flexibility in trying to devise items to tap into a candidate's grasp of the subject matter. A greater understanding of the skills we are measuring in the exam promises more clearly specified links between the latent construct of interest (accounting ability) and the skills as tapped by exam questions (i.e., Construct Maps). A more clearly specified link between the latent construct and the skills that represent varying levels of familiarity with the latent construct allows for the development of task model statements that clearly explicate the type of information provided by a particular class of items that could exist. Templates based on the task model statements allow for increased production of items known to provide information about a candidate's ability. In short, this work helps keep us oriented as we try to apply AE principles to the Uniform CPA Examination.

There are obvious limitations to this work as well as daunting obstacles that must be overcome for it to be practical. The most ominous of these limitations will be addressed briefly. First and foremost, attempting to clearly specify the interrelated concepts and knowledge for each and every accounting task relevant to the examination is an enormous undertaking. In fact, some individuals would likely think that clear specification of all relevant cognitive models would be impossible. The sheer volume of tasks deemed necessary to be a minimally proficient CPA alone indicates that this sort of work, if possible, would continue for many years. In addition to the sheer amount of work to cover existing accounting tasks, changes to policy and procedure would be an ongoing concern. Beyond simply classifying all tasks by way of our brand of CTA, we would

constantly need to stay abreast and ahead of changes to the profession. Any misspecifications in these skill phrases or the CTA on which they are based could derail the entire AE locomotive. While this process is not easy, it is certainly attemptable and would provide great reward for the cost if successful.

Also, a very practical obstacle at this point is trying to provide information about multiple skills from an assessment that has been designed to be unidimensional. The AERA/APA/NCME standards (1999) (http://www.aera.net/publications/Default.aspx?menu_id=32&id=1850) provide clear warning about the practice of reporting sub-scores when a single total score is available. Scores must be reliable and valid to be used in reporting meaningful information to stakeholders. If the skill phrases we develop do not rest on a firm foundation of sound psychometric theory, we may still be guilty of fielding skill definitions that do not provide any usefulness, and worse yet, provide misdirection. This is a real psychometric challenge, and it must be kept in mind at each step in the process of revising the skill definitions and putting them to work.

The future of this work lies in trying to make the theoretical practical. Knowledgeable experts are the ideal source of information concerning what is going on with the accounting tasks we present in the Uniform CPA Examination. Principled approaches to test development and design such as AE rely on the input of those who know the material best. The input of measurement specialists and cognitive scientists should focus on taking what the experts know and finding a way to soundly determine if others know it as well. To begin, the remaining TBSs in the item bank must be reviewed so that the skill phrases can be refined until they are representative of all operational items. These skill phrases must then be put to use to provide meaningful information. For example, what do these skill phrases really mean to the stakeholders? Can they tell candidates something to help them retake and pass a section of the exam in the future? If so, how will this be represented in

a score report? Validity evidence must be clearly identified and then sought. How will we know that we are measuring the right skills?

, a great deal of work remains to be done. While this work is being done, it will be crucial that we monitor our progress and align ourselves with what are almost certainly going to be changing goals and strategies. This work is important because this is an aspect of the exam that has not been utilized in any way thus far. To provide meaningful definitions of the skills involved in the Uniform CPA examination, and to be able to leverage these skill definitions to making the assessment more useful, is a challenge worth undertaking.

Recommendations

The work described herein is an involved undertaking requiring a considerable amount of time and resources. It offers the promise of added value to test development practices, but not immediately. The AICPA is accommodating this work because of its long term vision of exam development and an emphasis on continuing improvement.

Gaining a better understanding of 1) the construct itself and 2) the nature of the inventory that you use to provide evidence about the construct gives insight into what skills should be demonstrated (and the level at which they should be demonstrated). This is especially important in licensing exams such as ours as we provide a service, namely, protecting the public from unqualified accounting practitioners. Incorporating the elements of AE gives us a better sense of the evidence that needs to be provided to serve the public interest.

Our advice to others seeking to take a similar approach is to determine whether the environment you are in is amenable to a process frontloaded with planning, preparation, and research and whose benefits will come into play at a future date. If feasible, we strongly encourage others to engage in this work. If constraints are present and you cannot fully implement the sorts of

efforts described in this paper, we feel that a smaller scale implementation of some of the activities we perform (and are planning to perform) can provide added value to certain test development practices.

Item review

Consistent, comprehensive item review procedures can be implemented relatively easily to improve item development practices. The act of revisiting questions after data has been collected to compare the functioning of the item in actuality to the presumed role it was to play as laid out in the test specifications is absolutely essential. CTA is a relatively straightforward procedure that doesn't require any special technology (just training and practice). Additionally, there are multiple cognitive taxonomies to choose from which can help structure those conversations so that you can start to organize your understanding of your items and the skills they tap. Knowing your items are clearly connected to your test specifications and knowing specifically how they function is an essential part of establishing and demonstrating the validity of your exam. In the context of certification and licensure tests, demonstration of the skills associated with the construct is at least as important as the requisite content knowledge on which they draw. Using a taxonomy for skills helps you focus on the level and type of skill you are testing as opposed to the factual knowledge upon which those skills rest.

Item templates

Thorough item review procedures partnered with the use of item templates adds an additional feature; the potential mass production of items with similar psychometric characteristics. Employing AE's concept of task models and templates brings a nice structure to one of the principal concerns of any test development process, namely, item development. This essential task of test

development can be improved by effectively putting the empirical data gained in item review to use in providing item writing guidelines that translate to efficiency, accuracy, and consistency in item development.

Test Specifications

The framework of Assessment Engineering provides a nice way to represent the connection between content and skills believed to undergird a construct. The test specification is an ideal place to communicate this representation of the construct. A great deal of information can be captured and presented in test specifications in a way useful for multiple stakeholder groups. It is our opinion that well-developed, thorough test specifications should translate the construct (as defined through a Practice Analysis) faithfully into items that provide meaningful information about candidates' levels of proficiency. Combining a comprehensive item review procedure with the notion of Task Models and Templates puts one in an advantageous position to create test specifications that serve double duty as quality assurance mechanisms.

Putting it all together

What may be the nicest feature of AE is that if you start to implement these processes piecemeal, you can build on them with further elaboration of the framework in your day to day operations. As implied above, adding an item review procedure similar to the one described here would not be a huge undertaking. Once it is in place, other procedures/practices can be added as necessary to supplement the item review procedure. Once multiple pieces are in place, it is easier to consider linkages between test specifications, items, and test scores and what they mean. Implementing AE does not have to be an overarching change to the way day to day business is run, rather it can be a gradual transition to processes that have long term vision.

Ultimately, we recommend that you be prepared to ask hard questions of your testing practices and don't be surprised by the answers, or lack thereof. Are there things in your test specifications that are not amenable to testing in the form you do it? Do you know which skills are expected to be demonstrated in which content areas? What evidence can I collect to make decisions about changing practices for the better? AE doesn't guarantee an answer to these questions; it simply provides a framework for arranging things so that you can ask yourself tough questions.

References

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

Anderson, L. W., Krathwohl, D. R., Airasian, P. W., Cruikshank, K. A., Mayer, R. E., Pintrich, P. R., Raths, J., & Wittrock, M. C. (Eds.). (2000.) *A taxonomy for learning, teaching, and assessing: A Revision of Bloom's Taxonomy of Educational Objectives, Abridged Edition*. Boston, MA: Allyn & Bacon.

Bloom, B. S. (1956). *Taxonomy of educational objectives, handbook I: The cognitive domain*. New York, NY: David McKay Co., Inc.

Clark, R., Feldon, D., van Merrienboer, J., Yates, K., & Early, S. (2008). Cognitive task analysis. In J. M. Spector, M. D. Merrill, J. J. G. van Merrienboer, & M. P. Driscoll (Eds.), *Handbook of research on educational communications and technology* (3rd ed.) (pp. 577-593). New York, NY: Macmillan/Gale.

Cooke, N. J. (1994). Varieties of knowledge elicitation techniques. International Journal of Human Computer Studies, 41, 801–849.

Daniel, R. & Embretson, S. (2010). Designing cognitive complexity in mathematical problem solving items. Applied Psychological Measurement, 34(5), 348-364.

Examination Content Specifications: http://www.aicpa.org/BECOMEACPA/CPAEXAM/EXAMINATIONCONTENT/CONTENTANDSKILLS/Pages/default.aspx

Gierl, M., & Leighton, J (2010). Developing construct maps to promote formative diagnostic inferences using assessment engineering. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Denver, CO.

Luecht, R. M., & Gierl, M. (2007) *Defining skills constructs for the Uniform CPA Examination: An application of assessment engineering*. AICPA Internal Report.

Luecht, R. (2008, October). *Assessment engineering in test design, development, assembly, and scoring.* Invited Keynote presented at the Annual Meeting of the East Coast Language Testing Organizations (ECOLT), Washington, DC.

Luecht, R. M. (2009, June). *Adaptive computer-based tasks under an assessment engineering paradigm.* Paper presented at the 2009 GMAC CAT Conference, Minneapolis, MN (in press, Proceedings of the 2009 GMAC CAT Conference).

Shu, Z., Burke, M., & Luecht, R. (2010). *Some quality control results of using a hierarchical Bayesian calibration system for Assessment Engineering task models, templates, and items.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, Denver, CO.

van Someren, M., Barnard, Y., & Sandberg, J. (1994). *The think aloud method: A practical guide to modeling cognitive processes.* London, England; Academic Press.

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah, NJ: Lawrence Erlbaum Associates.