

*Incorporating Innovations in Psychometric and Cognitive Theory into Operational Tests:*

A special issue of the Journal of Applied Testing Technology

Matthew J. Burke

AICPA

Creating successful large-scale, high stakes operational tests is a balancing act. Adherence to established industry standards must be weighed against a constant need for modernization. Large-scale assessment is a field marked by continuous innovation for the explicit purpose of improving the measurement process. The gulf between what is practical and what is possible is continually evaluated by test developers in the hope of finding solutions where what theoreticians deem to be “best practice” can be implemented operationally. Two general classes of theory provide much of the guidance to operational testing companies and organizations: psychometric theory and cognitive theory. It is widely recognized that much work must be done to ensure that innovations in these theories translate into innovation in practice. The goal of the following papers is to communicate the lessons learned by three large-scale testing organizations when implementing principled assessment design and cognitive theory in an operational examination.

In recent years, a great deal of attention has focused on the use of modern psychometric theory, namely principled approaches to assessment (e.g., evidence-centered design [ECD]), to guide and govern test development, administration, and scoring. Unfortunately, the vast majority of this attention has been focused on theoretical or hypothetical treatments of these frameworks. Additionally, cognitive science is constantly pressed into service to provide explanations for the thought processes related to any of the myriad endeavors individuals undertake. Understanding the characteristics of human thought provides a greater opportunity to accurately represent

subject matter for the sake of assessment. These articles address some of the challenges associated with putting principled assessment design and cognitive science into operational practice.

The viability of these principled approaches to assessment has yet to be determined in operational settings. One of the reasons why the viability remains in question is that operational examinations tend to have unique characteristics arising from different practical development concerns. No two operational testing programs are exactly alike, and as such, the idiosyncratic operational constraints that characterize these tests muddy the water when test developers implement these theoretical techniques. In plain language, the unique needs of operational exams make it difficult to merge theory with practice in a uniform way. It is likely that every attempt to put theoretical innovations into place in an operational setting will have its own unique challenges. However, the identification and resolution of these unique challenges creates an opportunity to discuss the successes and setbacks faced when trying to operationalize innovative psychometric frameworks.

As operational testing programs attempt to incorporate these innovations, it is of tremendous practical importance that the obstacles faced and lessons learned be communicated to the testing community. Despite the unique character of the many programs (e.g., content, test format, item types, purpose) that may choose to implement principled assessment frameworks, a great deal of commonality still exists (e.g., the need for reliable and valid assessments), which may allow others to learn from the shared experiences of three large-scale examinations. These innovative approaches to testing offer the promise of adding value to the process of high stakes assessment. As such, the dissemination of findings related to the nature and efficacy of implementation should be disseminated to the testing community at large.

The article by Luecht presents a treatment of Assessment Engineering (AE) and how it applies to modern test development, delivery, and scoring. This article serves to inform readers about what AE is and how its application to test development offers potential practical benefits to the user. In many ways, AE is a highly structured and formalized manufacturing-engineering process for designing and implementing cognitively based assessments envisioned under an evidence-centered design (ECD) framework for summative or formative purposes (Mislevy, 1994, 2006; Mislevy, Steinberg, & Almond, 2003). The article serves to make the application of a principled assessment framework, like ECD, regimented and predictable.

The article by Hendrickson, Huff, Ewing, and Kaliski documents work being done by the College Board to implement ECD in the large-scale, high stakes Advanced Placement exams that are currently being redesigned. The redesign of these assessments involves two major challenges: the need to create courses and exams that reflect advances in what is understood about how students learn and the need to ensure comparability of scores within and across years, complicated by the complex nature of mixed-format assessments, which imposes significant pressure on equating methods. The College Board is addressing these two challenges simultaneously through ECD. The bulk of the article addresses the challenges and lessons learned from implementing ECD for this exam program, including logistical constraints (e.g., policy decisions, costs and resource constraints), as well as divergent approaches to and philosophies about testing. This article demonstrates how ECD offers the promise of adding value in the large-scale, multiple exam structure of the AP program.

Luebke's and Lorie's article details the Law School Admission Council's use – beginning in 1990 – of Bloom's Taxonomy (Bloom et al., 1956) to guide Reading Comprehension item development and test construction for the LSAT. The paper also provides data showing the

extent to which this use was operationally successful and discusses how some practical item development considerations impact its effectiveness. This article demonstrates the longevity of added value when cognitive models are used to aid in test development.

Finally, the article by Burke, DeVore, and Stopek looks at the work being done at the American Institute of Certified Public Accountants (AICPA) to implement aspects of AE (Luecht, this issue) in the Uniform CPA Examination. The article focuses on work underway to use Cognitive Task Analysis (Clark, Feldon, van Merriënboer, Yates, & Early, 2008) and the Revised Bloom's Taxonomy (Anderson & Krathwohl, 2001) as a way to review the skills measured in the CPA exam as part of implementing AE. This article shows how the AICPA is dealing with the challenge of retrofitting AE in an ongoing, operational exam.

The information presented in these articles is designed to elicit thoughtful inquiry on the part of the readers about the characteristics of their own tests and testing programs. Each of these articles offers some advice or guidance on how to better incorporate advances in cognitive and psychometric science in operational tests. Taken together, these articles offer readers some insight into how and why cognitive science and psychometric theory can be used to provide practical benefits in operational measurement settings.