

**KEY FACTORS TO CONSIDER WHEN ENGAGING IN A DEVELOPMENT OR
REDEVELOPMENT PROCESS FOR EXAMINATIONS**
(CALIFORNIA'S TWO-DAY BAR EXAMINATION PROPOSAL)

Chad W. Buckendahl, PhD
Alpine Testing Solutions, Inc.
July 15, 2013

This report provides a short overview of key factors to consider when engaging in a development or redevelopment process for examinations. Contained in the report are specific comments that are intended to clarify intuitive perceptions of validity and reliability that may unintentionally mislead the Committee of Bar Examiners (Committee) without further explanation.

There are a number of elements that can influence exam development practices, but for the purposes of this topic, the paper focuses on three broad categories:

- *Content representation*
- *Empirical characteristics*
- *Non-cognitive factors*

Tests are not valid or invalid; validity is a property of scores based on defined, intended interpretations and uses. Each of the elements in the categories noted above can influence the validity of inferences made about scores from an examination. Figure 1 below illustrates the multiple, integrated phases of test development and validation that collectively contribute to the overarching concept of validity. No one element stands alone. Content representation includes elements of program design, test design, domain analysis, blueprint development (which includes weighting), content development, and content review. Empirical characteristics are then reflected in the pretesting, analysis, operational test assembly, standard setting, and custom validation studies as part of test maintenance. In addition, non-cognitive factors such as practical, policy, legal, and business can contribute to or threaten valid score interpretations; these cannot be disentangled from the psychometric considerations and are generally disclosed and discussed in the program design phase of development or redevelopment.

ATTACHMENT D

Figure 1. Test Development and Validation Process



A validation framework is based on industry expectations promulgated in the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999) in conjunction with the intended interpretations and uses of scores for licensure testing programs and was used as a foundation for the comments that follow.

During the public forum that was held by the Committee in 2013, it was suggested that if validity and reliability were equivalent in the proposed 2-day exam when compared with the existing 3-day exam, certain commentators would support it. Because validity is the overarching goal for any credible testing program, this is a reasonable conclusion.

However, a comment that came up during the public forum warrants additional discussion. Specifically, the suggestion that if the Cal Bar Exam does not include one of the performance tests, it is a delegation of responsibility to the law schools, is over-interpreting the intent of the purpose for reconfiguring the examination. One of the primary purposes of a professional licensure examination is to provide independent evidence that candidates possesses sufficient competency for entry-level practice. It would be inappropriate to confound that intent with the purposes of educational training programs or accreditation activities associated with that program. It would be important for the Cal Bar Exam to communicate that any changes to the examination are not intended to abdicate nor delegate its responsibility to adequately sample from the domain to support the licensure process.

The following comments focus on the feedback provided by another commentator. A commentator suggested that the Cal Bar Exam should oppose changes that would *negatively* impact the validity of inferences made from scores in the examination.

Although the comment was non-directional, the implication was that the proposed changes would negatively impact the program without acknowledging that the proposed changes could also have a *positive* impact on the validity of inferences that Cal Bar seeks to make regarding candidates' competency. Statements were made related to four areas of concern: impact on curriculum and instruction at law schools, changing the weight of the examination, reducing the number of essay questions, and reducing the length of time on the performance test. I provide some general comments regarding the feedback and then respond to each specific area of concern separately.

General Comments: There were assertions that the Cal Bar Exam is the “best in the country” and that the proposed changes would “dumb it down.” Evidence to support these assertions was not provided and the comments are driven by an intuitive perception of validity. Further, although one source of measurement error is related to decisions made on scores, these are not the only sources of error that examination programs consider. Other sources of measurement error that the Cal Bar Exam needs to consider include the sampling of content from the domain, prompt/task selection, rater error, and number of observations. Collectively, this evidence contributes to information about reliability which then further contributes to the overarching concept of construct validity. Within the feedback, there is also a confusion of reliability (i.e., the estimation of measurement error from multiple sources) and equating (i.e., the content and empirical equivalence of exam scores across test forms and years).

Impact on curriculum and instruction at law schools: When tests are designed to align with educational programs, the inter-relationships among curriculum, instruction, and assessment are important. However, for licensure and professional certification programs, from the perspective of the testing program, this connection is much less important because the intent of the examination program is to reflect entry-level *practice*, not any particular school's curriculum and instructional practices. Although often misused for such purposes, licensure testing program scores are not intended to serve as a comprehensive evaluation of a program's curriculum and instruction. It is appropriate to gather input from training programs in the process of defining the domain, however, practice drives the content and design of licensure testing programs.

Changing the weights of the examination from 65% constructed response/35% selected response to 50% constructed response/50% selected response: The first of the subsequent three criticisms relates to the proposed shift in the weighting of the examination from the existing distribution of item types. The validity argument for licensure testing programs begins with

evidence that the test is representative of the intended, job-related domain. That representation includes considerations of both content and cognitive demand (e.g., memorization, application, analysis), but must align with the expectations of the target candidate for the program. These weights are most often informed through a combination of informed judgment and surveys of practitioners that will yield empirical evidence. The methods that systematically collect this evidence are called practice, occupational, or job analysis. Because the Cal Bar Exam does not appear to have results of a recent practice analysis to inform these weights, the examination committee will make the determination of how content should be represented on the examination.

All examinations are a sample of a larger domain of interest with a challenge to balance breadth and depth of measurement. Although selected response items can be written to measure higher order thinking skills (e.g., application, analysis), they are more often used to more efficiently measure the breadth of the domain to broadly sample from the important content areas. Conversely, constructed response items do not generally provide a much cost-benefit with respect to breadth of coverage of the domain and are typically used to sample from important components where candidates can demonstrate the depth of their abilities on a narrow range of topics. The current weighting of the examination is heavily weighted to the constructed response components. In addition to some of the practical considerations for reducing the overall length of the examination, it is likely that the overweighting of a smaller number of topics is having a negative impact on validity. To repeat, by shifting the weighting of the examination to a balance of breadth and depth of cognitive measurement, the Cal Bar Exam will improve the validity of the inferences it is making on the examination due to an increased representation of the domain.

As Dr. Klein's white paper correctly noted, the shift in distribution will likely improve reliability as estimated by internal consistency and decision consistency. Although these are empirical considerations, reliability broadly refers to estimation of measurement error. A lack of content representation of the domain of interest with respect to the overall decision negatively impacts validity. Because reliability is a necessary, but insufficient source of evidence for valid interpretations, the increase in reliability coupled with the better representation of content ultimately improves the overall validity of the program. The increase in reliability is a nice artifact, but even if reliability were to remain unchanged, the increased sampling of the domain would contribute to more valid decisions based on the scores.

As a final thought on this topic, there are very few things that statisticians dislike more than attempts to use anecdotes as evidence that should be generalized; attorneys would probably characterize this as “hearsay.” Anecdotal comments from a range of individuals concerned about entry-level practitioners’ writing skills are not based on systematic collection of evidence that supports the assertion. Further, the Cal Bar Exam is not intended to be a test of writing skills, regardless of intuitive perspectives to the contrary. Even assuming the writing skills of candidates are as poor as were asserted, knowing that the exam is not intended to target writing skills, adjusting the distribution of item type in any direction will not have an influence on an evaluation of the skill. Because decisions on the examination are proposed to be based on 50% constructed response items, the message to law schools is that the Cal Bar Exam is going to increase the breadth of coverage while still retaining a substantive part of the exam that will be based on responses that candidates produce.

Reducing the number of essay questions from six (6) to five (5) and performance tests from two (2) to one (1): As an extension of the concern raised regarding the shift in weighting, the focus on item types as opposed to domain representation is misleading. Constructed response items are not inherently better or worse than selected response items; the use of these is a function of the intended measurement target – content and cognitive demand. The number of essay questions and performance tests are still sufficient to provide multiple opportunities to evaluate the depth of candidates’ abilities on selected topics. And to reiterate a point made above: the Cal Bar Exam is not intended to measure writing skills (e.g., ideas, organization, conventions); rather the intended target focuses on a candidate’s ability to communicate their legal reasoning and analysis.

Reducing the length of time for the performance test from three (3) hours to one and a half (1.5) hours: When designing an examination, it is important to consider the amount of measurement information that can be collected for a given item type – again considering the breadth and depth of the intended domain. The potential for reducing a single task from 3 hours to 1.5 hours permits the Cal Bar Exam to collect some information at a given level of depth, while balancing the competing need for broader representation of content. These item types are not intended to be analogous to a classroom assessment or used for programmatic evaluation. They have a specific purpose to target measurement around the point of distinction between minimally competent or not. No one has argued that constructed response items should be removed from

the exam entirely, only that their use be more efficient when considering psychometric, practical, administration, scoring, and business factors.

The Cal Bar Exam is a power, not a speeded test. As a result, if there is a factor of fatigue that represents a significant component of the examination, this is problematic. Introducing factors like speed or fatigue into the process represents construct irrelevant variance threatening the validity of the intended inferences. Although it was communicated as an anecdote, if candidates are unable to complete the examination or if there is a real or perceived factor of “stamina” being part of what is being measured, this is something the committee should strive to correct as unintended.

As a final comment, although every state and testing program has a certain level of pride associated with the efforts to produce and maintain their programs, to suggest that other states are looking to California’s exam as a model for the nation is overzealous. If this were the case, then one conclusion would be that only Louisiana and South Carolina are sufficiently progressive to follow California’s lead. Because there have been advances in both testing and technology since the current structure was developed, it is more likely that as most states have moved to more efficient measurement models to inform their licensure process, this is an instance where California is in a position to progress to where other programs have already gone.