The Testing Column

Testing Basics: What You Cannot Afford Not to Know

by Joanne E. Kane, Ph.D., and Andrew A. Mroch, Ph.D.

Editor's Note: This article is based on a presentation made by the authors, "Testing Basics for Bar Examiners: What You Cannot Afford Not to Know," at the 2017 NCBE Annual Bar Admissions Conference held on May 4–7, 2017, in San Diego, California. Measurement terms that the authors have identified as important testing concepts are bolded at their first appearance in this article and where they are defined; these terms and their definitions are also collected in the sidebar on page 36.

Why Do We Test Examinees at All?

Simply put, the purpose of licensure is to protect the public by identifying individuals who are not adequately prepared for entry-level practice. Licensure exams are used to determine whether candidates have attained a minimum threshold of essential knowledge, skills, and abilities for entry-level practice. The bar exam cannot realistically assess *all* the knowledge, skills, and abilities needed for entry-level legal practice. Passing the bar exam is intended to ensure that candidates can meet a minimum threshold of necessary knowledge and skills for entry-level legal practice but is not sufficient to ensure that a candidate is adequately prepared or that the candidate will be a competent lawyer in all situations.

It is important to keep in mind that this view of the purpose of licensure is not limited to legal licensure. The purpose of licensing examinations for doctors or accountants, for example, is similarly to protect the public. The website for the United States Medical Licensing Examination (USMLE) states that the series of examinations "assesses a physician's ability to apply knowledge, concepts, and principles, and to demonstrate fundamental patient-centered skills, that are important in health and disease and that constitute the basis of safe and effective patient care."[1] According to the American Institute of CPAs, "[t]he Uniform CPA Examination protects the public interest by helping to ensure that only qualified individuals become licensed as U.S. Certified Public Accountants (CPAs)."[2]

Why Do We Test in the Way That We Do?

The overarching goals in designing the bar exam are that the exam adequately represents the content domain of knowledge, skills, and abilities necessary for entry-level practice (a **validity** issue), that it is accurate enough and consistent enough to be fair (a **reliability** issue), and that it is reasonable in terms of time and resource requirements (a practicality issue). The variety of exam components used by most jurisdictions—multiple-choice questions, essay questions, and performance tests, each of which has advantages and limitations—are combined to balance validity, reliability, and practicality.

Test Content

We match the content of our examinations to the job requirements for entry-level lawyers. Prior to 2011, subject matter experts, including practitioners and academicians, provided guidance in developing test blueprints and subject matter outlines. In 2011 and 2012, Applied Measurement Professionals Inc. conducted a job analysis at the request of NCBE to determine what new lawyers do, and what knowledge, skills, and abilities newly licensed lawyers believe that they need to carry out their work. This job analysis has subsequently been used in concert with the opinions of subject matter experts to shape the test blueprints and subject matter outlines.[3]

The key guiding principle in establishing the connection or "bridge" between the content of the examination and actual entry-level practice in the profession is validity. **Validity** is the extent to which a test measures what it purports to measure, and the degree to which evidence and theory support the interpretations of test scores for particular uses.[4]

The reason we believe that those involved in bar admissions cannot afford not to know about the concept of test validity is that many attacks on the bar examination have validity issues at their core. For example, critics have claimed that the bar exam "does nothing to measure lawyering skills."[5] This critique is calling the validity of the interpretation of bar exam scores into question.

Beyond the NCBE Job Analysis, other research supports the validity of the interpretation of scores from the bar exam. For example, studies indicate that bar examination scores and law school grade point averages (GPAs) are positively correlated.[6] Scores on the Multistate Bar Examination (MBE) and the Law School Admission Test (LSAT) are also positively correlated, but interestingly, past research revealed that this is true only among students who have completed law school; among students taking the MBE within the first month of their law school training, LSAT scores do not correlate with MBE scores, refuting the suggestion that MBE scores are driven mainly by chance or "test-wiseness," and supporting the inference that MBE scores are a measure of legal knowledge.[7] Given that at least one major purpose of law school should be to prepare students for the entry-level practice of law, we would expect that a test designed to measure preparation for the entry-level practice of law should have some correspondence with both having attended law school and the grades received in law school. Our research reveals that it does.

Test Security and Standardization

The validity of the interpretation and use of bar exam scores can be threatened in a variety of ways. One major threat is cheating. If exam materials are stolen and distributed, or if examinees collude with one another, test scores can no longer be taken as reflective of given examinees' knowledge, skills, or abilities. Bar admission administrators and others who administer the bar examination are an absolutely integral part of maintaining and ensuring test security and, by extension, the validity of bar exam scores. NCBE and all bar examining professionals are trained and educated on the secure monitoring of test materials and testing environments to help ensure that the scores for individuals and groups of examinees are valid.

Another threat to the validity of the interpretation and use of bar exam scores is a lack of standardization in testing materials and conditions. If testing conditions are not the same across testing centers and individual examinees, bar exam scores might not be directly comparable. For example, if a test administrator invited a subset of examinees to discuss answers with one another during the test administration, the scores from that subgroup would not have the same meaning or be equally valid for the purpose of determining which individual examinees were adequately prepared for entry-level practice. NCBE relies on administrators to follow their jurisdictions' and NCBE's procedures to help ensure uniformity in testing conditions.

There are other threats to the validity of the interpretation and use of bar exam scores. NCBE employs careful scoring and scaling practices and conducts research behind the scenes to help ensure that the scores may be meaningfully applied to the licensure decision. We will describe some of these scoring considerations and mechanics in the next section.

Mechanics and Scoring

To ensure public protection and to give examinees a fair opportunity regardless of when they sit for the examination, bar examination scores must have equivalent meaning over time. When an examinee tests (February or July, one year or another) should have no impact on that examinee's score, because score interpretation should remain stable over time. NCBE achieves this stability in two important ways: first by achieving high reliability on the examinations, and second through **equating** the MBE and then **scaling** the written components of the bar examination to the MBE to capitalize on the MBE's equating.

In a testing context, **reliability** is a measure of how likely it is that a group of examinees would be rank-ordered in the same way over multiple (theoretical) testing sessions. In general, longer exams are more reliable than shorter exams. There is diminishing utility to adding exam questions beyond a certain point, but in general, longer tests are preferred to shorter ones from the perspectives of both reliability (i.e., score stability) and validity (i.e., content coverage). There is thus a reason that the bar examination is as long as it is; both high reliability and adequate content coverage are crucial. NCBE's former Director of Testing, Susan Case, pointed out that although there might be room on a short examination for examinees to complain that a particular topic they had studied was not included, or one they had failed to study was overrepresented, no examinee could reasonably complain that we happened to pick the 200 MBE scenarios that they had happened to fail to study.[8]

Equating has been given a thorough treatment in the *Bar Examiner* relatively recently.[9] Why does NCBE (like all other high-stakes testing organizations) equate its examinations? Essentially, **equating** makes adjustments to examinees' scores to help ensure that no examinee is unfairly penalized or unfairly rewarded—to even a very small degree—by taking a form of the test that is easier or more difficult than another year's form of the test. Equating is a central part of every large-scale testing operation, including familiar examinations like the LSAT, the GRE, the SAT, and the ACT.

Scaling is similar to equating. Both processes help ensure that examinees are not unfairly penalized or advantaged due to fluctuations in test difficulty over time. Both processes are also technical enough as topics to be difficult to understand at first blush—but, in a nutshell, the process of **scaling** used with written bar exam scores is a linear transformation of raw essay scores onto the MBE scale, which preserves the rank-ordering of examinees, allows for anchoring to an equated examination, and further allows easy combination of examination elements.[10]

What Is NCBE Doing Behind the Scenes?

Staff members in NCBE's Testing and Research Department engage in a variety of activities to support NCBE exams. However, there are several research-related activities that we engage in behind the scenes that may be helpful for those in the bar admissions community to understand.

Operational Analysis

One category of behind-the-scenes activities we do can be described as operational analysis, which consists of statistical and psychometric analysis to support NCBE exams. Equating, which is described above, is an important operational activity for the MBE. An activity analogous to equating but applied to written components of the bar exam is scaling, also described above. Equating and scaling are important behind-the-scenes activities that are regularly discussed in presentations that we give and have been addressed in this publication on many occasions. A third operational analysis we do that doesn't receive quite as much attention as equating and scaling but is nonetheless an important behind-the-scenes activity is **item analysis.**

Item analysis is the process of statistically analyzing each test question (also referred to as a test item). Typically, we use item analysis to ensure that each question is functioning statistically in a way that is consistent with what we would expect of a high-quality test question. For example, for the MBE items we look at the percentage of examinees answering the question correctly as an indicator of the difficulty of the test question. In addition, we look at how scores on a question differentiate those with low and high scores on the entire exam and whether those who tend to correctly answer a question also tend to score higher on the exam (or vice versa). We also study how each of the incorrect options (also referred to as distractors) perform compared to the correct option for a multiple-choice question.

Item analysis is conducted for all MBE questions but is particularly important for MBE pretest questions, which are unscored questions being tried out for inclusion on future MBE forms. Questions that appear on an MBE test form have been through an extensive review and revision process by committees of experts,[11] but we also secure statistical evidence of each question's performance using item analysis. For a question to be included on an MBE and to contribute to an examinee's score, in addition to being extensively vetted from a content perspective, it is vetted from a statistical perspective using item analysis.

Research for Jurisdictions

A second category of activities that we do behind the scenes includes research for jurisdictions. Often this research is in response to an inquiry or a special request from a board of bar examiners to conduct analysis on a particular topic for the jurisdiction. Most of these requests involve consultation with NCBE staff to determine what the jurisdiction needs and, possibly, what data are available to address the needs. Examples of research for jurisdictions include analysis of possible effects of testing irregularities on examinee performance, bar exam performance trends across exam administrations, implications of changing a jurisdiction's cut score, and implications of making other changes to the bar exam. In addition, we sometimes receive requests that do not involve research in the form of data analysis but involve questions about topics like test scores, graders, grading, or testing concepts. The research that we do for jurisdictions can range from a relatively quick response based on previous work we have conducted to an ongoing project for a jurisdiction across multiple bar exam administrations.

Exploratory Studies

A third category of activities we do involves conducting exploratory studies. Continuous improvement is a theme for these types of studies. Consistency over time is important for fairness and score comparability.

That said, we also look for ways to improve what we test, how we test, and the processes and procedures we use with our exams. Exploratory studies help us evaluate available options (or even develop new options) before making any changes to our exams. Examples of exploratory studies could include topics like online testing, automated scoring of written responses, standard setting, and automated test assembly.

We also occasionally research topics of particular interest to NCBE or jurisdictions, often to inform policy. For example, to inform discussions of the essay regrading policies within jurisdictions that have adopted the Uniform Bar Examination, NCBE staff reviewed the regrading policies in UBE jurisdictions that conduct regrading and researched the effects of regrading on examinee performance and bar passage. This research contributed to a *Bar Examiner* article on regrading[12] and informed jurisdiction policies on regrading.

Broader Psychometric Research

Finally, we also conduct basic research that contributes to the field of educational measurement and psychometrics. Often we present this research at conferences related to assessment and measurement. Sometimes this type of research helps us explore new or alternative psychometric methods for our testing programs, so it may overlap with exploratory studies we conduct.

Are We Alone in the Licensing Universe?

No. There are many other licensing authorities for a variety of professions, and the bar exam is one of many licensure exams that exist. Other professions that have a licensure exam include medicine, accounting, architecture, commercial aviation, teaching, nursing, and pharmacology. Similar to the legal profession, many professions issue a general

Measurement Term Definitions

Validity is the extent to which a test measures what it purports to measure, and the degree to which evidence and theory support the interpretations of test scores for particular uses.

Reliability is the degree to which scores for a group of examinees would be consistent over multiple (theoretical) testing sessions.

Equating is the process of making adjustments to examinees' scores to help ensure that no examinee is unfairly penalized or unfairly rewarded—to even a very small degree—by taking a form of the test that is easier or more difficult than another year's form of the test.

Scaling is the process of placing one exam's scores onto the scale of another exam. When scaling the written component of the bar exam, raw essay scores are transformed linearly onto the MBE scale, which preserves the rank-ordering of examinees but allows for anchoring to an equated examination.

Item analysis is the process of statistically analyzing each test question (also referred to as a test item) to ensure that each question is functioning statistically in a way that is consistent with what is expected of a high-quality test question.

license and require a multiple-choice exam as part of their licensure exam. It is common for the exam to assess the breadth of practice covered by the license.

Fundamental testing concepts like reliability and validity are important for any exam, including other licensing exams. We are in good company!

Where Can Those Involved in Bar Admissions Go for More Information or Help?

The *Bar Examiner* is a rich source of information covering a range of topics of importance to the bar admissions community. Many questions and concerns pertinent to bar exams and bar examining have been addressed in prior issues of the magazine over the years, and the archived issues are available on NCBE's website at www.ncbex.org/publications/the-bar-examiner. NCBE's website in general is also an excellent source of information for candidates, law schools, and others with questions about the bar exam.

One of the wonderful aspects of the bar admissions community is that it is a community. Peer jurisdictions are often a helpful source of information and assistance, as any given problem or question in one jurisdiction has likely been dealt with in another jurisdiction at some point. And for those tricky or unusual situations, NCBE staff members are available to assist.

Notes

1. United States Medical Licensing Examination, http://www.usmle.org (retrieved May 25, 2017).
2. American Institute of CPAs, http://www.aicpa.org/BecomeACPA/CPAExam/Pages/default.aspx (retrieved May 25, 2017).
3. See Susan M. Case, Ph.D., "The Testing Column: The NCBE Job Analysis: A Study of the Newly Licensed Lawyer," 82(1) *The Bar Examiner* (March 2013) 52–56; Susan M. Case, Ph.D., "Summary of the National Conference of Bar Examiners Job Analysis Survey Results," January 2013, *available at* http://www.ncbex.org/pdfviewer/?file=%2Fdmsdocument%2F55.
4. For an expansion of this topic, *see* Michael T. Kane, Brian E. Clauser, and Joanne Kane, "A Validation Framework for Credentialing Tests," in *Testing in the Professions: Credentialing Policies and Practice* (Eds. Susan Davis-Becker and Chad W. Buckendahl, National Council on Measurement in Education 2017).
5. Kristin Booth Glen, law professor and former dean of the City University of New York School of Law, quoted in Elizabeth Olson, "Bar Exam, the Standard to Become a Lawyer, Comes Under Fire," *New York Times,* March 19, 2015, *available at* https://www.nytimes.com/2015/03/20/business/dealbook/bar-exam-the-standard-to-become-a-lawyer-comes-under-fire.html?hp&action=click&pgtype=Homepage&module=mini-moth&region=top-stories-below&WT.nav=top-stories-below&r=1&mtrref=blogs.findlaw.com.
6. *See, e.g.,* Susan M. Case, Ph.D., "The Testing Column: Identifying and Helping At-Risk Students," 80(4) *The Bar Examiner* (December 2011) 30–32, at 31; Michael T. Kane, Andrew A. Mroch, Douglas R. Ripkey, and Susan M. Case, *Impact of the Increase in the Passing Score on the New York Bar Examination* (National Conference of Bar Examiners 2006) 124–126, *available at* https://www.nybarexam.org/press/ncberep.pdf; and Stephen P. Klein, *Summary of Research on the Multistate Bar Examination* (National Conference of Bar Examiners 1993) 25–27.
7. Klein, *supra* note 6, at 27–28.
8. Susan M. Case, Ph.D., "The Testing Column: What Everyone Needs to Know about Testing, Whether They Like It or Not," 81(2) *The Bar Examiner* (June 2012) 29–31.
9. Mark A. Albanese, Ph.D., "The Testing Column: Equating the MBE," 84(3) *The Bar Examiner* (September 2015) 29–36.
10. For more on scaling, *see, e.g.,* Mark A. Albanese, Ph.D., "The Testing Column: Scaling: It's Not Just for Fish or Mountains," 83(4) *The Bar Examiner* (December 2014) 50–56; Susan M. Case, Ph.D., "The Testing Column: Frequently Asked Questions About Scaling Written Test Scores to the MBE," 75(4) *The Bar Examiner* (November 2006) 42–44; and Judith A. Gundersen, "It's All Relative—MEE and MPT Grading, That Is," 85(2) *The Bar Examiner* (June 2016) 37–45.
11. For details, *see* C. Beth Hill, "MBE Test Development: How Questions Are Written, Reviewed, and Selected for Test Administrations," 84(3) *The Bar Examiner* (September 2015) 23–28.
12. Mark A. Albanese, Ph.D., "The Testing Column: Regrading Essays and MPTs—and Other Things That Go Bump in the Night," 85(1) *The Bar Examiner* (March 2016) 58–61.

Joanne E. Kane, Ph.D., is the Associate Director of Testing for the National Conference of Bar Examiners.

Andrew A. Mroch, Ph.D., is a Research Psychometrician for the National Conference of Bar Examiners.