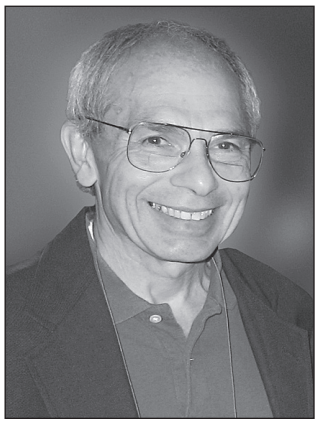
The Testing Column

Standards? We Don't Need No Stinking Standards!

by Mark A. Albanese, Ph.D.



Imagine . . . a world of people who all have outstanding innate academic and real-world talents and who all receive the best education possible, and where opportunity and jobs are plentiful in all occupations. Anyone who wants to go to law school can go to law school and will graduate with honors. There is no need for licensing tests, because all law school graduates are so far above average that they can't even see average, it is so far below them; and all law school graduates go on to become Supreme Court justices . . . .

The reality is that not everyone has the innate academic and real-world talents to succeed in a profession like law, even if they have a burning desire to be a lawyer. Given that the numbers of students applying to law school are at lows not seen since the 1970s, and that there are now far more law schools than there were back then, academic standards for admission and graduation at some law schools do not signify what they may have in the past. Although bar passage has been generally declining for over a decade, it started to cascade downward in July 2014. In response to the declining passing rate, law school deans in many jurisdictions have called for jurisdictions to consider re-evaluating their passing standards; and two large jurisdictions, Texas and California, have formally begun to do so. Clearly, incompetent lawyers can have disastrous consequences for their clients. In this environment, the bar admissions process is the last line of defense against incompetent lawyers being allowed to practice.

But that is not the only perspective. An opposing view comes from graduating law students who have put in years of study and have often paid massive amounts of money for obtaining their law degrees. If they cannot practice law, their dreams of becoming a lawyer will be crushed, many will have no way to make a living sufficient to repay their educational debt, and the debt load may be a lifelong millstone. The only thing that keeps them from attaining their dream and avoiding this scenario is the bar admissions process.

So, a lot is riding on the process of admission to the bar, and what makes or breaks that process is the standard set for passage of the bar examination—the minimum score that determines passage. (There is, of course, also a character and fitness component that must be satisfied, in most cases before the bar examination.) Depending on the jurisdiction, the bar exam can take two to three days, beginning with essays and performance tests and, in some cases, jurisdiction-specific multiple-choice tests, and ending with the Multistate Bar Examination (MBE) on the last Wednesday of the month (February or July).[1] The jurisdiction generates a score from the written component answers and combines the score in some form with the scaled score generated from the MBE to produce the score used to determine the pass/fail result on the bar examination. On the MBE scale,

where scores can range from 0 to 200, passing scaled scores range from 129 to 145 across jurisdictions.

With the continuing decline in performance on the bar examination for the past few years, many jurisdictions have questioned whether their passing standards are set at the appropriate level. The purpose of this article is to provide information on several approaches that have been used to set standards on high-stakes licensing examinations and to highlight some of the challenges that exist in arriving at standards that ensure the protection of the public and are fair to law school graduates.

Standard-Setting Methods

The methods of setting standards all employ judgments on the part of a group of knowledgeable experts. The selection of these individuals is crucial to the credibility of the standards that result. The experts should be respected, and anyone questioning the credibility of the standard-setting process should, upon viewing the credentials of the experts, conclude that the experts are appropriate for the assignment. Depending upon the standard-setting approach used, the standard-setting panel may be required to make judgments about exam content, examinees, or task requirements. Those on the panel should also be free from any conflicts of interest, such as setting standards for students they have taught.

The approaches that have been used can be grouped into three types: arbitrary standard setting, test-centered methods, and examinee-centered methods.[2]

Arbitrary Standard Setting

The arbitrary method of standard setting is probably the most common approach and involves setting standards without systematically examining either the task requirements or samples of examinee performance. Susan M. Case, Ph.D., cites the BOGSAT approach (the acronym standing for Bunch of Guys Sitting at a Table).[3] The problem with arbitrary standards is that they are not easily defended, which might make them acceptable for low-stakes decisions like grades in a single course, but makes them ill-suited for making high-stakes decisions like bar passage. If no one currently involved with bar admissions in a jurisdiction has any idea of how that jurisdiction's standard was set, it should be reevaluated through a process other than the arbitrary method.

Test-Centered Methods

Psychometricians have been developing methods of setting defensible standards since at least the middle of the 20th century. The early methods were primarily test-centered, and it could be said that the operating principle undergirding these methods is that a standard would be more defensible than an arbitrary standard if the experts actually examined the test in terms of expectations of how the minimally competent examinee would perform.

Nedelsky Method

The earliest method in the psychometric literature is the Nedelsky method, named after its originator, Leo Nedelsky. It derived a standard, also known as a minimum pass level or MPL, for multiple-choice items by having the experts determine which of the incorrect answers the minimally competent examinee would be able to eliminate for any given multiple-choice item, assuming that the examinee would then guess among the remaining answers. For instance, if an item had four answer options and experts estimated that a minimally competent examinee could eliminate two of them, the MPL would be 1/2 (or 50%) since the examinee would be guessing between the correct answer and the remaining incorrect answer. The Nedelsky method was obviously limited to multiple-choice items.

Angoff Method

The Angoff method was subsequently developed by William Angoff with an eye toward broader applications. It or one of its variants has been applicable to almost any type of item, score, or task. This versatility is one of the factors that probably has motivated the Angoff method's use in many professional settings, such as medical licensure.[4] The Angoff approach has the experts first define characteristics of the minimally competent examinee. With a clear picture of what this minimally competent individual can do imprinted on their minds, the experts review the task at hand and give their estimate of the likelihood of this minimally competent individual being successful on the task. Because experts sometimes find it difficult to give an estimate of the likelihood of an individual being successful on a task, it is often reframed as how many of a group of 100 minimally competent examinees the expert would expect to be successful on this task.

M. Friedman Ben-David adapted the Angoff approach to use with a performance exam where examinees rotate between different stations and where, at each, they must demonstrate a specific skill or set of skills, and are often graded by a point system as they demonstrate aspects of the skill(s).[5] The adaptation of the Angoff method was made by having the experts determine the number of scoring points an individual borderline candidate would receive in order to pass the station. If an essay question (or performance test) is substituted for the skill(s) to be demonstrated at the station, the approach has direct applicability to the bar examination.

Examinee-Centered Methods

There are a number of methods that focus on the performance of examinees. The two most commonly encountered approaches are the Contrasting Groups method and the Bookmark method (although the Bookmark method has forms that are strictly test-centered). A third method with different application is the Hofstee method.

Contrasting Groups Method

There are variations of the Contrasting Groups method, but Michael T. Kane, Ph.D., describes this method in general as having experts determine if examinees have the knowledge, skills, and judgment needed to practice, based upon a sample of their performance; categorizing examinees into two groups (i.e., those who have met the requirements and those who have not); and then selecting a passing score that differentiates between the two groups as well as possible.[6]

Bookmark Method

The Bookmark method involves making judgments about either the tasks or the performance of examinees on a task, but it requires actual data. The easiest case for illustration purposes is for multiple-choice test performance. Items would be ordered from easiest to hardest (based upon actual data), and the experts would start with the easiest item and stop when they reach the point where they think a minimally competent examinee would have a specified probability of answering the item correctly (e.g., 50%). The difficulty of the item at this point would become the standard.

An alternative for performance-based assessments would be to order a sample of performances (e.g., essays) according to the grade awarded from lowest to highest. Experts would start at the lowest-graded performance and work their way up through the higher-graded performances until reaching the point where they find the first performance that a minimally competent examinee would have the specified probability of reaching. The grade of that performance would then be the passing standard.

In practice, the ordering of items/performances is not perfect, and one needs to have experts go up for a few more items/performances to be certain that some of the higher-graded ones were consistent with the stopping point and that where they stopped initially was not at an item/performance out of order. The challenge with the Bookmark method is to get the ordering of the items/performances before the standard-setting session, since it generally requires actual data to make the ordering.

Hofstee Method

The Hofstee method of standard setting does not make assessments of performance at the individual item level but requires experts to give their impressions of what the minimum and maximum failure rates should be for the exam, as well as what the minimum and maximum percent correct scores should be. These minimum and maximum failure rates and percent correct scores are averaged across experts and projected onto the actual score distribution to derive a passing score. Because it operates at the overall test level, it can be combined with other standard-setting methods as a cross-check. In fact, having experts go through the standard-setting process with, say, the Angoff method can be a good training approach for experts before they apply the Hofstee method.

Challenges

Generally, the methods used to derive standards seem relatively straightforward conceptually. However, the devil is in the details. Ronald K. Hambleton, Ph.D., provides the following 11 steps for setting performance standards on educational assessments, which can be applied to any of the standard-setting methods discussed.

1. Choose a panel (large, and representative of the stakeholders).
2. Choose one of the standard-setting methods, and prepare training materials and finalize the meeting agenda.
3. Prepare descriptions of the performance categories (e.g., basic, proficient, and advanced [or, in the case of the bar exam, fail and pass]).
4. Train panelists to use the method (including practice in providing ratings).
5. Compile item ratings and/or other rating data from the panelists (e.g., panelists specify expected performance of examinees at the borderlines of the performance categories).
6. Conduct a panel discussion; consider actual performance data (e.g., item difficulty values, item characteristic curves, item discrimination values, distractor analysis) and descriptive statistics of the panelists' ratings. Provide feedback on interpanelist and intrapanelist consistency.
7. Compile item ratings a second time that could be followed by more discussion, feedback, and so on.
8. Compile panelist ratings and obtain the performance standards.
9. Present consequences data to the panel (e.g., passing rate).
10. Revise, if necessary, and finalize the performance standards, and conduct a panelist evaluation of the process itself and their level of confidence in the resulting standards.
11. Compile validity evidence and technical documentation.[7]

Each of these 11 steps also seems relatively straightforward. However, as I said earlier, the devil is in the details. Taking the first step, "Choose a panel (large, and representative of the stakeholders)," here are some of the devilish details to think about. Who are the stakeholders? Are there some stakeholders who must actually be on the panel as opposed to

simply being represented? How many experts are needed? (There is large and then there is LARGE.) It has been recommended in the psychometric literature that 15 to 20 experts be used in setting the standard for a high-stakes examination like the bar examination. Choosing and recruiting experts is no small challenge. Representing a stakeholder group is not enough. What background and experiences does an expert need to be credible with not only the stakeholder group but the public at large? Serving on a standard-setting panel is a major time commitment. Are the experts willing to serve, and are they available to do so when needed? Setting standards is a very important task, and it requires careful thought at each step. Enlisting assistance from someone who has experience with the standard-setting process will help avoid major problems.

So, you might have noticed that the title of this article is really a double negative that translates into "Standards? We do need standards!" Whether they stink or not will depend upon how they are set.

References

Academy of Medical Royal Colleges, Guidance for Standard Setting: A Framework for High Stakes Postgraduate Competency-Based Examinations (October 2015), *available at* https://www.aomrc.org.uk/wp-content/uploads/2016/05/Standard_setting_framework_postgrad_exams_1015.pdf (accessed May 2, 2017).

John J. Bowers, Ph.D., and Russelyn Roby Shindoll, "A Comparison of the Angoff, Beuk, and Hofstee Methods for Setting a Passing Score," ACT Research Report Series No. 89-2, May 1989, *available at* https://forms.act.org/research/researchers/reports/pdf/ACT_RR89-2.pdf (accessed May 2, 2017).

S.M. Downing, A. Tekian, and R. Yudkowsky, "Procedures for Establishing Defensible Absolute Passing Scores on Performance Examinations in Health Professions Education," 18(1) *Teaching and Learning in Medicine* (Winter 2006) 50–57.

S.L. Fowell, R. Fewtrell, and P.J. McLaughlin, "Estimating the Minimum Number of Judges Required for Test-Centred Standard Setting on Written Assessments. Do Discussion and Iteration Have an Influence?" 13(1) *Advances in Health Sciences Education: Theory and Practice* (March 2008) 11–24.

W.K.B. Hofstee, "The Case for Compromise in Educational Selection and Grading," in *On Educational Testing* 109–127 (S.B. Anderson and J.S. Helmick eds., Jossey-Bass 1983).

National Board of Osteopathic Medical Examiners, Standard Setting, The Approach to Standard Setting, http://www.nbome.org/standardsetting_app.asp.

M.R. Raymond and J.B. Reid, "Who Made Thee a Judge? Selecting and Training Participants for Standards on Complex Performance Assessments," in *Setting Performance Standards: Concepts, Methods, and Perspectives* 119–157 (G.J. Cizek ed., Lawrence Erlbaum Associates 2001).

Notes

1. For the July 2017 and February 2018 bar examinations, Massachusetts will continue to administer its 10 essay questions on the Thursday following the administration of the Multistate Bar Examination. Effective with the July 2018 bar examination, Massachusetts will administer the Uniform Bar Examination, which ends with the MBE on Wednesday.
2. Michael T. Kane, Ph.D., "Standard Setting for Licensure Examinations," 70(4) *The Bar Examiner* (November 2001) 6–9.
3. Susan M. Case, Ph.D., "The Testing Column: Sometimes BOGSAT Is Just Not Good Enough," 81(3) *The Bar Examiner* (September 2012) 31–33.
4. R.J. Nungester, G.F. Dillon, D.B. Swanson, N.A. Orr, and R.D. Powell, "Standard-Setting Plans for the NBME Comprehensive Part I and Part II Examinations," 66(8) *Academic Medicine* (August 1991) 429–33.
5. M. Friedman Ben-David, "AMEE Guide No. 18: Standard Setting in Student Assessment," 22(2) *Medical Teacher* (2000) 120–130.
6. Kane, s*upra* note 2.
7. Ronald K. Hambleton, "Setting Performance Standards on Educational Assessments and Criteria for Evaluating the Process," Laboratory of Psychometric and Evaluative Research Report No. 377, School of Education, University of Massachusetts, Amherst, MA, *available at* http://www.nciea.org/publications/SetStandards_Hambleton99.pdf (accessed May 2, 2017).

Mark A. Albanese, Ph.D., is the Director of Testing and Research for the National Conference of Bar Examiners.