A Review of Assessment Engineering Principles with Select Applications to the Certified Public Accountant Examination

Technical Report
October 2009
W0903

Jiawen Zhou

Centre for Research in Applied Measurement and Evaluation
University of Alberta

Prepared For:
The American Institute of Certified Public Accountants

AICPA®
American Institute of Certified Public Accountants

Introduction

Modern technology has infiltrated almost all areas in contemporary society and, in turn, generated new opportunities for interdisciplinary research. The discipline of educational measurement is no exception. One consequence of these interdisciplinary influences is the emergence of a new research area called *assessment engineering* (Luecht, 2006a, 2006b; see also Luecht, Gierl, Tan, & Huff, 2006). Assessment engineering is an innovative approach to measurement where principled test design concepts are used to direct the design and development as well as the analysis, scoring, and reporting of assessment results. Assessment engineering requires four explicit stages. In the first stage, the measurement specialist defines the construct of interest using specific, empirically-derived cognitive models of task performance. In the second stage, item models are created to produce replicable assessment tasks. In the third stage, automated test assembly procedures are employed to build assessments that function to exacting specifications. In the fourth stage, psychometric models are applied to the examinee response data collected using item models to produce scores that are both replicable and interpretable. The purpose of this research project is to describe my work in applying the principles from the first two stages in assessment engineering to the Certified Public Accountant (CPA) Examination. This attempt helps link cognitive theory and psychometric practice to facilitate assessment development for the CPA Examination. In the **first section** of this paper I review assessment engineering in educational measurement and explain why it is important in the development and analysis of assessments. In the **second section** I briefly present the applications of the first two stages of assessment engineering towards CPA Examination. In the **third section**, I provide a summary of the study and identify areas where additional research is required.

Section I

Assessment Engineering and Educational Measurement

Assessment engineering is a research area where engineering-based principles are used to direct test development as well as the analysis, scoring, and reporting of test results. The concept of assessment engineering reflects the central idea described by Drasgow, Luecht, and

Bennett (2006) in their seminal chapter in *Educational Measurement* (4th Edition) on technology and testing:

> *Our vision of a 21st-century testing program capitalizes on modern technology and takes advantage of recent innovations in testing. Using an analogy from engineering, we envision a modern testing program as an integrated system of systems. (p. 471)*

Assessment engineering differs from more traditional approaches to test development and analysis in four important ways; namely, construct definition, model-driven item development, automated item assembly, and model-data fit assessment. These four aspects are reviewed next.

Stage 1: Construct Definition—Cognitive Model of Task Performance

Assessment engineering relies on a cognitive model, of some type, rather than content blueprints, to develop items and analyze examinee item response data, generate scores, and guide score interpretations. In educational measurement, a cognitive model refers to a, "simplified description of human problem solving on standardized educational tasks, which helps to characterize the knowledge and skills examinees at different levels of learning have acquired and to facilitate the explanation and prediction of examinees' performance" (Leighton & Gierl, 2007, p. 6). A cognitive model of task performance is designed to specify the knowledge requirements and processing skills underlie examinees' task performance applied to solve items. A cognitive model is organized with fine-grain components because it represents detailed knowledge structures and thinking process. The grain size of each component in a cognitive model can be specified to reflect specific problem-solving skills so we can connect test performance and score interpretations to understanding how the examinees' cognitive skills can produce their responses and, subsequently, test scores. Often, this type of specificity cannot be obtained using a post-hoc or retrofitting approach to test design (e.g., coding existing items for cognitive attributes) because items with these specific cognitive characteristics are unlikely to exist on a test developed without a cognitive model.

Among a variety of procedures (e.g., judgmental and logical analyses, generalizability studies, and analyses of group differences; Messick, 1989) that may be used to generate a cognitive model, verbal report methods provide one way to study information-processing skills (Leighton & Gierl, 2007). Researchers can develop a cognitive model by administering tasks to a group of examinees who represent the intended population, having them think aloud as they solve items, and then conducting protocol or verbal analysis with the corresponding verbal data (Chi, 1997; Ericsson & Simon, 1993; Leighton, 2004; Leighton & Gierl, 2007; Taylor & Dionne, 2000).

Only a valid cognitive model can be used to empirically confirm the thinking processes individuals use to answer or solve classes of test items. An existing cognitive model can be validated using the same method as it is generated, for example, by verbal reports. A cognitive model can also be evaluated by checking how examinees' observed response data fits it or by comparing model-data fit across competing models using statistical procedures such as structural equation modeling.

The influence of cognitive models in educational and psychological measurement appears in many forms. One beneficial feature of using a cognitive model is that it is a viable guide for item and test development after the model is developed and validated. With the description of specific, fine-grain cognitive skills provided by a cognitive model, test developers can create items and test with the control over the knowledge and skills specifically measured by each item. That is, the assessment principles used in test construction are much more specific allowing items to be created quickly and efficiently during the development cycle.

Another benefit of a cognitive model is its facility to provide detailed cognitive diagnostic feedback to examinees about their problem-solving strengths and weaknesses as the model provides an explicit framework necessary to link cognitively-based inference with specific, fine-grain test score interpretations (Leighton & Gierl, 2007). A cognitive model tracks the underlying knowledge requirements and thinking processes for solving a task. Hence, fine-grain size diagnostic inferences about examinees' cognitive skills in a specific domain can be made. Examinees will be able to discern their strengths and weaknesses through the diagnostic inferences provided and therefore be able to improve their learning.

One other benefit of using a cognitive model is the potential for linking cognition theory with learning and instruction. Instructional principles are decided on the basis of how examinees reason and solve items. The diagnostic inferences associated with examinees' knowledge and thinking processes may help instructors identify examinees' strengths and weaknesses and fine-tune, if necessary, their instructional strategies. Cognitive models provide one means to report examinees' cognitive skills on tasks of interest which could be used to associate their test score with instructional procedures designed to improve the examinees' skills (National Research Council, 2001; Pellegrino, 2002; Pellegrino, Baxter, Glaser, 1999).

To visually represent a cognitive model, a construct map (Wilson, 2005) can be used. Functioning as a cognitive model, the construct map is an approach for classifying levels of proficiency in a domain and the corresponding item responses in a coherent and succinct manner. To overcome the weaknesses of *scaling* and *item mapping* which are conducted post-hoc and restricted to specific inferences associated with a small set of test items, construct maps can guide item development and test construction and, therefore, connect the test content with interpretations about examinees' scores. Wilson (2005) provided this summary of a construct map:

> *The type of construct is one that is particularly suitable for a visual representation—it is called a construct map. Its most important features are that there is (a) a coherent and substantive definition for the content of the construct; and (b) an idea that the construct is composed of an underlying continuum—this can be manifest in two ways—an ordering of the respondents and/or an ordering of item responses. (p. 26)*

As presented in Figure 1, a generic construct map contains three parts: a double-sided arrow, the respondents, and their corresponding responses. The construct of interest, X, is composed of an underlying continuum extending from one extreme to another (Wilson, 2005). The left side of the map indicates qualitatively distinct groups of respondents, ranging with an ordering of knowledge and skills mastery levels on X domain. The right side of the map represents qualitative differences in item responses, ranging from responses indicating low mastery to those reflecting high mastery of X domain. Because both the respondents and item responses are conjunctively stated, higher X test performance implies that lower level of X knowledge and skills has been successfully mastered.

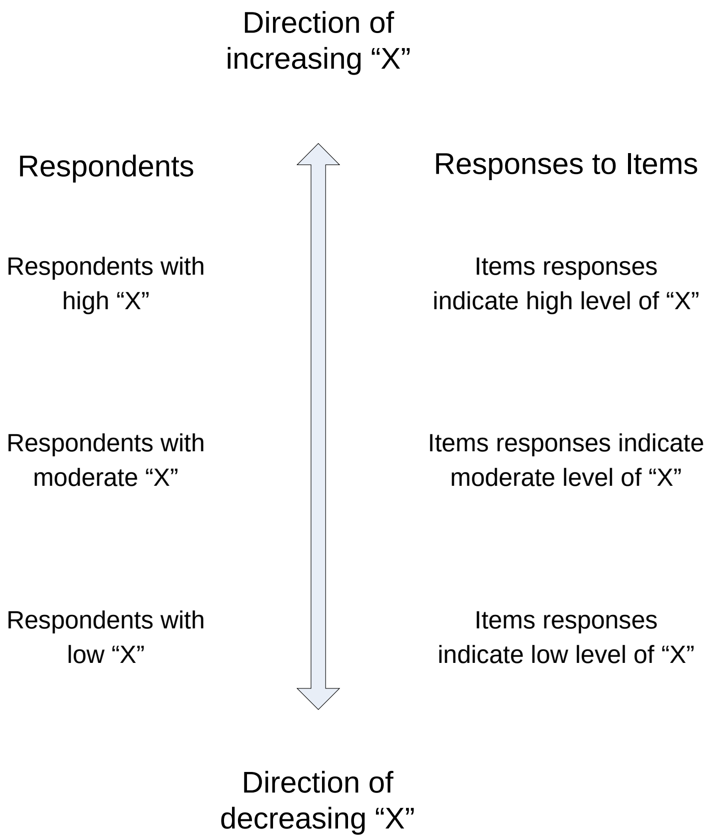


Figure 1. A generic construct map of construct “X”.

Stage 2: Model-Driven Item Development — Task and Item Models

Traditional item development using manual processes can be inefficient, largely because items are treated as isolated entities that are individually created, reviewed, and formatted. Because the items are individually authored, they yield unpredictable statistical outcomes (and, therefore, require field testing) because stimulus elements and conditions are not easily identified or well understood. Traditional item development can also pose security risks for a testing program because the costs associated with construction, calibration, and maintenance limit the number of operational items that are available at any one time—with fewer operational items available, exposure risks may increase because more examinees are being exposed to each item. Drasgow et al. (2006) describe this problem, as follows:

> *The demand for large numbers of items is challenging to satisfy because the traditional approach to test development uses the item as the fundamental unit of currency. That is, each item is*

> *individually hand-crafted—written, reviewed, revised, edited, entered into a computer, and calibrated—as if no other like it had ever been created before. A second issue with traditional approaches is that it is notoriously hard to hit difficulty targets, which results in having too many items at some levels and not enough at other levels. Finally, the pretesting needed for calibration in adaptive testing programs entails significant cost and effort. (p. 473)*

With assessment engineering, item development includes two continuous steps, task modeling and item modeling. To closely connect corresponding construct map which provides detailed descriptions of specific, fine-grain cognitive skills with item development stage, *task models* are created at specified levels of a construct map. Task modeling differ from traditional test development approach because it provides theoretical backing for item development by regulating features of assessment tasks such as stimulus elements and conditions. Next, item models are developed to represent relevant task models. Item models serve as an efficient and accurate item generation engine because multiple items produced by each item model adhere to strict quality controls and meet high psychometric standard consistently.

Task Model

One principled way to test knowledge and skills specified by a construct map is to first develop relevant task models. A task model is a generic profile of an assessment task which contains descriptions of knowledge and skills, descriptions of key features (e.g., objects and their properties, variables for difficulty variation) of the task, specifications of task representation material and any required condition, and classifications of response actions returned for scoring. Task models are created at different locations along a construct map and, in turn, each model provides measurement information in a particular region of the construct map.

A task model should specify features that control content dimensionality and task difficulty, such as task objects and their properties, nature of the relationships among objects, and cognitive level of the action(s) required by the task. Figure 2 presents a generic template for a task model. To build these models, the knowledge and skill measured by tasks and the required representation materials are provided. Elements that are incidental and radical are also listed[1]. In addition, classifications of response actions returned for scoring should be enumerated for score reporting purpose.

According to Luecht (2007), developing a task model is an iterative process. A task model only becomes useful if it is validated, which indicates that real data is needed to show that: (a) the task models can order themselves as expected (this is more or less control for difficulty with respect to the construct for the target population); (b) extraneous nuisance dimensionality is controlled; (c) each task model is capable of creating multiple item models and, in turn, to create multiple items; (d) what information is scored as well as which scoring evaluators are used.

Task Model

Knowledge and Skill Specification:

Rationale:

Candidate's Task:

Given Information:

Radical Elements:

Incidental Elements:

Auxiliary Information:

Tools and Resources Required:

Response Format:

Scoring:

Figure 2. A Generic Template of A Task Model.

[1] *Incidental* elements are the surface features of an item that do not alter item difficulty. To measure content at similar difficulty levels, incidental elements should be manipulated to generate items which are *isomorphic*. Conversely, *radical* elements are the deep features that alter item difficulty, and may also affect the psychometric properties of the test such as dimensionality. In order to measure content at different difficulty levels, besides the incidental elements, one or more radical elements must be manipulated to generate items that are *variant*.

Item Model

Task modeling provides theoretical backing for item development in the assessment engineering system while item modeling offers operational foundation in terms of efficient and valid item development. An item model[2] (LaDuca, Stamples, Templeton, & Holzman, 1986; Bejar, 1996, 2002) serves as an explicit representation of the clauses listed in a corresponding task model, which are embodied as the *stem*, the *options*, *key*, and oftentimes *auxiliary information* (Gierl, Zhou, & Alves, in press). The *stem* is the part of an item which formulates context, content, and/or the question the examinee is required to answer. The o*ptions* contain the alternative answers with one correct option and one or more incorrect options or distractors. When dealing with a multiple-choice item model, both stem and option are required. With an open-ended or constructed-response item model, only the stem is created. The *key* either specifies the correct option for a multiple-choice item model or lists criteria for an open-ended or constructed-response item model. *Auxiliary information* includes any additional material, in either the stem or option, required to generate an item, including texts, images, tables, and/or diagrams.

The stem and options can be divided further into *elements*. These elements are often denoted as strings (S) which are non-numeric values and integers (I) which are numeric values. By systematically manipulating these elements, which are incidental and/or radical elements, measurement specialists can generate large numbers of instances, at similar or different difficulty levels, for each model. One generic template of an item model is presented in Figure 3.

[2] Item models have been described in different ways. For example, they have been called schemas (Singley & Bennett, 2002), blueprints (Embretson, 2002), templates (Mislevy & Riconscente, 2006), forms (Hively, Patterson, & Page, 1968), and shells (Haladyna & Shindoll, 1989).

ITEM MODEL

Stem: xxx; **Options: xxx; Auxiliary Information: xxx**

xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx, xxxxxxxxxxxxxxxxxxxxxxxxxxxxxx?
A. xxxx
B. xxxx
C. xxxx
D. xxxx

STEM:

S1 xxxxxxxxxxxxx **S2** xxxxxxxxxxxxxxxxxxxxxxxx **I1**. xxxxxxxxx **S2** xxxxxxxxxx?

ELEMENTS:

S1 Range: “xx”, “xx”, “xx”, “xx”
S2 Range: “xx”, “xx”
I1 Range: xx to xx by xx

OPTIONS:

A. xxx
B. xxx
C. xxx
D. xxx

AUXILIARY INFORMATION:

xxx

KEY:

x

Figure 3. A Generic Template of An Item Model

Features of Task Modeling and Item Modeling

A task model concerns features for connecting a task with specific knowledge and skills. It provides a framework for constructing and describing the context in which examinees do, say, or make something to provide data about what they know or can do, more generally (Mislevy & Riconscente, 2006). Stated differently, a task model functions as a connector between a cognitive model and multiple item models because a task model describes the knowledge and skill specified in a construct map to an operational description required for item development.

Task models that provide the theoretical foundation for item development can proceed from either strong or weak theory. If strong theory is used, then task models are produced using the design principles articulated in a cognitive model (i.e., stage 1 in assessment engineering framework; see also Leighton & Gierl, 2007). The cognitive model provides a detailed description of the elements that affect examinee performance which, in turn, can help pinpoint the item difficulty features. The obvious benefit of strong theory is that the cognitive features of item elements are identified and articulated in such detail that difficulty can be predicted and controlled. Unfortunately, few strong theories currently exist to guide educational and psychological measurements. As a result, strong theory has been limited to specific tasks in narrow domains such as mental rotation (Bejar, 1990) and spatial ability (Embretson & Yang, 2007).

In the absence of strong theory, weak theory must be used. Task models generated according to weak theory using design guidelines (rather than design principles) discerned from a combination of experience, theory, and research (rather than cognitive models) (Drasgow et al., 2006). In other words, the way a task model is designed based on measurement specialists' existing experience. The benefit of weak theory for task modeling stems from its practicability. Task models can often be identified by reviewing items from previously administered exams. Weak theory is also well-suited to broad content domains where few theoretical descriptions exist about the cognitive knowledge and skills used by examinees to solve items. The main drawback of weak theory is that difficulty is neither easily predictable nor controlled.

Important features required for item development such as content domain, item difficulty, and cognitive level are considered in a task model. Therefore, each task model should be capable of generating multiple item models. Items produced by one item model should behave similarly in terms of their psychometric properties if only incidental elements are included in an item model; otherwise, items measuring different difficulty levels should be generated by an item model that contains radical elements.

Item models that efficiently assist item development can help overcome some of the limitations of the traditional approach thereby enhancing assessment development in two important ways. For example, item modeling is cost-effective. The purpose is to create multiple item models, where each item model yields many items. Hence, banks can be created quickly which will minimize item exposure because larger pools of operational items are available for each test administration. The logic behind item modeling can also lead to more cost-effective practices because items are treated as classes which require a systematic and strategic development approach compared with treating each item as a single unit. Therefore, the cost per item is lower because the unit of analysis is multiple instances per model rather than single instance per content specialist. Also, costly, yet common, errors in item development—including omissions or additions of words, phrases, or expressing as well as spelling, punctuation, capitalization, item structure, typeface, formatting, and language (e.g., English to French translation) problems—can be avoided because only specific elements in the stem and options are manipulated across large numbers of items. That is, the item model serves as a *template* where content specialists manipulate specific, well-defined, elements. The remaining components in the template, once finalized, are not altered during item development. As a result, item modeling should allow content specialists to quickly create large numbers of high-quality operational items that require few revisions during the development stage.

Item models also provide the foundation necessary for *automatic item generation*. Automatic item generation is a procedure for using item models to create instances with known item characteristics, often in real-time as the examinee is writing the test. The procedure has two requirements. An item class must be described in enough detail to permit a computer to create instances of the class automatically. Also, the variables that affect item difficulty must be controlled across instances so the generated items do not require separate calibration (Drasgow et al., 2006). One key benefit of automatic item generation is that it minimizes, if not eliminates, the need for extensive field testing because the instances generated from the parent model are *pre-calibrated* and, thus, do not need to be field tested. In short, item modeling can enhance test development practices and provide the necessary foundation for sophisticated psychometric procedures such as automatic item generation.

Stage 3: Automated Item Assembly

No test is developed using statistical criteria alone. A complex set of specifications based on content and statistical requirements must be considered for operational test assembly. Efficiently selecting items to develop multiple highly constrained test forms that meet these specifications from large item banks has become challenging, when using traditional manual test assembly procedures. Computer-based procedures for test assembly have been developed to address this test assembly problem (see a review by van der Linden, 1998; see also van der Linden, 2005). These procedures, referred to as *automated test assembly,* use computer algorithms to manage and construct tests in a much more efficient way than traditional manual assembly procedures.

Automated test assembly ensures that specification requirements (e.g., target information function) are optimally achieved by selecting the best set of items from the population of available items. Automated test assembly can be used to efficiently generate multiple test forms that are parallel in terms of both statistical and content specifications. Automated test assembly procedures contain three general continuous steps. The first step involves defining the characteristics of each item so that all pertinent psychometric features, content categories, and any other relevant item feature are mutually coded. This step is essentially the process of developing a well defined item bank. In the next step, a mathematical model that incorporates all psychometric and content specifications of the test is developed. Once the model is defined, optimization algorithms are applied to evaluate every possible solution relative to the target until the optimal or best possible combination of items is achieved. One problem that has prevented the wide-spread use of automated test assembly procedures is attributable to small, ill-defined item banks used for many testing agencies. By implementing task and item modeling procedures, this problem can be overcome.

Stage 4: Model-Data Fit Assessment

Traditionally, a model is explored through a validation study which occurs after a test is administered as well as examinees' response are collected to characterize the latent trait of interest. The model generated in such an exploratory analysis helps provide the information about the latent trait of interest, which reflects a single score presented on a unidimensional scale. No connection between cognitive psychology and educational testing is established.

From the perspective of assessment engineering, psychometric models focusing on the psychological features are employed in a *confirmatory*—versus exploratory—manner to assess the model-data fit relative to the intended underlying structure of the constructs the test is designed to measure. A model that is developed prior to test administration is to be confirmed by assessing the consistency between expected and observed responses to ensure cognitive principles are closely aligned with measurement practice, thereby providing a direct connection between cognitive theory and educational measurement.

Data collected from field test can be used for model-data fit analysis. The models should be statistically confirmed before formal test administration to ensure the psychology underlying task performance is well defined and organized by the specific psychometric approach. The outcomes from these model-data fit analyses also provide developers with guidelines for specific modifications to the cognitive and item models, as needed, to facilitate the acquisition of data that supports the intended assessment inferences (Luecht, Gierl, Tan, & Huff, 2006). The target score scale is therefore rich in meaning so fine-grained inferences can be made about examinees' knowledge and skills.

To summarize, assessment engineering differs from traditional approaches to test development and analysis in four fundamental ways. First, cognitive models guide item development, rather than content blueprints. Hence, the assessment principles used in test construction are much more specific allowing items to be created quickly and efficiently during the development cycle. Second, explicit item models are created to control and manipulate both the content and difficulty of the items. Content experts use the item models during development thereby producing assessment tasks that adhere to strict quality controls and that meet high psychometric standards consistently. Third, automated test assembly procedures are employed to build assessments that function to exacting specifications. Therefore, multiple test forms can be created from a bank of items very efficiently according to both content and statistical specifications. Fourth, psychometric models are employed in a *confirmatory* manner to assess the model-data fit relative to the intended underlying structure of the constructs or traits the test is design to measure. The outcomes from these model-data fit analyses also provide developers with guidelines for specific modifications to the cognitive and item models, as needed, to facilitate the acquisition of data that supports the intended assessment inferences. In short, assessment engineering directs the design and development as well as the analysis, scoring, and reporting of assessment results using engineering-based principles.

Section II

Applying Assessment Engineering Principles to the CPA Examination

The computer-based Uniform CPA Examination is a licensure test for Certified Public Accountants (CPAs). The examination consists of multiple-choice questions and Simulations (case studies) that evaluate the knowledge and skills required of entry-level CPAs. For the purposes of this study, we worked with a draft outline of skills. The two main skill areas assessed in Simulations are *Communication* and *Application of the Body of Knowledge* (*ABK*). Skills for these two areas are evaluated by three sections: Audit and Attestation (AUD), Financial Accounting and Reporting (FAR), and Regulation (REG).

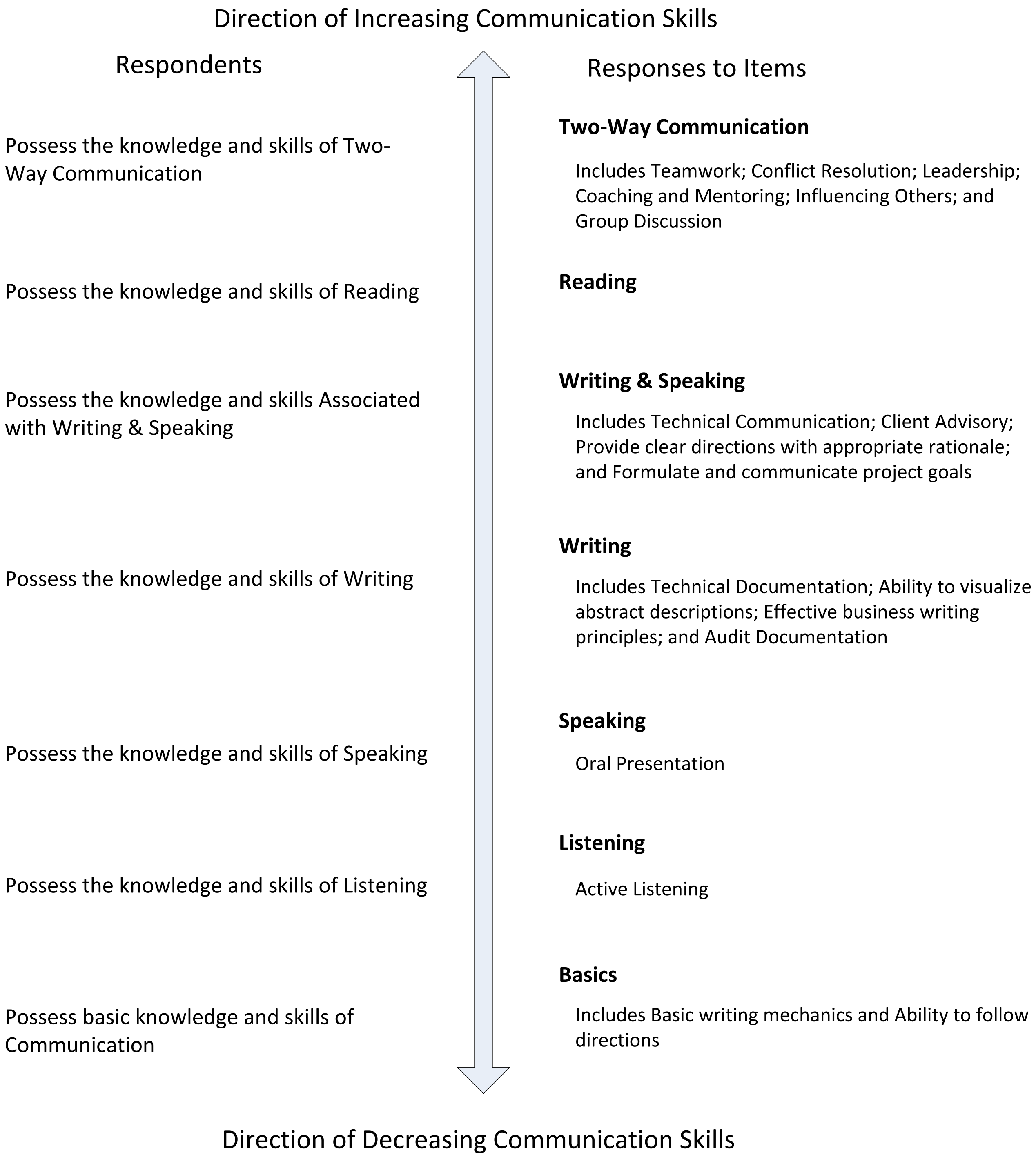
Stage #1: Developing Construct Maps for CPA Examination

Knowledge and skills measured in *Communication* and *ABK* are classified by their cognitive complexity. However, only three levels—low, medium, and high—are included in the cognitive category. The three-level category is not sufficient to categorize cognitive complexity and therefore provides only a vague classification of knowledge and skills assessed in Simulations.

To apply the principles of assessment engineering to the Simulations in the CPA Examination, construct maps were created. As a general rule, knowledge and skills are jointly described at each level in a construct map. Higher performance assumes that lower level knowledge and skills have been successfully mastered.

Because the skill areas assessed in Simulations are different in their cognitive nature, different cognitive categories were applied to the two areas. The skill area of *Communication* was classified using a six-level communication category. The categories include: Basics, Listening, Speaking, Writing, Reading, and Two-way Communication. Tasks classified in Basics level indicate they are basic in nature and are prerequisite to skills in all other levels. Listening, Speaking, Writing, and Reading are natural categories of communication. They are ordered from relatively low cognitive complexity, listening, to relatively high cognitive complexity, reading. Skills classified in Two-way Communication require back-and-forth communication for assessment and therefore are treated as the highest level in the category.

With content specialists and psychometricians' careful review, evaluation, reasoning, and discussion of 18 different knowledge and skills assessed in *Communication*, these knowledge and skills were classified using the communication category. Four skills combine the skills of Writing and Speaking. A construct map was then developed by mapping classified knowledge and skills along the left side of the construct map continuum, from lowest (Basics) to highest (Two-way Communication) proficiency, and the corresponding item response performance on the right side of the continuum. Figure 4 presents the construct map of *Communication* measured in Simulations section of CPA Examination.



Direction of Increasing Communication Skills

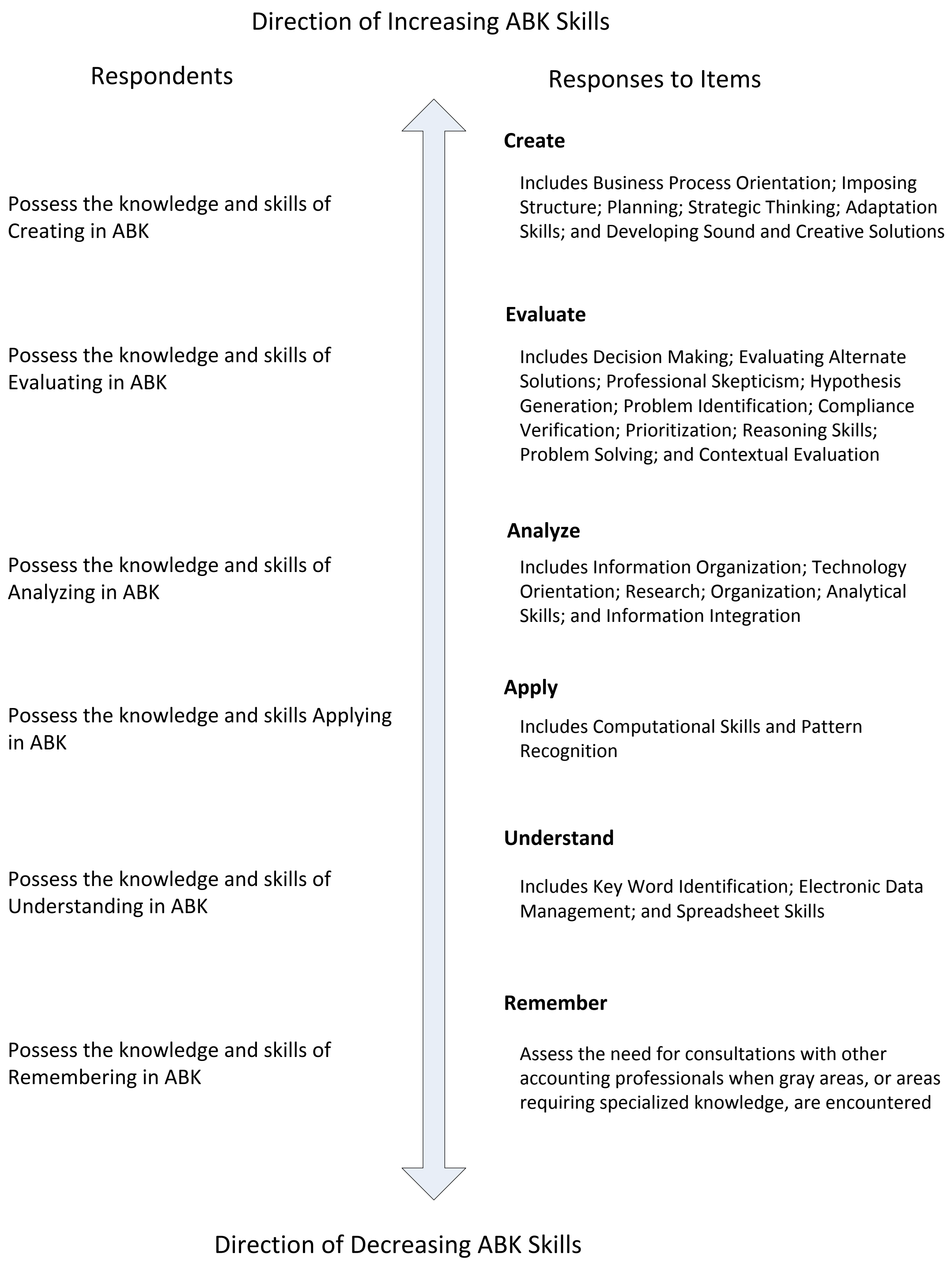
Respondents	Responses to Items
Possess the knowledge and skills of Two-Way Communication	**Two-Way Communication** Includes Teamwork; Conflict Resolution; Leadership; Coaching and Mentoring; Influencing Others; and Group Discussion
Possess the knowledge and skills of Reading	**Reading**
Possess the knowledge and skills Associated with Writing & Speaking	**Writing & Speaking** Includes Technical Communication; Client Advisory; Provide clear directions with appropriate rationale; and Formulate and communicate project goals
Possess the knowledge and skills of Writing	**Writing** Includes Technical Documentation; Ability to visualize abstract descriptions; Effective business writing principles; and Audit Documentation
Possess the knowledge and skills of Speaking	**Speaking** Oral Presentation
Possess the knowledge and skills of Listening	**Listening** Active Listening
Possess basic knowledge and skills of Communication	**Basics** Includes Basic writing mechanics and Ability to follow directions

Direction of Decreasing Communication Skills

Figure 4. A Construct Map of Communication Skills in CPA Examination

A revised Bloom’s Taxonomy (RBT, Anderson & Krathwohl, 2001) was used to classify cognitive levels of complexity for the area of *ABK* in Simulations section of CPA Examination. The RBT still contains six categories of processes. Bloom's six major categories, however, were changed from noun to verb forms. The lowest level of the original version, knowledge, became remembering; comprehension and synthesis were changed to understanding and creating. In addition, the order of the top two levels is changed in the revised version. Appendix 1 provides a brief definition and example of each cognitive process in the RBT, lists their alternative names, and indicates the category to which it belongs. These 19 specific cognitive processes are intended to be mutually exclusive; together they delineate the breadth and boundaries of the six categories.

Again, with content specialists and psychometricians’ review of 28 different knowledge and skills assessed in *Application of the Body of Knowledge*, these knowledge and skills were mapped to the six categories defined in RBT. A construct map (see Figure 5) was then developed by mapping classified knowledge and skills along the left side of the construct map continuum, from lowest (Remembering) to highest (Creating) proficiency, and the corresponding item response performance on the right side of the continuum.



Direction of Increasing ABK Skills

Respondents

Responses to Items

Create

Possess the knowledge and skills of Creating in ABK

Includes Business Process Orientation; Imposing Structure; Planning; Strategic Thinking; Adaptation Skills; and Developing Sound and Creative Solutions

Evaluate

Possess the knowledge and skills of Evaluating in ABK

Includes Decision Making; Evaluating Alternate Solutions; Professional Skepticism; Hypothesis Generation; Problem Identification; Compliance Verification; Prioritization; Reasoning Skills; Problem Solving; and Contextual Evaluation

Analyze

Possess the knowledge and skills of Analyzing in ABK

Includes Information Organization; Technology Orientation; Research; Organization; Analytical Skills; and Information Integration

Apply

Possess the knowledge and skills Applying in ABK

Includes Computational Skills and Pattern Recognition

Understand

Possess the knowledge and skills of Understanding in ABK

Includes Key Word Identification; Electronic Data Management; and Spreadsheet Skills

Remember

Possess the knowledge and skills of Remembering in ABK

Assess the need for consultations with other accounting professionals when gray areas, or areas requiring specialized knowledge, are encountered

Direction of Decreasing ABK Skills

Figure 5. A Construct Map of Application of the Body of Knowledge (ABK) Skills in CPA Examination

Stage #2: Developing Task and Item Models

Item development is the second task required to apply the principles of assessment engineering towards the CAP Examination. Currently, items in the CPA Examination are individually created by test developers. In this project, task models and item models were designed and created using model-driven item development. To begin, task models were designed and developed. A task model is a generic profile of an assessment task which contains descriptions of knowledge and skills intended to measure, descriptions of key features (e.g., objects and their properties, variables for difficulty variation) of the task, specifications of task representation material and any required condition, and classifications of response actions returned for scoring. A task model considers features that control content dimensionality and task difficulty, such as task objects and their properties, nature of the relationships among objects, and cognitive level of the action(s) required by the task.

Because of the absence of strong theory, task models in this project were constructed by reviewing items from previously administered items. Item models were then developed with the guidance of each task model. Item stem, variable elements, and their possible constraints are designated in each item model to provide the conditions for generating multiple instances. Because the incidental and radical elements were defined in the parenting task model, psychometric characteristics of instances generated from one item model are known. Due to security considerations however, an example is not presented in this technical report.

The two assessment engineering stages applied to CPA Examination in this project are indispensible. Cognitive model development, the first stage, specifies knowledge and skills underlie examinees' task performance and, therefore, facilitates accurate and efficient item development. Next, item development is driven by task and item modeling. Task models developed closely connect to corresponding cognitive model and describe important features of tasks designed for assessing certain skills. Item models designed to embody elements in task models can function like an engine to efficiently generate multiple instances.

Section III

Summary, Limitations, and Future Research Directions

Summary of Current Study

The purpose of the current study was to apply principles of assessment engineering to CPA Examination focusing, specifically, on the Simulations component. The skills statements on *Communication* and *ABK*, the two main skill areas intended to assess with Simulations tasks, were carefully reviewed, evaluated, and organized as construct maps. The skill area of *Communication* was classified using a six-level communication category. The categories include: Basics, Listening, Speaking, Writing, Reading, and Two-way Communication. A revised version of Bloom's Taxonomy was used for *ABK* knowledge and skills classification. The six categories in the taxonomy, from remembering to creating, were ordered by cognitive levels of complexity.

Task and item models development is the second stage in the framework of assessment engineering framework. Task models, which can direct item model development, and item models, which can generate either isomorphic or variant instances, were developed. The benefits of using task and item models for test bank development lie in the facility to, first, control content representation and, second, have a sense of the difficulty levels of items. In addition, a test bank with large numbers of high quality items can be created using task and item models.

Directions for Future Research

In this research project, construct maps on the skill areas of *Communication* and *ABK* were developed by content specialists but has not been validated using a sample of examinees from the target population. Validating the cognitive model is a critical step before the model can be used in practice. To address this problem, examinee response data should be collected with verbal think-aloud methods to evaluate the knowledge structures and processing skills used by a sample of CPA test-takers to solve the Simulations items. Once the model is validated with the population of interest, task and item models can be created to generate items that measure specific components of the model thereby providing developers with a way of controlling the cognitive attributes measured by the test.

For each specific skill measured in the Simulations section, several tasks were designed. These tasks were used as parent items for guiding task and item models development. These connections between skill and tasks were, again, set up by content specialists and have not been validated. One suggestion is to evaluate whether the tasks measure the skills correctly.

When cognitive model and task loading are validated, the link from cognitive model to skills and then to specific tasks is established. Recall, one merit of using a cognitive model in test developments lies in its facility to yield detailed cognitive diagnostic feedback to the examinees about their problem-solving strengths and weaknesses (Leighton & Gierl, 2007). In addition, a cognitive model has the potential of linking cognition theory with learning instruction. The diagnostic inferences associated with examinees' mastery of knowledge and thinking processes assist instructors to observe examinees' strengths and weaknesses and thereby fine-tune, if necessary, the instructional principles and strategies to improve the examinees' skills. Hence, score reporting on CPA Examination can include cognitive diagnostic information such as skills that tasks are intended to measure and test-takers' cognitive strengths and weaknesses. With these cognitive inferences, CPA Examination-takers can identify their proficiency levels of specific tasks examined.

References

Anderson, L. W., & Krathwohl, D. R. (Eds.). (2001). *A taxonomy for learning, teaching and assessing: A revision of Bloom's Taxonomy of educational objectives.* New York : Longman.

Bejar, I. I. (1990). A generative analysis of a three-dimensional spatial task. *Applied Psychological Measurement, 14*, 237-245.

Bejar, I. I. (1996). *Generative response modeling: Leveraging the computer as a test delivery medium* (ETS Research Report 96-13). Princeton, NJ: Educational Testing Service.

Bejar, I. I. (2002). Generative testing: From conception to implementation. In S. H. Irvine & P. C. Kyllonen (Eds.), *Item generation for test development* (pp.199-217). Hillsdale, NJ: Erlbaum.

Chi, M. T. H. (1997). Quantifying qualitative analyses of verbal data: A practical guide. *Journal of the Learning Sciences, 6*, 271-315.

Drasgow, F., Luecht, R. M., & Bennett, R. (2006). Technology and testing. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 471-516). Washington, DC: American Council on Education.

Embretson, S. E. (2002). Generating abstract reasoning items with cognitive theory. In S. H. Irvine & P. C. Kyllonen (Eds.), *Item generation for test development* (pp. 219-250). Mahwah, NJ: Erlbaum.

Embretson, S. E., & Yang, X. (2007). Automatic item generation and cognitive psychology. In C. R. Rao & S. Sinharay (Eds.) *Handbook of Statistics: Psychometrics*, Volume 26 (pp. 747-768). North Holland, UK: Elsevier.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data*. Cambridge, MA: The MIT Press.

Gierl, M. J., Zhou, J., & Alves, C. (in press). *Developing a taxonomy of item model types to promote assessment engineering.* Paper submitted to Journal of Technology, Learning, and Assessment.

Haladyna, T., & Shindoll, R. (1989). Items shells: A method for writing effective multiple-choice test items. *Evaluation and the Health Professions, 12*, 97-106.

Hively, W., Patterson, H. L., & Page, S. H. (1968). A “universe-defined” system of arithmetic achievement tests. *Journal of Educational Measurement, 5*, 275-290.

LaDuca, A., Staples, W. I., Templeton, B., & Holzman, G. B. (1986). Item modeling procedures for constructing content-equivalent multiple-choice questions. *Medical Education, 20,* 53-56.

Leighton, J. P. (2004). Avoiding misconceptions, misuse, and missed opportunities: The collection of verbal reports in educational achievement testing. *Educational Measurement: Issues and Practice, 23*, 6-15.

Leighton, J. P., & Gierl, M. J. (2007). Defining and evaluating models of cognition used in educational measurement to make inferences about examinees' thinking processes. *Educational Measurement: Issues and Practice, 26,* 3-16.

Luecht, R. M. (2006a, May). *Engineering the test: From principled item design to automated test assembly*. Paper presented at the annual meeting of the Society for Industrial and Organizational Psychology, Dallas, TX.

Luecht, R. M. (2006b, September). *Assessment engineering: An emerging discipline.* Paper presented in the Centre for Research in Applied Measurement and Evaluation, University of Alberta, Edmonton, AB, Canada.

Luecht, R. M. (2007, February). *Assessment engineering workshop*. Presented at Association of Test Publishers Conference. Palm Spring, CA.

Luecht, R. M., Gierl, M. J., Tan, X., & Huff, K. (2006, April). *Scalability and the development of useful diagnostic scales.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement*. Washington, DC: American Council on Education.

Mislevy, R. J., & Riconscente, M. M. (2006). Evidence-centered assessment design. In S. M. Downing & T. Haladyna (Eds.), *Handbook of test development* (pp. 61-90). Mahwah, NJ: Erlbaum.

National Research Council (2001). *Knowing what students know: The science and design of educational assessment.* Committee on the Foundations of Assessment. J. Pellegrino, N. Chudowsky, and R. Glaser (Eds.). Board on Testing and Assessment, Center for Education. Washington, DC: National Academy Press.

Pellegrino, J. W. (2002). *Understanding how students learn and inferring what they know: Implications for the design of curriculum, instruction, and assessment*. In M. J. Smith (Ed.), NSF K-12 Mathematics and Science Curriculum and Implementation Centers Conference Proceedings (pp. 76-92). Washington, DC: National Science Foundation and American Geological Institutue.

Pellegrino, J. W., Baxter, G. P., & Glaser, R. (1999). Addressing the "two disciplines" problem: Linking theories of cognition and learning with assessment and instructional practices. In A. Iran-Nejad & P. D. Pearson (Eds.), *Review of Research in Education* (pp. 307-353). Washington, DC: American Educational Research Association.

Singley, M. K., & Bennett, R. E. (2002). Item generation and beyond: Applications of schema theory to mathematics assessment. In S. H.

Irvine & P. C. Kyllonen (Eds.), *Item generation for test development* (pp. 361-384). Mahwah, NJ: Erlbaum.

Taylor, K. L., & Dionne, J-P. (2000). Accessing problem-solving strategy knowledge: The complementary use of concurrent verbal protocols and retrospective debriefing. *Journal of Educational Psychology, 92*, 413-425.

van der Linden, W. J. (1998). Optimal assembly of psychological and educational tests. *Applied Psychological Measurement, 22*, 195-211.

van der Linden, W. J. (2005). *Linear models for optimal test design*. New York: Springer.

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

Appendix 1

The Cognitive Process Dimension in the Revised Bloom's Taxonomy

Categories & Cognitive Processes	**Alternative Names**	**Definitions and Examples**
1. REMEMBER – Retrieve relevant knowledge from long-term memory		
1.1 Recognizing	Identifying	Locating knowledge in long-term memory that is consistent with presented material (e.g., Recognize the dates of important events in U.S. history)
1.2 Recalling	Retrieving	Retrieving relevant knowledge from long-term memory (e.g., Recall the dates of important events in U.S. history)
2. UNDERSTAND—Construct meaning from instructional messages, including oral, written, and graphic communication		
2.1 Interpreting	Clarifying Paraphrasing Representing Translating	Changing from one from one form of representation (e.g., numerical) to another (e.g., verbal) (e.g., Paraphrase important speeches and documents)
2.2 Exemplifying	Illustrating Instantiating	Finding a specific example or illustration of a concept or principle (e.g., Give examples of various artistic painting styles)
2.3 Classifying	Categorizing Subsuming	Determining that something belongs to a category (e.g., concept or principle) (e.g., Classify observed or described cases of mental disorders)
2.4 Summarizing	Abstracting Generalizing	Abstracting a general theme or major point(s) (e.g., Write a short summary of the events portrayed on a videotape)
2.5 Inferring	Concluding Extrapolating Interpolating Predicting	Drawing a logical conclusion from presented information (e.g., In learning a foreign language, infer grammatical principles from examples)

2.6 Comparing	Contrasting Mapping Matching	Detecting correspondences between two ideas, objects, and the like (e.g., Compare historical events to contemporary situations)
2.7 Explaining	Constructing models	Constructing a cause-and-effect model of a system (e.g., Explain the causes of important 18-century events in France)
3. APPLY – Carry out or use a procedure in a given situation		
3.1 Executing	Carrying out	Applying a procedure to a familiar task (e.g., Divide one whole number by another whole number, both with multiple digits)
3.2 Implementing	Using	Applying a procedure to an unfamiliar task (e.g., Use Newton's Second Law in situations in which it is appropriate)
4. ANALYZE – Break material into its constituent parts and determine how the parts relate to one another and to an overall structure or purpose		
4.1 Differentiating	Discriminating Distinguishing Focusing Selecting	Distinguishing relevant from irrelevant parts or important from unimportant part of presented material (e.g., Distinguish between relevant and irrelevant numbers in a mathematical word problem)
4.2 Organizing	Finding coherence Integrating Outlining Parsing Structuring	Determining how elements fit or function within a structure (e.g., Structure evidence in a historical description into evidence for and against a particular historical explanation)
4.3 Attributing	Deconstructing	Determine a point of view, bias, values, or intent underlying presented material (e.g., Determine the point of view of the author of an essay in terms of his or her political perspective)
5. EVALUATE – Make judgments based on criteria and standards		

5.1 Checking	Coordinating Detecting Monitoring Testing	Detecting inconsistencies or fallacies within a process or product; determining whether a process or product has internal consistency; detecting the effectiveness of a procedure as it is being implemented (e.g., Determine if a scientist's conclusions follow from observed data)
5.2 Critiquing	Judging	Detecting inconsistencies between a product and external criteria, determining whether a product has external consistency; detecting the appropriateness of a procedure for a given problem (e.g., Judge which of two methods is the best way to solve a given problem)
6. CREATE – Put elements together to form a coherent or functional whole; reorganize elements into a new pattern or structure		
6.1 Generating	Hypothesizing	Coming up with alternative hypotheses based on criteria (e.g., Generate hypotheses to account for an observed phenomenon)
6.2 Planning	Designing	Devising a procedure for accomplishing some task (e.g., Plan a research paper on a given historical topic)
6.3 Producing	Constructing	Inventing a product (e.g., Build habitats for specific purpose)