RAISING THE BAR: A SOCIAL SCIENCE CRITIQUE OF RECENT INCREASES TO PASSING SCORES ON THE BAR EXAM

Deborah J. Merritt[*]
Lowell L. Hargens[**]
Barbara F. Reskin[***]

At least a dozen states have raised the score required to pass their bar exams during the last decade, with several more evaluating proposed increases.[1] Partly as a result of these changes, the percentage of test takers passing the bar has dropped sharply since 1994. In that year, 74% of examinees nationwide passed the bar.[2] In 1995, the pass rate dropped to 70%, while in 1998 it fell to 66%.[3] Declines in some states have been even more precipitous; in Ohio, the passing rate fell from 85% in 1994 to 69% in 1998.[4]

Passing rates fluctuate partly due to applicant quality; if exam takers during the late 1990s were less qualified than those taking the exam in earlier years, the recent decline in passing rates was appropriate. Statistics released by the National Conference of Bar Examiners, however, suggest that applicant quality was *higher* in the 1990s than it was during the 1980s.[5] Today's passing scores are excluding prospective lawyers who would have passed the same bar exam a decade ago.

---

* Director, The John Glenn Institute for Public Service and Public Policy, The Ohio State University; John Deaver Drinko/Baker & Hostetler Chair in Law, The Ohio State University. B.A., Harvard University; J.D., Columbia University. We thank James Brudney, Michael Masinter, Andrew Merritt, and Allan Samansky for their very helpful comments on an earlier draft.

** Professor of Sociology, The Ohio State University. B.A., University of Minnesota; MA., University of Wisconsin; Ph.D., University of Wisconsin.

*** Professor of Sociology, Harvard University. B.A., University of Washington; MA., University of Washington; Ph.D., University of Washington.

1. States that have raised their passing score include Arizona, Georgia, Illinois, Kansas, Maine, Missouri, Nebraska, North Carolina, Ohio, Oklahoma, Pennsylvania, Rhode Island, Texas, and Wisconsin. Other states have implemented more complex changes that may have had the effect of making bar passage more difficult. Only one state, New Mexico, plainly lowered its passing score during the 1990s. Two others, Mississippi and New Jersey, both raised and lowered their passing scores with the net effect unclear. Florida and Minnesota currently are considering proposals to raise their passing scores, while Pennsylvania (which raised its score earlier in the decade) is considering a possible decline. *See infra* notes 34, 60-69 and accompanying text.

2. *See 1994 Statistics*, B. EXAMINER, May 1995, at 7, 10.

3. *See 1995 Statistics*, B. EXAMINER, May 1996, at 23, 26; *1998 Statistics*, B. EXAMINER, May 1999, at 6, 8. In 1996 and 1997, the nationwide passing rate held steady at 70%. *See 1996 Statistics*, B. EXAMINER, May 1997, at 15, 17; *1997 Statistics*, B. EXAMINER, May 1998, at 17, 19.

4. *See 1994 Statistics*, *supra* note 2, at 7, 10; *1998 Statistics*, *supra* note 3, at 6, 8.

5. *See infra* note 23 and accompanying text.

The decline in bar passage rates has serious implications for both individual applicants and the public. For individuals, bar failure may bring unemployment, a deep sense of professional failure, and financial insecurity. The risk is especially high for graduates with steep educational debts. Even if the graduate passes the bar on a second or third try, any delay in bar admission hinders repayment of those debts.

For the public, increases in the bar passing score may be anti-competitive. Bar exams are intended to keep unqualified individuals from practicing law, but they also reduce the number of attorneys serving the public and raise the price of legal services. Recent rises in bar passing scores followed both a legal recession during the early 1990s and the admission of a record number of new attorneys in 1994.[6] Although the bar examiners and state supreme court justices who set passing standards undoubtedly act in good faith, this anti-competitive context makes it essential to scrutinize recent rises in the passing score carefully.

Increased passing scores may also threaten the diversity of the legal profession. Although law school graduates today are more demographically diverse than at any time in our nation's history, minority test takers fail the bar exam at a higher rate than do white examinees.[7] Under these circumstances, raising the bar's passing score—especially without sound evidence that former standards failed to weed out incompetent practitioners—undermines the profession's goal of increasing diversity. The implications are particularly troubling when the hurdle set for today's demographically diverse graduates is higher than the one set for less diverse examinees ten or twenty years ago.

Ironically, inflated bar standards can also diminish the quality of new lawyers. As bar exams become more difficult to pass, students devote more time to memorizing the rules tested on the bar. Less time is available for subjects—like alternative dispute resolution—that are

---

6. In 1987, the states admitted 43,481 attorneys by examination. *See 1990 Statistics*, B. EXAMINER, May 1991, at 13, 22. That figure climbed to 45,420 in 1990, *see id.*, and 51,139 in 1992, *see 1994 Statistics*, *supra* note 2, at 7, 16. In 1994, the number of attorneys admitted by examination peaked at 53,039. *See id. But see 1998 Statistics*, *supra* note 3, at 6, 15 (showing the 1994 number to be 52,962). Since then, the figure has fallen to 49,168 in 1998. *See 1998 Statistics*, *supra* note 3, at 6, 15.

Attorneys admitted by examination include some experienced attorneys seeking bar admission in a new state. Likewise, some new attorneys take more than one bar exam at the start of their career. The number of lawyers admitted to the bar by examination each year, therefore, is not the same as the number of new attorneys licensed each year. Even with this caveat, 1994 appears to have represented a dramatic peak in newly licensed lawyers joining the profession.

7. *See, e.g.*, Stephen P. Klein & Roger Bolus, *The Size and Source of Differences in Bar Exam Passing Rates Among Racial and Ethnic Groups*, B. EXAMINER, Nov. 1997, at 8; Linda Wightman, *Summary of the Report*, LSAC NATIONAL LONGITUDINAL BAR PASSAGE STUDY 8, 9 (1998).

highly responsive to public needs but not represented on the bar exam. By forcing students to attain high levels of knowledge in a few areas, heightened bar passage standards may distract new lawyers from developing essential skills.

Given these costs, how have states justified increases in their bar passing scores? Courts and bar examiners in some states have offered no rationale at all. The Illinois Supreme Court, for example, raised the passing score on its bar exam without inviting public comment or revealing the method by which it had chosen the new passing score.[8] Several other states have pursued more complex standard-setting procedures to choose their new passing scores. These procedures, developed by psychologist Stephen Klein, attempt to introduce scientific rigor into the standard-setting process.[9] Regrettably, however, the Klein method suffers from a fundamental flaw that produces an arbitrary passing score. Indeed, the process may mislead bar examiners, supreme court justices, and the public by falsely suggesting that the state has adopted a scientifically defensible passing score.

In this Article, we critique both the general movement toward higher passing scores and the specific approach used by states adopting Klein's method. The first section of the Article briefly reviews the structure and scoring of bar examinations. In the second section, we examine general claims that previous passing scores were too low. The third section explains and critiques in some depth Klein's method of setting passing scores to which some states are turning. In a final section of the Article, we examine the impact of raised passing scores on the admission of minority attorneys, an issue of special concern to many practitioners and members of the public.

The process used to set bar exam passing scores has elicited little previous attention from legal academics. But consideration of this issue, particularly the Klein process, is crucial for three reasons. First, for the

---

8. *See, e.g.*, Chris Klein, *Illinois Deans' Dilemma—Is the Bar Exam (A) Sentry or (B) Aptitude Test?*, NAT'L L.J., Dec: 30, 1996, at A16; Jackson Williams, *Irreconcilable Principles: Law, Politics, and the Illinois Supreme Court*, 18 N. ILL. U. L. REV. 267, 318 (1998).

9. Klein's process was used in Ohio, Pennsylvania, Florida, and Minnesota; it is currently being used in New York. *See* Florida Board of Bar Examiners, Reply to Comments app. at 1, Case No. SC96869 (Fla. May 10, 2000) (reprinting letter from Stephen P. Klein to Thomas A. Pobjecky, General Counsel, Florida Board of Bar Examiners (May 8, 2000)) [hereinafter Reply to Comments]. Klein first used the process in Puerto Rico during the 1980s. *See id.* For further discussion of the process pursued in these states, and the results rising out of them, see *infra* notes 32-69 and accompanying text.

As citations throughout this article attest, Klein is a frequent consultant and recognized expert in the bar examination field. Much of his work related to bar examinations has been excellent. The method critiqued in this Article, however, suffers from the serious flaws we identify. Klein's status as consultant on bar exam issues should not lead bar examiners to overlook the flaws in this particular method, which Klein himself calls an "eclectic model." *See infra* note 32 and accompanying text.

reasons indicated above, bar exam passing scores have important policy implications for both the public and the profession. Second, an increasing number of states may be drawn to the apparently scientific nature of Klein's method without recognizing its flaws. Indeed, New York is using Klein's process now.[10] Finally, adoption of Klein's method in several states illustrates the way in which legal issues increasingly depend upon sophisticated social science techniques—and on how a failure to understand those techniques can produce flawed outcomes. Unless legal scholars engage social science methods in our own journals, we will be unable to identify and correct defective legal rules based on those techniques.

I. THE STRUCTURE AND SCORING OF BAR EXAMINATIONS

The bar exam in most states consists of two parts. The Multistate Bar Exam (MBE), designed by the National Conference of Bar Examiners (NCBE), is a multiple-choice test of 200 questions in six subjects. Forty-seven states and the District of Columbia require this exam.[11] The remainder of the bar, often dubbed the "state section," may include essays, performance items, and/or additional multiple-choice questions. These questions often, but not always, focus on state law.[12] For convenience, we refer to this second portion of the bar exam as the "state" or "essay" section of the exam.

The NCBE assures state bar examiners and examinees that a given score on the MBE reflects the same level of knowledge from year to year. To achieve this result, the National Conference includes both new and old (i.e., "repeat") questions on each exam with the latter known as "equators." The NCBE compares the average scores on the equator

---

10. *See* Reply to Comments, *supra* note 9, at 1.

11. *See* ABA SECTION OF LEGAL EDUCATION AND ADMISSIONS TO THE BAR & THE NATIONAL CONFERENCE OF BAR EXAMINERS, COMPREHENSIVE GUIDE TO BAR ADMISSION REQUIREMENTS 2000, Chart V (2000) (available on the web at <http://www.abanet.org/legaled/publications/Compguide2000/cgchart5.html>) [hereinafter 2000 COMPREHENSIVE GUIDE]. Indiana, Louisiana, and Washington are the only states that do not currently use the MBE. *See id.* (indicating Puerto Rico also does not use the MBE).

12. An increasing number of states now use the Multistate Essay Exam or Multistate Performance Test, both provided by the National Conference of Bar Examiners, as part or all of the "state" portion of their exams. Although these tests are drafted centrally, responses are graded by examiners in individual states. States may instruct examinees to answer these exam questions either by applying the distinctive law of their states or by applying general principles of law.

In addition to the traditional bar exam, all but three states and Puerto Rico require applicants to pass the Multistate Professional Responsibility Exam (MPRE). *See* 2000 COMPREHENSIVE GUIDE, *supra* note 11, at Chart VI (showing that only Maryland, Washington, Wisconsin, and Puerto Rico do not require applicants to pass the MPRE). States set a separate passing score for the exam, which is administered separately from the traditional bar exam. We do not discuss passing scores for the MPRE in this Article.

questions from each exam with past scores on those same questions. If the current test takers perform better, on average, than past examinees on the equator questions, the NCBE scales up the current examinees' total scores on the MBE; it adds the difference in average scores on equator questions to every examinee's MBE score. If current performances are worse, on average, than those of previous test takers, the NCBE scales down the current scores by subtracting the difference from every examinee's MBE score.[13]

As a result of this equating, MBE scores are comparable both over time and across states. An equated score of 131 on the MBE reflects the same level of performance today as it did ten or twenty years ago, and it signals the same level of competence in Maine as in California. All variation in average equated MBE scores across states and over time reflects variation in applicants' performance, not fluctuations in test difficulty or grading standards.[14]

Raw scores on the state sections of bar examinations, in contrast, are not comparable across states or over time. This is true for at least three reasons: Different states use different questions on their exams, states change their questions over exam administrations, and states may use different graders from year to year. As a result, although scores on the state portion of the bar examination may measure the relative performances of those who took the exam on a given occasion,[15] they cannot be used to compare performances across administrations of the exam.

When combining state-section scores with MBE scores, most states first calibrate the combined state (or "essay") scores to match the distribution of that state's MBE scores on the same exam. State bar examiners do this by transforming the combined state-portion scores for each test taker so that the combined scores' average, as well as their standard deviation (a measure of dispersion around the average), equal those for the MBE portion. If, for example, examinees who took the Ohio bar exam in July 1999 averaged 142 points on the MBE and had

---

13. For further discussion of equating, see Stephen Klein, *Options for Combining MBE and Essay Scores*, B. EXAMINER, Nov. 1995, at 38, 38-40, and Julia C. Lenel, *Issues in Equating and Combining MBE and Essay Scores*, B. EXAMINER, May 1992, at 6, 6-8.

14. As a result, changes in MBE scores achieved in one large state, California, "have tracked almost perfectly changes in [the] same applicants' mean LSAT scores." Klein, *supra* note 13, at 40; *see also* Stephen P. Klein, *On Testing: Establishing Pass/Fail Standards*, B. EXAMINER, Aug. 1986, at 18, 19 ("MBE equated scores . . . are not affected by differences between exams in the average difficulty of the questions asked and the leniency with which the answers to them are graded.").

15. We say "*may* measure relative performance" because there is a reasonable chance that they do not do so if raw scores on individual questions are not standardized before they are combined into a total score for the state portion of the exam. States vary in whether they standardize essay scores.

a standard deviation of 15 points on that portion of the exam, their scores on the second part of the exam would be transformed so that those scores also averaged 142 with a standard deviation of 15.[16]

Most states, finally, sum the MBE and transformed state scores to obtain a single combined score for each examinee.[17] Most also pass examinees who meet an announced passing score for the full exam; test takers in these states can compensate for poor performance on one part of the exam with a stronger performance on the other. A few states require examinees to meet separate passing scores for the MBE and state portions of the exam.[18] Others use hybrid rules, such as one requiring a minimum combined score plus a passing grade on a designated percentage of the state's essays.[19] Many states also have procedures for regrading exams that fall just short of the passing line.[20]

Whatever the state's passing rule, almost all scale their essays or other "state" items to MBE scores in the manner described above. This practice rests on the assumption that relative performances on the two portions of the exam are equivalent. Thus, it assumes that an average performance on the state portion of a bar exam demonstrates the same level of competence as an average performance among that state's MBE scores. Similarly, it assumes that performances that are one standard deviation above the mean on the state portion are equivalent to scores

---

16. *See* Klein, *supra* note 13, at 38-42; Lenel, *supra* note 13, at 8-14. Some states use a variation of this method called the "equipercentile method" to scale state scores to the MBE. Under that method, the examinee with the highest raw score on the state portion of the exam receives a scaled state score equivalent to the highest MBE score in the state, the examinee with the second highest raw score on the state portion receives a scaled score equal to the second highest MBE score, etc. Complex calculations are used to resolve ties. *See* Klein, *supra* note 13, at 38-39.

17. When combining scores from the two sections, some states weight the two portions differently. If a state follows that path, it will multiply the MBE and/or state scaled scores before combining them. Texas, for example, has four parts to its bar exam. It multiplies both the scaled MBE score and the scaled Texas essays score by two, then combines those scores with scaled scores for the Texas Procedure and Evidence portion of the exam and the Multistate Performance Test portion. *See* 2000 COMPREHENSIVE GUIDE, *supra* note 11, at Chart VII & notes. The MBE and Texas essays thus each contribute one-third to the examinee's final score, while the MPT and Procedure/Evidence questions each contribute one-sixth.

18. Rhode Island, for example, requires applicants to obtain a scaled score of 140 on the MBE and to pass seven out of twelve essay questions; South Carolina requires applicants to obtain a scaled score of 125 on the MBE and a score of 70 on the essays; Vermont requires a scaled MBE score of 135 and a score of 36 on the essays. *See* 2000 COMPREHENSIVE GUIDE, *supra* note 11, at Chart VII & Supplemental Remarks.

19. Delaware, for example, requires examinees to obtain a scaled score of 130 on the MBE, an average score of 65 on twelve essay questions, and a grade of at least 65 on at least five of those essays. *See* 2000 COMPREHENSIVE GUIDE, *supra* note 11, at Chart VII Supplemental Remarks. Nevada requires a combined MBE/essay score of 75 scaled points and a scaled score of 75 on at least three out of nine essays. *See id.*

20. *See, e.g.*, Stephen P. Klein, *On Testing: How to Respond to the Critics*, B. EXAMINER, Feb. 1986, at 16, 20.

that are one standard deviation above the mean on the state's distribution of MBE scores.

Because MBE scores are constructed so that a given score reflects a constant level of performance over time, scaling essay (or "state") scores to the MBE also assumes that any change over time in a state's average scores on the MBE is mirrored in a corresponding change in the essay score. This assumption rules out the possibility that average performance on the state portion of the bar exam might be declining at the same time that average performance on the MBE is improving or staying constant.

If the above assumptions about scaling are met, two important consequences follow. First, the scaling process eliminates any impact of grade inflation or changes in test difficulty on the state portion of the exam. MBE scores are adjusted so that a given score's meaning is constant from year to year; scores on the state portion of the exam, in turn, are standardized to match the mean and standard deviation of the state's MBE scores. If the overall quality of performance on a bar exam in a state, as measured by average scores on the MBE, remains constant from one year to the next, then the average standardized scores on the essay portion of that state's exam will also remain constant, regardless of whether the raw scores awarded on that portion of the exam change. Even if the state's essays were easier than essays in a previous year, so that raw scores rose, the standardized scores would remain the same as in the previous year. Similarly, even if essay graders became extravagantly lenient and "add[ed] . . . 100 points to every applicant's essay raw score," that grade inflation would "have absolutely no effect on an applicant's [standardized] essay . . . score and thereby on that applicant's chances of passing."[21]

Second, scaling both MBE and essay scores insures that the level of performance required to pass the bar remains constant across successive administrations. States set an absolute standard that successful applicants must achieve to pass; they do not pledge to pass a predetermined percentage of test takers. If applicant quality declines, the equator questions on the MBE will reveal that fact and MBE scaled scores will fall. Essay scores, scaled to the MBE, likewise will decline and fewer candidates will meet the state's passing score. Conversely, if applicant quality rises, scaled scores will rise and a higher percentage of candidates will exceed the state's passing score. Scores on the bar exam, in other words, follow a constant metric rather than a curve.

---

21. Klein, *supra* note 13, at 39.

II. WERE PASSING SCORES TOO LOW?

Once set, a passing score on the bar exam represents a constant level of competence. Those scores, however, still must be set. How does a state know if its passing score is too high or too low? Passing scores that are too low risk the admission of incompetent practitioners. Thresholds that are too high may hurt individual applicants, deprive the public of competent attorneys, raise the price of legal services, and limit increases in minority lawyers.

Most recent arguments in favor of raising bar passing standards either invoke unsubstantiated generalizations or misapprehend the nature of bar exam scoring. One set of arguments claims that recent bar examinees are less qualified than their predecessors. In 1996, for example, the president of Illinois's board of admissions charged that "People are getting into law schools who aren't qualified, and law schools are graduating people who aren't qualified to be lawyers."[22]

Empirical evidence, however, suggests that just the opposite is true—at least if one compares bar examinees during the 1990s with their peers in the 1980s. As Table I shows, average MBE scores since 1992 have been consistently higher than they were before that time. The average nationwide score in 1999, the latest year for which data are available, is higher than the average score for any year before 1992. As explained in the previous section, these scores represent a constant measure of applicant quality. Recent examinees have not achieved those high scores because today's MBE is easier than earlier versions of the test. Instead, those examinees are more competent than their predecessors—as measured by the bar exam itself.[23]

---

22. Klein, *supra* note 8, at A16 (quoting Stuart Duhl).

23. Other measures support this conclusion. Average LSAT scores of matriculating law students peaked among students entering law school in fall 1991, the same group of students who would achieve peak MBE scores three years later in 1994. *See* Klein, *supra* note 8 (quoting Erica Moeser, president of the National Conference of Bar Examiners). Indeed, variation in MBE scores over time closely tracks changes in LSAT scores. *See supra* note 14.

Table I: National Mean Scaled Scores for MBE, July Exams*

Year	Mean MBE	Year	Mean MBE
1980	140.6	1990	141.4
1981	140.8	1991	141.1
1982	139.7	1992	142.9
1983	141.5	1993	142.8
1984	139.2	1994	145.2
1985	140.6	1995	143.7
1986	140.3	1996	143.2
1987	140.3	1997	143.9
1988	139.8	1998	142.1
1989	142.0	1999	142.3

* Data drawn from *1999 Statistics*, BAR EXAMINER, May 2000, at 6, 20. Trends for the February exam are comparable, but scores are lower in each year.

Arguments based on declining applicant quality, moreover, overlook the manner in which scores on both portions of the bar exam are scaled. As explained above, the NCBE uses equator questions to adjust scores on each version of the MBE, while states scale raw scores from their state portions to the MBE. Together, these processes produce exam scores that represent a constant measure of performance. If applicant quality declines, scaled scores will decline as well and fewer applicants will meet the state's existing passing score. Scaling, in other words, controls for applicant quality without any need to change the passing score.

A related argument in favor of increased passing scores suggests that the quality of essay answers has declined while that of MBE performances has remained constant or risen. In Ohio, for example, the Board of Bar Examiners noted that it "had been concerned for some time about the quality of answers to the essay portion of the exam."[24] The current practice of scaling essay scores to MBE scores would mask any such decline; applicants might meet the old passing score even though their essay answers uniformly were inferior to those of a previous generation.

This concern about unique declines in essay quality, however, is at odds with the method most states use to scale essay scores to the MBE.

24. *Board of Bar Examiners Increases Bar Exam Passing Score*, ASSOCIATE NEWS, June 1996.

As explained in the previous section, that method assumes both that an average performance on the essay portion of the bar denotes a similar level of competence as an average score on the MBE, and that these reflections of competence vary in the same way over time. If today's examinees are, in fact, writing worse essays than their predecessors, while their MBE scores have remained constant or risen, then the scaling process—not the passing score—should be reassessed.

Raising the overall passing score on the bar, moreover, is a particularly inapt response to any concern over essay quality. As long as states continue to calibrate essay scores to MBE scores, applicants can meet higher passing standards by improving their performance on the MBE portion alone. Indeed, if applicants as a group improve their MBE performances sufficiently, they can surpass a higher minimum passing standard even with poorer quality essays.[25]

Anecdotal impressions of essay performance, finally, provide slim evidence that the quality of written answers has actually declined. In Ohio, as in most states, each examiner reads answers to only one of many essays on the exam; until recently the Ohio exam included eighteen essay questions.[26] The subjective impressions of individual examiners, therefore, rest on inspection of a small portion of each examinee's performance. Most examinees understand some subjects better than others; even the best applicants write poor answers to some questions.[27] Even among a group of highly qualified applicants, each essay question will generate some failing answers. Observers who read only answers to a single question are likely to conclude that a portion of the applicants is unqualified. If the examiners read the entire battery of essays written by each applicant, a more reliable measure of those applicants' quality, their assessment of the applicants' overall quality might well rise.

In addition to these concerns, some bar examiners may have favored higher passing scores because of worries about easier tests or grade inflation. As explained above, however, neither of these possibilities can affect bar scores. Equator questions on the MBE insure that the multistate exam maintains a consistent difficulty level from year to year. Scaling essay scores to that constant metric, in turn, assures that variations in question difficulty or grader leniency do not affect scores.

---

25. *Cf.* David M. White, Comments of David M. White, Testing for the Public, Case No. 96,869 (Fla. Apr. 6, 2000) (making complementary point that, even if applicants genuinely improve their essay answers, their overall scores—and the passing rate—will not rise unless MBE scores rise as well).

26. Starting with the July 2000 exam, Ohio reduced the number of essay questions to twelve and added the Multistate Performance Test.

27. For empirical support of this point, see *infra* notes 83-85 and accompanying text.

Easy tests and grade inflation may affect grades in law school classes, but they have no impact on the bar exam.

Perhaps the most common justification for raising passing scores has been simply to keep up with the standards in other states. Ohio, for example, announced that it had raised its passing score in part because "Ohio's standard was one of the lowest in the country, placing Ohio 43rd out of 47 jurisdictions."[28] Its board further noted that "[s]ince 1993, approximately 10 states with higher pass/fail standards ha[d] raised their passing scores even more."[29] To keep pace with that trend, Ohio decided to examine its passing score and ultimately to raise that score.

A simple desire to keep up with passing scores in other states, however, can have anti-competitive effects. Unless the states have independent evidence of attorney incompetence, this race to raise passing scores may produce unnecessary restrictions on entrance to the legal profession and higher priced legal services nationwide. States interested in competing to provide high-quality legal services would be better advised to match one another's continuing education programs, training programs for new attorneys, or disciplinary processes. High passing scores do not in themselves serve the public; in fact, they can have the opposite effect.

Although bar examiners have proffered these general arguments about attorney competence and the need for stricter bar admission standards, none has produced concrete evidence that existing standards are ineffective in preventing unqualified individuals from practicing law. Boards, for example, have not cited evidence that disciplinary complaints based on competence have been unacceptably high under current passing standards.[30] Indeed, Florida's records show that only six out of 365 disciplinary actions in the most recent year involved incompetence.[31] Nor have states cited other evidence of incompetence among lawyers passing the bar. In sum, states have raised bar passing scores without evidence that prevailing standards were inadequate, and despite evidence that examinees' average performance was improving.

A final argument that might justify raising bar exam passing scores is that law practice has become more difficult, so attorneys must be more competent today than in previous years. We are unaware of states that

---

28. *Board of Examiners Increases Bar Exam Passing Score*, supra note 24.

29. *Id.*

30. The question of what level of complaints is "unacceptable" requires a subjective judgment that each bar association would have to make for itself. We note only that the "unacceptable" level should account for the fact that some complaints may be unfounded or stem from a layperson's misunderstanding of applicable legal standards. A zero tolerance for competence complaints thus would be unrealistic.

31. *See* Noel G. Lawrence, Minority Report, Case No. 96,869, at 5-6 (Fla. Nov. 29, 1999).

have justified increases in the passing score on this basis. It would be extremely difficult, moreover, to judge whether on balance law practice today is more difficult than in previous decades. On the one hand, practice undeniably is more complex: complicated statutes, international rules, and alternative methods of dispute resolution exist today that did not exist twenty or thirty years ago. But on the other hand, attorneys have much more sophisticated means of research, database management, and communication than they had in those earlier years. One may need to know more to practice law today, but it may be easier to acquire and manage that knowledge.

Even if law practice is more difficult today than it was twenty years ago, raising the passing standard on current bar examinations is poorly tailored to measure any needed increase in competence. The content of most bar examinations has changed little in the last twenty years. In particular, the exam entirely omits most of the ways in which law practice has become more complex. Most bar exams do not test knowledge of civil rights statutes, environmental regulations, ERISA, or other modern statutory schemes. Nor do they touch upon international trade agreements, alternative methods of dispute resolution, or other forces that have revolutionized law practice during the last quarter century. Requiring applicants to answer correctly a few more multiple-choice questions about proximate cause or the rule against perpetuities—traditional subjects that are mainstays of the bar exam—in no way tests any increased competence needed for a twenty-first century law practice. On the contrary, the heightened need to memorize rules in these traditional subjects may discourage aspiring lawyers from taking the law school courses (*e.g.*, environmental law, international trade, alternative dispute resolution) that would better prepare them for a sophisticated practice.

In sum, all of the justifications offered to support higher bar passage standards lack empirical support, overlook controls already in place, prescribe the wrong remedy for an ill-defined disease, or restrict competition. Bar exams are not graded on a curve; constant standards rigorously maintained from year to year will identify (and fail) any increasing proportion of unqualified applicants. The scaling process likewise eliminates concerns about grade inflation or declines in test difficulty. Bar examiners have cited no concrete evidence of growing lawyer incompetence; on the contrary, MBE scores provide firm evidence that lawyers taking the bar exam during the last decade were more qualified than their predecessors. Concerns about essay quality or alterations in law practice would, if substantiated, require changes in scaling methods or the content of bar exams—not in the passing score. And the simple desire to match passing scores in other states, without

real evidence of attorney incompetence, risks reducing the supply of able attorneys available to serve the public without any countervailing benefit.

III. A SCIENTIFIC APPROACH TO CHOOSING A PASSING SCORE?

Prompted by the general concerns described above, several states have attempted to find an objective or scientific method to choose a passing score for their bar exams. The attempt to set standards objectively is laudable, but at least one of the methods used so far has serious defects. We describe this method, designed by psychologist Steven Klein and used in at least four states so far, below. We then identify the flaws in this method with the hope that states both will reexamine scores set by this method and work to devise more accurate standard-setting procedures in the future.

A. The Klein Method

Klein's method of setting a passing score appears unique, both among bar processes and in the general literature on setting educational or testing standards.[32] Klein first described his process in 1986, as one he had used to recommend a new passing score for the Puerto Rico bar exam.[33] More recently, Klein has used the process to recommend passing scores for the Ohio, Florida, Minnesota, and Pennsylvania bar exams.[34]

In brief, Klein collects expert judgments from regular bar graders, practicing attorneys, judges, and law professors about the quality of essays written on a recent bar exam. He then uses those judgments to estimate the percentage of examinees on that bar who would have failed the exam if the expert judgments had been applied. Once he has estimated that percentage, Klein determines the scaled score that would have produced that percentage of failing exams. For example, if Klein estimates from the expert judgments that thirty percent of the examinees

---

32. Klein himself describes the process as "eclectic." Klein, *supra* note 14, at 26.

33. *See id.* at 26-29.

34. These processes are described in reports prepared by Klein. *See* Stephen P. Klein, *Panelist and Reader Judgments Regarding the Passing Score on the Florida Bar Exam* (Aug. 12, 1999) [hereinafter "Florida Study"]; Stephen P. Klein, *Panelist and Reader Judgments Regarding the Passing Score on the Minnesota Bar Exam* (Feb. 10, 1998) [hereinafter "Minnesota Study"]; Stephen P. Klein, *Independent Panelist and Reader Judgments Regarding the Passing Score on the Ohio Bar Exam* (Jan. 5, 1996) [hereinafter "Ohio Study"]. We attempted to obtain a copy of Klein's study for the Pennsylvania bar examiners, but were not able to obtain that report. We base our comments about the Pennsylvania study on a letter authored by Klein. *See* Reply to Comments, *supra* note 9, app. at 1.

on a July 1998 exam should have failed that exam, and thirty percent of those examinees earned scaled scores below 135, then he will recommend 135 as the state's passing score.

Klein's method, notably, does not mean that thirty percent of future examinees will fail the bar exam. Bar exam scores, as explained above, establish a constant level of competence. If test takers study harder or take more bar preparation courses, a higher percentage of them may surpass the new hurdle. Conversely, if law schools graduate less qualified students, the failure rate could rise above thirty percent.

Many features of Klein's general approach are reasonable elements for choosing a passing score. Bar examiners, judges, practicing lawyers, and law professors are qualified to assess the quality of bar exam performances—although, without training, some of them may have unrealistic expectations about the type of essays examinees can produce in thirty or forty minutes. Using those judgments as part of a procedure to estimate the percentage of examinees who pass a threshold of minimal competence is a plausible, and recognized, method of setting a passing score.[35] The errors in Klein's method lie in the details, particularly in the way in which he uses judgments about answers to individual essay questions to predict performance on the bar as a whole.[36] Those flaws, unfortunately, undermine the entire approach, yielding an arbitrary passing score. To understand the flaws in Klein's process, and the unsound passing scores it produces, we detail the steps in his process below.

1. Post Hoc Judgments by Exam Graders

Exam graders in most states do not assign passing or failing scores to essays. Instead, they assign raw scores according to a predetermined scale. Graders in Ohio and Minnesota use a scale of 1 to 7; in Florida, they use one of 1 to 100.[37] In the first stage of his process, Klein asked

---

35. This is an "examinee centered" method of determining a cut score. *See* Allan S. Cohen, et al., *A Generalized Examinee-Centered Method for Setting Standards on Achievement Tests*, *in* 12 APPLIED MEASUREMENT IN EDUC. 343, 344, 345-46; Richard M. Jaeger, *Certification of Student Competence*, *in* EDUCATIONAL MEASUREMENT 485, 496-500 (Robert L. Linn ed., 3d ed. 1989). Consultants using this method ask experts to divide examinees into passing and failing groups on some basis other than their overall exam score. The basis of classification, however, may be a holistic assessment of the candidate's performance on the exam. After the candidates have been classified in this manner, a statistician determines the cut score for the exam that best matches the experts' categorization of applicants.

36. As explained further below, Klein essentially used recognized methods to develop cut scores for each of the individual essays he reviewed. The flaw in his technique lies in the way he combined these individual cut scores to generate an overall passing score.

37. *See* Florida Study, *supra* note 34, at 1; Minnesota Study, *supra* note 34, at 1; Ohio Study, *supra* note 34, at 1. Readers in Ohio occasionally assign scores of zero, *see* Ohio Study, *supra* note 34, at 1; it is

graders who had graded essays for a recent exam to designate the raw score that they felt most closely represented the minimum passing score on that batch of essays. For each question, Klein then checked bar records to determine the percentage of examinees who had achieved that score or better. Those percentages became the readers' "passing rates" for each question. Klein then averaged the passing rates for all questions on the exam to obtain an overall passing rate (as we show below, this is a fundamental error). Finally, he checked bar records to determine the overall scaled score that would have generated that percentage of passing exams.[38] Figure one illustrates this process for a hypothetical bar exam encompassing three essay questions graded on a scale of 1 to 5.

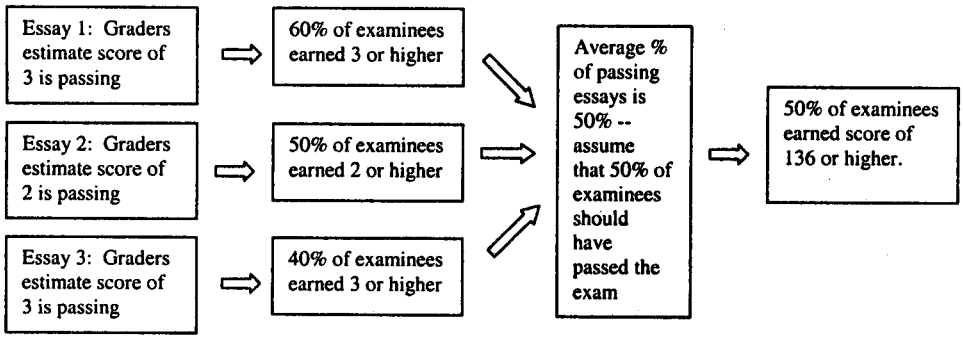

Figure 1

2. Judgments by Expert Panels

In the second part of his process, Klein convened expert panels of judges, practicing attorneys, and law professors. Klein worked with ten panels in Ohio, eight in Minnesota, and six in Florida. The Ohio panels included two to four members, while the Florida and Minnesota panels included four or five people.[39] The assigned task for each panel was to

---

not clear whether readers in the other two states follow the same practice or assign "one" as the lowest possible score.

38. Klein carried out this procedure for four different bar examinations in Ohio, for one examination in Florida, and for one in Minnesota. *See* Ohio Study, *supra* note 34, at 2; Florida Study, *supra* note 34, at 2; Minnesota Study, *supra* note 34, at 2. After filing his primary report, he carried out a second exam grader study in Florida. *See* Reply to Comments, *supra* note 9, app. at 10 (reprinting letter from Stephen P. Klein, to Missy A. Gavagni, Director of Examinations, Florida Board of Bar Examiners (Oct. 19, 1999)).

39. *See* Ohio Study, *supra* note 34, at 3; Florida Study, *supra* note 34, at 2; Minnesota Study, *supra* note 34, at 2.

assess the quality of essay answers to a single question from a recent bar exam.

Before panels met for the evaluation session, each member received a copy of the question they would evaluate. Klein encouraged each panelist to develop his or her own scoring guide for the question.[40] After developing their scoring guides, the panelists for each question convened for a standard-setting session in which they first discussed "the criteria they felt would be appropriate for their question, such as what an answer would have to cover to receive a passing grade."[41] Although all of the panelists discussed these criteria, they did not have to agree on them.

In Florida, each panelist then received 40 essays to grade; in Minnesota, the panelists graded 35 essays apiece; and in Ohio, 30 essays.[42] These essays were actual answers written for the bar exam and were chosen to represent a range of those answers. Each panelist gave each of the essays one of four grades: 1 ("clear fail"), 2 ("marginal fail"), 3 ("marginal pass"), or 4 ("clear pass"). The panelists, who were not told the scores the regular exam graders had assigned to their essays, worked independently in grading.[43]

Klein repeated this panelist process in Ohio for ten of the eighteen essay questions from that state's July 1995 bar exam.[44] In Florida, he used all three of the essay questions from the July 1998 and February

---

40. In Ohio, Klein apparently provided panelists with no information about how regular readers had scored the question. His report for that state notes, "[t]o insure their complete independence, the panelists were not given the scoring guide the regular readers used for that question." Ohio Study, *supra* note 34, at 3. This differed from Klein's original design in Puerto Rico, where he provided panelists with a "copy of the . . . readers' scoring guide" and "advised [them] that they [could] modify this guide as needed." Klein, *supra* note 14, at 26. Klein's descriptions of the Minnesota and Florida studies do not indicate whether the panelists received any information about the regular readers' scoring criteria. A copy of the instructions sent to Florida panelists, however, indicates that they received a copy of the model answer for the question they assessed. *See* Reply to Comments, *supra* note 9, app. at 9.

41. Ohio Study, *supra* note 34, at 3; Florida Study, *supra* note 34, at 2; Minnesota Study, *supra* note 34, at 2.

42. *See* Florida Study, *supra* note 34, at 3; Minnesota Study, *supra* note 34, at 2; Ohio Study, *supra* note 34, at 3.

43. It appears that panelists were allowed to discuss their judgments with one another, but that they were encouraged to reach independent decisions on each essay they graded. *See* Reply to Comments, *supra* note 9, app. at 9 (reprinting directions to Florida panelists) ("You can discuss answers with the other members of your team, such as whether an issue that is raised by an applicant is relevant, but please use your own judgment in deciding whether the answer as a whole merits a passing grade.").

In Ohio, the panelists evaluated essays in batches of ten, pausing after each batch to discuss their ratings. *See* Ohio Study, *supra* note 34, at 3. It does not appear that the panelists could change their ratings after these discussions, because they first turned in their score sheets. *See id.* Klein's reports do not mention whether this process was used in Florida and Minnesota.

44. *See* Ohio Study, *supra* note 34, at 3. Until July 2000, the Ohio bar exam included the MBE and eighteen essay questions, with the total essay score given twice the MBE's weight in calculating the final score. *See id.* at 1-2; *supra* note 26.

1999 exams, for a total of six essay questions.[45] And in Minnesota, he used all eight of the essay questions from the state's July 1997 exam.[46]

Once the panelists had graded the essays for their given question, Klein calculated the average of the panelists' grades for each essay. He assumed that essays with an average score of 2.5 or higher were passing essays, while those with average scores below 2.5 were failing essays. He then determined what raw score from the range of scores used by the regular exam graders corresponded to a panelist average score of 2.5.[47] In Ohio, for example, where graders score essays on a scale of 1 to 7, he found that on one of the essay questions the regular grader's scores of 1, 2, and 3 most closely corresponded to what the panelists designated as failing answers, while scores of 4, 5, 6, and 7 corresponded to passing answers.[48] In Florida, where essays can receive up to 100 points, he found that scores under 47 corresponded to failing answers on one essay question, while those of 47 or more corresponded to passing answers.[49]

Klein then determined the percentage of the actual bar examinees who received scores equal to or above each essay's new "passing score." As one would expect, given the variety of subjects tested on the bar and the variable difficulty of essay questions, these percentages differed among questions.[50] In Ohio, the passing rate across the ten questions that Klein had panelists assess varied from a low of 57% to a high of 89%.[51] In Florida, the passing rates for individual questions varied from 36% to 87%.[52] And in Minnesota, they varied from 58% to 92%.[53]

---

45. *See* Florida Study, *supra* note 34, at 3-4. The Florida bar exam consists of the Multistate Bar Exam, three essay questions, and 105 to 120 additional multiple choice questions. The MBE and state portion (*i.e.* the three essay questions and remaining multiple choice items) are weighted equally in computing a final score. *See id.* at 1-2.

46. *See* Minnesota Study, *supra* note 34, at 1. Minnesota's bar exam includes the MBE and eight essay questions, with the two portions weighted equally in computing a final score. *See id.*

47. *See* Florida Study, *supra* note 34, at 3; Minnesota Study, *supra* note 34, at 3; Ohio Study, *supra* note 34, at 4.

48. *See* Ohio Study, *supra* note 34, at 4. Similarly, in Minnesota (which uses the same scoring scale as Ohio for essays), Klein found that scores of 1, 2, or 3 on one essay question corresponded with the panelists' collective judgment of failing while scores of 4 and above corresponded with passing. *See* Minnesota Study, *supra* note 34, at 3.

49. *See* Florida Study, *supra* note 34, at 3.

50. The percentages also varied depending on the method Klein used to calculate them. Klein used two different methods to generate these percentages. As his reports show, the two methods yielded similar results overall–although they varied considerably in the percentages of passing applicants they predicted for individual questions. *See* Ohio Study, *supra* note 34, at 10-11; Florida Study, *supra* note 34, at 10-11; Minnesota Study, *supra* note 34, at 8-10. For simplicity, we use percentages from Klein's cross-classification method throughout this Article.

51. *See* Ohio Study, *supra* note 34, at 10.

52. *See* Florida Study, *supra* note 34, at 10.

53. *See* Minnesota Study, *supra* note 34, at 9.

In the next stage of the expert panel process, Klein averaged the question-specific passing rates for all of the essays his panels had evaluated. In Florida, the average passing rate across all three essay questions for both exams Klein evaluated was 56%.[54] In Ohio, the average passing rate for the ten essays evaluated was 69%.[55] And in Minnesota, it was 78%.[56]

In a key step in his process, Klein assumed that the average of the passing rates for each essay equals the passing rate for the exam as a whole. He then used these overall averages to find the scaled bar exam score that would match that passing rate. In Ohio, for example, he found that 69% of the candidates who took the July 1995 exam (the exam for which panelists judged essays) scored 410 or higher on that state's scaled scoring system.[57] Similarly, in Minnesota, he determined that 78% of the applicants who took the July 1997 exam scored 271 or more on that state's scale.[58] The Florida study yielded somewhat different cut scores for the two exams assessed: 139.5 for the July 1998 exam and 135 for the February 1999 test.[59] We illustrate the stages of Klein's expert panel studies in Figure 2, again using a hypothetical bar exam consisting of three essay questions each graded on a scale of 1-5.

54. *See* Florida Study, *supra* note 34, at 10. The identical rates, however, led to somewhat different cut scores for the two exams. *See id.*

55. *See* Ohio Study, *supra* note 34, at 10.

56. *See* Minnesota Study, *supra* note 34, at 9.

57. *See* Ohio Study, *supra* note 34, at 10.

58. *See* Minnesota Study, *supra* note 34, at 3.

59. *See* Florida Study, *supra* note 34, at 10.

Essay 1: 4 panelists each assess 30 essays
Average 4 scores for each essay; 2.5 is "passing"
Reader grade of 2 best matches panelists' passing line
60% of examinees received grade of 2 or higher
Essay 2: 4 panelists each assess 30 essays
Average 4 scores for each essay; 2.5 is "passing"
Reader grade of 3 best matches panelists' passing line
40% of examinees received grade of 3 or higher
Essay 3: 4 panelists each assess 30 essays
Average 4 scores for each essay; 2.5 is "passing"
Reader grade of 3 best matches panelists' passing line
50% of examinees received grade of 3 or higher
Average percentage of passing essays is 50% -- assume that 50% of examinees should have passed exam
50% of examinees earned score of 238 or higher

Figure 2

3. Recommending a New Passing Score

To recommend a new passing score, Klein reported both the passing scores gleaned from the exam graders and those generated from the expert panels. The different studies sometimes generated a wide range of results. In Ohio, for example, the four studies of exam graders yielded overall passing scores of 389, 392, 405, and 416, while the study based on expert panels produced a score of 410.[60] Similarly, the recommended scores in Florida ranged from 133.5 to 141.[61] In Minnesota, on the other hand, the three predicted scores clustered at 270, 271, and 272.[62]

In all three of these states, the scores generated by Klein's studies exceeded the state's existing passing score.[63] In Pennsylvania, on the other hand, the projected score was one point lower than the score that state recently had adopted.[64] The latter score, however, had been raised by six points during the mid-1990s.[65]

In addition to reporting the passing scores from his studies, Klein noted in each of his reports how the state's passing score compared with that of other states—and included a chart showing the state's relative ranking among other states. In Ohio, Florida, and Minnesota, he pointed out that those states had passing scores well below scores maintained in other states.[66] His final recommendation that these three states raise their passing scores was based on both the results of the

---

60. *See* Ohio Study, *supra* note 34, at 2, 4.

61. The four recommended scores were 133.5, 135, 139, and 141. *See* Florida Study, *supra* note 34, at 4; Reply to Comments, *supra* note 9, app. at 10.

62. *See* Minnesota Study, *supra* note 34, at 4. Scores in all three states use somewhat different scales. Values on the Minnesota scale are twice as large as those on the Florida scale, while the Ohio scale uses values that are three times as large as those on the Florida scale. Converting the Minnesota and Ohio numbers to the Florida scale yields recommended scores ranging from 129.7 to 138.7 in Ohio and 135 to 136 in Minnesota.

63. In Ohio, the passing score was 375. *See* Ohio Study, *supra* note 34, at 5. In Florida, it was 131. *See* Florida Study, *supra* note 34, at 4. And in Minnesota, it was 260. *See* Minnesota Study, *supra* note 34, at 4.

64. Reply to Comments, *supra* note 9, app. at 4 (noting that Klein study yielded passing score of 136 rather than the 137 used by the state). The process used by Puerto Rico in the 1980s also generated a lower score than the one that jurisdiction had used. *See id.*

65. *Compare* ABA SECTION OF LEGAL EDUCATION AND ADMISSIONS TO THE BAR AND THE NATIONAL CONFERENCE OF BAR EXAMINERS, COMPREHENSIVE GUIDE TO BAR ADMISSION REQUIREMENTS 1990, at Chart VII (1990) (showing that Pennsylvania required a combined scaled score of 129), *with* 2000 COMPREHENSIVE GUIDE, *supra* note 11, at Chart VII (requiring combined score of 270, which is equivalent to 135 on the scale previously used by Pennsylvania). Several commentators noted the dramatic rise in Pennsylvania's passing score.

66. *See* Ohio Study, *supra* note 34, at 5; Florida Study, *supra* note 34, at 4; Minnesota Study, *supra* note 34, at 4.

studies he had conducted and the low scores used by those states compared to other states.

After considering Klein's report, the Ohio Board of Bar Examiners proposed raising the passing score on that state's exam from 375 to 405. The Ohio Supreme Court approved that change in 1996, with the increase taking place in two steps.[67] Similarly, the Florida Board of Bar Examiners proposed raising that state's passing score from 131 to 136. The Florida Supreme Court solicited public comments on the proposal and has it under advisement.[68] The Minnesota Board of Law Examiners likewise proposed raising that state's passing score from 260 to 270 points, and the Minnesota Supreme Court is weighing that proposal.[69]

B. Flaws in the Klein Method

Although Klein's method attempts to set bar passing rates rigorously, the method has serious flaws. In particular, as we explain in more detail below, Klein's process confuses the percentage of passing *essays* with the percentage of passing *test takers*. His implicit assumption, that each essay represents an independent measure of a unitary "competence" dimension, is not true for the bar exam. On the bar exam, individual questions tap different bodies of knowledge. In this type of exam the overall percentage of passing essays has no reliable relationship with the percentage of passing test takers. The passing score generated by both Klein's study of exam graders and his study of expert panelists thus is an arbitrary figure.[70]

Klein's method is problematic precisely because it represents a good faith attempt to set passing scores using statistical methods. The apparent rigor of the process, combined with statistical techniques that most bar examiners and lawyers do not understand, creates a deceptive aura of scientific objectivity. It is difficult for examiners and other decisionmakers, who must rely upon statistical consultants, to see the flaws in the process. Statistical techniques are powerful and we favor

---

67. *See Board of Examiners Increases Bar Exam Passing Score*, *supra* note 24.

68. Reply to Comments, *supra* note 9, at 1-2, 6-8.

69. Letter from John D. Kelly, President, and Margaret Fuller Corneille, Director, Minnesota Board of Law Examiners, to Harry J. Haynesworth, Dean, William Mitchell College of Law (Aug. 11, 1999) (on file with the *University of Cincinnati Law Review*).

70. The figure is arbitrary both in the three states (Ohio, Florida, and Minnesota) in which Klein recommended raising the passing score and in the one state (Pennsylvania) in which he suggested a modest decline. *See also supra* note 65 (noting that recommended decline in Pennsylvania score followed dramatic rise in that score).

their use in many contexts. They must be used properly, however, to produce valid results.

We concentrate in this section on the panelist studies Klein conducted, because they were the basis for the bulk of his reports in Florida, Minnesota, and Ohio. As we note briefly below, however, the first and most fundamental defect in Klein's method affected the regular grader exercises as well as the panelist studies.[71] Neither one of these procedures, therefore, supports the passing scores Klein recommends.

1. Confusing Passing Essays with Passing Exam Takers

In each of his panelist studies, Klein sampled essays written in response to a single question, submitted those essays to expert panelists for review, and used their judgments to estimate the percentage of passing essays on that question. He repeated this process for several different questions, using different panelists for each question,[72] and then averaged the passing rates from each question to yield an overall passing rate. Klein assumed that the latter figure was a valid estimate of the percentage of *test takers* who should pass the bar. In fact, however, Klein's average is merely an estimate of the percentage of passing *essays* produced by all applicants on all questions. To generate a useful estimate of the percentage of passing *test takers* (the percentage Klein then used to compute a passing score), one would also have to know (1) the distribution of passing essays across test takers,[73] and (2) the number

---

71. *See infra* text preceding note 96. Other concerns could be raised about the reader exercises, although we do not have space to explore them here. From the brief description in Klein's reports, for example, it appears that he asked readers to name the score that corresponded with "passing" on their essays without having them review a representative group of those essays. Retrospective recollections of this nature may be quite unreliable, especially because the grades readers assigned were *not* keyed to explicit pass/fail standards. Instead, those grades merely constituted comparative judgments among the essays reviewed.

72. Klein also drew the essays for each panel from different examinees. *E.g.*, the essays reviewed for the first essay question in each state were written by different candidates than the essays reviewed for other questions in that state.

73. We use "distribution" in the everyday sense of that word. That is, if one arrayed the essay scores of every candidate in a table, what would the pattern of scores look like? How many candidates would have high scores on all of the essays and how many would have high scores on none of them? Would scores on one essay show a relationship to scores on other essays; would candidates who scored highly on the first essay be likely to score highly on the second? Would the relationship between scores on the first and third essays be stronger or weaker than the relationship between scores on the first and second essays?

This everyday notion of distribution incorporates the statistical notions of both "distribution" and "correlation." In statistics, a "distribution" summarizes how frequently each possible outcome occurs. For Minnesota, for example, the distribution of passing answers for test takers would indicate how many of the test takers passed all eight questions, how many passed seven questions, how many passed six questions, and so forth, down to how many test takers failed all eight questions.

A "correlation" summarizes the strength of the relationship between two variables. If scores on two essays show a high positive correlation, that means candidates who scored highly on one essay were quite

of essays that a minimally qualified applicant could fail while still passing the bar.

In this section, we first show that Klein's method estimates the percentage of passing *essays* written by all applicants on all questions. We then explain why that percentage of passing *essays* is not a reliable predictor of the percentage of passing *test takers*. Although the flaw in Klein's process is subtle, it negates any value in his recommended passing scores.

a. Percentage of Passing Essays

Recall that Klein asked each of his panels to review a sample of essays responding to a single question on the bar exam. Based on those judgments, Klein estimated the percentage of passing essays responding to that question among all candidates who took the exam. Florida's July 1998 exam, for example, included three essays. The panelists who reviewed the first essay made judgments that, under Klein's calculations, suggested that 43% of all applicants had passed that question. A second set of panelists produced judgments suggesting that 59% of all applicants had passed the second question. And a third set of panelists yielded judgments suggesting that 65% of the applicants had passed the final question.[74]

Florida's records reveal that 2,077 candidates took the July 1998 bar exam.[75] Those test takers each wrote three essay answers, for a total of 6,231 essays.[76] From his panelist studies, Klein determined that 43% of the answers to the first essay (893 of those 2,077 essays) passed; 59% of the answers to the second essay (1,225 of those 2,077 essays) passed; and 65% of the answers to the third essay (1,350 of those 2,077 essays) passed. In all, therefore, 3,468 of the 6,231 essays written for that exam

---

likely to score highly on the second essay as well. If scores show a high negative correlation, that means that candidates who scored highly on one essay were quite likely to score poorly on the second essay. And if scores show no correlation, that means that candidates who scored well on the first essay might have scored well or poorly on the second essay; knowing the first essay score would not allow us to predict the second essay score.

74. *See* Florida Study, *supra* note 34, at 10.

75. *See* 1998 Statistics, *supra* note 3, at 6, 7.

76. It is possible that one or more applicants skipped an essay question. The failure to answer a question, however, is still an "essay answer" in the sense that the failure contributed to the scoring. An applicant who skips a question presumably receives a zero for that essay question.

were passing essays while 2,763 were failing ones.[77] Another way to say this is that about 56% of the 6,231 essays were passing essays.[78]

When Klein averaged the passing rates for these three essay questions, he merely obtained this percentage of passing essays written by all exam takers on all three questions: 56%.[79] Overall, according to the expert panelists, about 56% of the 6,231 essays written by the 2,077 applicants demonstrated minimal competence. Klein assumed this meant that about 56% of the applicants were minimally competent. He assumed, in other words, that the average percentage of passing essays equaled the percentage of passing candidates. But is this true?

b. The Percentage of Passing Test Takers

Klein's assumption, that the average percentage of passing essays equals the percentage of passing test takers, is valid under certain rare conditions. For example, imagine a situation where all test takers who passed any one question on the exam also passed all other questions, while test takers who failed any one question also failed all other questions. In this case the test takers would be clearly split into two groups, one that passed all of the questions and one that failed all of them. Under this condition the average proportion of passing essays would also equal the proportion of passing test takers. We doubt that this situation has ever occurred—and it could not have occurred in any of the states using Klein's process because the panelists in those states estimated that different percentages of test takers passed each of the essay questions.

Another condition under which equating the proportions of passing essays and passing test takers would be justifiable would be if all questions on the bar exam measured the same skill, although not perfectly. In this situation candidates' performances on the different essays would be highly correlated (but not perfectly correlated as in the situation discussed above). That is, candidates who excelled on one essay would usually excel on the others, while candidates who failed one essay would be likely to fail others. On the Florida exam discussed above, for example, the 893 candidates who passed the first (and most difficult) question would almost all have passed the other two questions.

---

77. If we add the number of passing essays from each question (893 + 1,225 + 1,350) we find the total number of passing essays (3,468). If we then subtract the passing essays (3,468) from the total number of essays (6,231) we find that 2,763 were failing essays.

78. 3,468 divided by 6,231 yields .56.

79. *See* Florida Study, *supra* note 34, at 10.

Likewise, the 727 candidates who failed the third (and easiest) question would almost all have failed both of the other two questions as well.

Klein's reports suggest that he assumes that questions on a bar exam have this property. In his Ohio report, he noted that "the estimate of the overall pass/fail standard that was derived from the panelists ratings was based on ten separate and completely independent mini-studies; i.e., each question was its own mini-study. Put another way, there were essentially ten replications of the study design."[80] Klein assumed, in other words, that "competence" as measured by the bar exam is a one-dimensional trait and that each essay measures that same trait.[81] If that were true, it would be plausible to consider the results of each panel a "mini-study" estimating the percentage of test takers possessing that one-dimensional competence.

Importantly, however, the bar exam is not structured to measure one dimension of competence—or even to measure the same set of competencies in every question. Exam drafters tap *different* bodies of knowledge with different questions, as well as measuring some common reading, writing, and analytic skills in all questions.[82]

In part because questions draw upon different bodies of knowledge, candidates' scores on different questions show *low* correlations. We examined a sample of score reports for the July 1998 Ohio bar exam,[83] an exam that contains eighteen different essay questions. As Table II shows, inter-question correlations were usually small. Of the 153 separate correlations, only eighty-four (54.9%) were statistically significantly different from zero. None of the correlations were as high

---

80. Ohio Study, *supra* note 34, at 4; *see also* Florida Study, *supra* note 34, at 4 ("The results above provide a reliable basis for setting Florida's passing score because they were derived from six replications of the same basic study design . . . ."); Minnesota Study, *supra* note 34, at 4 ("The foregoing estimates of the panelists' overall pass/fail standard were derived from eight separate and completely independent mini-studies; i.e., each question was its own mini-study. Put another way, there were essentially eight replications of the study design.").

81. Alternatively, he could have assumed that the bar measures a package of competencies, but that each question taps the entire package. In either case, the key point is that Klein assumed that all essay questions on the bar measure the same skills. Otherwise, it would not make sense to speak of the different questions as "mini-studies" of the same competence.

82. *See, e.g.*, Marcia Kuechenmeister, *Admission to the Bar: We've Come a Long Way*, B. EXAMINER, Feb. 1999, at 25; Julia C. Lenel, *The Essay Examination Part I: The Problem of Reliability*, B. EXAMINER, Feb. 1990, at 18; Julia C. Lenel, *The Essay Examination Part II: Construction of the Essay Examination*, B. EXAMINER, May 1990, at 40.

83. Our sample includes all 158 graduates of The Ohio State University College of Law who took that exam. This is not a random sample of all Ohio examinees. The sample, however, includes exam takers in every decile of overall performance; indeed, one sample member was in the 0 percentile for final score while another ranked in the 100 percentile (according to the Board of Bar Examiners' reports). The sample contains a disproportionate percentage (17.7%) of examinees in the top decile of overall performance and some underrepresentation of examinees in the bottom four deciles. Of the 158 sample members, however, at least nine fell in every decile of statewide performance.

as .50, and only ten (6.5%) were .30 or higher. Six of the correlations were *negative*, suggesting that test takers who scored well on one question actually tended to score poorly on the second question in that pair.[84]

A more rigorous way to show that individual essay questions measure different competencies is to conduct a factor analysis of the scores from those individual questions. Factor analysis is a statistical technique that identifies the number of independent dimensions accounting for the correlations among a set of scores.[85] When we factor analyzed the scores on the 18 essay questions obtained by the sample of 158 candidates who took the July 1998 Ohio bar exam, we found that *six* different factors were necessary to account for the correlations among the scores on those questions. The competencies measured by different essay questions, in other words, vary widely.

Under these circumstances, estimating the percentage of bar examinees who are minimally competent by averaging the percentages who demonstrate minimal competence on individual essay questions is analogous to determining the percentage of physically fit college students by averaging the percentage who pass a sit-up test, the percentage who pass a pull-up test, and the percentage who pass a sprinting test. Abdominal strength, upper arm strength, and sprinting capacity are all components of physical fitness, but they are different dimensions of that fitness.[86] An individual might demonstrate minimal physical fitness by excelling in one of these dimensions, performing moderately well at another, and failing a third.

---

84. Similarly, correlations between individual essay scores and a candidate's total score on the MBE portion of the bar exam vary widely. In our sample of Ohio examinees, scores on three of the eighteen essay questions showed no correlation with MBE scaled scores. These three correlations were all less than .10 and lacked statistical significance in our sample. Four other correlations fell below .30, but were statistically significant; three fell between .30 and .40; six fell between .40 and .50; one reached .51 and another reached .57. Notably, the correlation between an applicant's *total* raw essay score and total scaled MBE score was much higher than the correlation for any one essay score: .71. This is consistent with many other studies of bar exams. *See, e.g.*, Klein, *supra* note 20, at 18-19. Finding that performance over a range of essays, which tap many dimensions of legal competence, is consistent with performance on a multiple choice exam likewise tapping those many fields, however, is different from concluding that performance among subareas is highly correlated. Our analyses of the Ohio data confirm that the latter correlations are relatively low.

85. *See, e.g.*, LAWRENCE C. HAMILTON, STATISTICS WITH STATA 5, 279-90 (1998).

86. The dimensions, of course, share some common base. A certain degree of athleticism or training may contribute to an enhanced ability to complete sit-ups, pull-ups, or a hundred-yard dash. The three tests, therefore, both measure separate dimensions of fitness and tap some common sources of fitness. Similarly, bar exam questions both tap common skills (reading, writing, and analytic ability) and specialized bodies of knowledge (torts, trusts, commercial transactions) or reasoning ability (statutory interpretation, counseling, open ended reasoning).

Table II: Inter-question Correlations on July 1998 Ohio Bar Exam
N = 158

	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14	Q15	Q16	Q17	Q18
Q1	.01	-.07	.13	.09	.16	.22	.19	.25	-.08	.16	.19	.31	.27	.13	.08	.23	.24
Q2		.13	.26	.15	.10	.21	.23	.10	.11	.14	.15	.20	.25	.18	.15	.12	.21
Q3			.13	.15	.12	.04	.10	.10	.07	-.02	.06	.05	.09	.15	.09	.08	.15
Q4				.23	.28	.47	.17	.21	.10	.11	.17	.14	.20	.19	.22	.23	.11
Q5					.12	.23	.02	.25	-.08	.11	.07	.21	.16	.25	.04	.18	.15
Q6						.28	.16	.23	.13	.12	.13	.23	.13	.08	.23	.04	.24
Q7							.18	.28	.23	.24	.18	.19	.18	.14	.21	.25	.22
Q8								.23	.13	.16	.23	.31	.21	.14	.29	.14	.24
Q9									.23	.22	.28	.36	.43	.12	.24	.18	.31
Q10										.06	.16	.08	.06	-.02	.02	.03	.07
Q11											.30	.27	.23	.04	.16	.15	.11
Q12												.25	.31	-.12	.23	.11	.22
Q13													.45	.06	.28	.16	.26
Q14														.09	.29	.27	.29
Q15															.05	.25	.25
Q16																.20	.27
Q17																	.30

Correlations shown in bold were significant ($p \leq .05$) in a two-tailed test.

If we tested a group of college students on these three dimensions of fitness, we might find that 70% passed the sit-up test, 60% passed the pull-up test, and 80% passed the sprinting test. This would *not* mean, however, that just 70% of the students (the average of these three passing rates) were fit. To know the percentage of fit students, we would have to know how these passing scores were distributed and whether a student had to pass all three tests to demonstrate minimal physical fitness. If we decided that an individual demonstrates minimal fitness by passing two of the three tests, and if performances on these tests are negatively correlated (so that students who fail one test pass the other two), then 100% of the students could be minimally fit. Conversely, if we decided that individuals must pass all three tests to be minimally fit, then no more than 60% of the students (the percentage passing the most difficult test) could be fit.

The three dimensions of physical fitness are analogous to the multiple essays on bar exams. Different questions measure different dimensions of an overall competence to practice law, while also tapping some general legal skills. Averaging the percentage of test takers who demonstrated minimal competence on each essay tells us the percentage of *essays* that were minimally competent, but it does not tell us the percentage of *test takers* who achieved that mark. To determine the percentage of minimally competent test takers, which in turn determines Klein's passing score, we need to know both the distribution of passing essays among test takers and the number of essays an applicant must pass to demonstrate minimal competence. Ignoring these two points overlooks the multi-dimensional nature of legal competence and the bar exam itself. Fitness to practice law embraces many competencies, and the bar exam measures those competencies through different questions.

As an example, consider the eight essay questions on Minnesota's bar exam. Klein's panelist studies determined the percentage of passing essays written in response to each of those questions on the July 1997 exam. The percentages are summarized in Table III. To compute a passing score for Minnesota, Klein simply averaged the passing rates for the eight questions and assumed that this average (78%) represented the percentage of passing test takers.

Table III: Passing Rates for Eight Essay Questions on the July 1997 Minnesota Bar Exam (as determined by Stephen Klein through Panelist Studies)

Question	Passing Rate	Question	Passing Rate
1	91%	5	75%
2	82%	6	86%
3	92%	7	67%
4	58%	8	72%

Average Passing Rate: 78%

We know from our discussion of inter-question correlations and multiple competencies, however, that the percentage of passing essays does not necessarily equal the percentage of passing test takers. How were the passing essays on the Minnesota bar exam distributed? In other words, how likely were candidates who passed one essay to pass the other essays as well? And how many essays did a candidate have to pass in order to be judged minimally competent? Clearly some competent candidates failed one or more essays; otherwise the passing rate in Minnesota could be no higher than 58% (the percentage who passed the hardest essay question).

One could generate hundreds of possible distributions of the passing and failing essays on the Minnesota bar exam. We, however, quickly generated a hypothetical sample distribution that fits the information we have about bar exam performance and inter-question correlation.[87] In this simulation, about one quarter (27%) of the candidates passed all eight essays; one fifth (22%) passed seven of the essays; another fifth (22%) passed six essays; a somewhat smaller percentage (17%) passed five essays; 4% passed four essays; 4% passed three essays; and 4% passed two essays. None of the candidates in this simulation passed zero or one essay.[88]

---

87. Our simulation is based on a round number of 100 hypothetical exam takers to simplify calculations. The percentage of passing essays on each of the eight questions matches the percentages computed by Klein from his panelist studies. For example, 91 of the hypothetical exam takers in our simulation passed the first question while 82 passed the second one. A spreadsheet showing the full simulation is available from the authors.

88. This distribution is quite consistent with the distribution of passing and failing essays on the Ohio bar exam, the only exam for which we have detailed information about actual score distributions. In our sample of 158 examinees who took that exam, *see supra* note 83, all ten who ranked in the bottom decile statewide passed at least three of the eighteen essay questions. Indeed, six of these low scoring examinees passed at least *six* of the eighteen essay questions. Conversely, among the 28 sample members who ranked

The inter-question correlations in this simulation are realistic. Table IV shows those twenty-eight separate correlations. As the table indicates, the correlations range from -.15 to .51. Only five of the correlations are negative and all of these are barely distinguishable from zero. Six of these correlations (21%) are higher than .30, and one is above .50. Overall, the essay scores in this simulation appear more highly correlated than the scores from the actual Ohio distribution described above.[89]

Table IV: Inter-Question Correlations for Hypothetical Distribution of 100 Candidates Taking July 1997 Minnesota Bar Exam

	Q2	Q3	Q4	Q5	Q6	Q7	Q8
Q1	.22	.42	.09	.22	.28	.15	.19
Q2		.25	.02	.51	.19	.39	-.06
Q3			.05	.17	.31	.11	-.10
Q4				.21	-.05	.39	.37
Q5					.10	.14	.10
Q6						.02	-.12
Q7							-.15

Within this simulated population, which uses the per-question passing rates calculated by Klein and a realistic distribution of passing essays, what percentage of examinees should have passed the bar exam? Klein's method would estimate 78%, the average passing rate across the eight questions (or the percentage of passing essays written for the exam). His estimate disregards both the actual distribution of passing and failing essays (whatever that might be) and the number of questions experts believe a passing candidate may fail.

---

in the top decile statewide, twelve (42.9%) failed at least one essay. Indeed, one of these high ranking examinees failed *three* of the eighteen essays. In making both of these calculations about Ohio performance, we used conservative estimates of passing and failing essays. That is, when counting passing essays in the bottom decile, we counted only essays receiving a score of 4 or higher. When counting failing essays in the top decile, we counted only essays with a score of 2 or lower. *See* Ohio Study, *supra* note 34, at 10 (showing that panelists' assessment of passing essays corresponded with scores of 3 or 4 depending on the question).

The distribution of passing and failing essays we simulated for Minnesota actually shows a somewhat greater concentration of passing and failing essays than we identified in Ohio. Thus, more than a quarter of the candidates in this simulation passed all eight essays. Our estimate of the percentage of Minnesota candidates passing the bar exam thus is probably somewhat low.

89. We do not calculate significance for the correlations in this table, because that calculation depends upon the number of exams in the simulation, and our choice of 100 exams was merely a matter of convenience. As the number of exams increases (with the same distribution of answers), the correlations would remain the same but more would be statistically significant.

Suppose, however, that the panelists in Klein's study indicated that they believed a minimally competent candidate could fail three of Minnesota's eight essay questions while still passing the bar.[90] Klein did not obtain this information from his panelists but, as we have explained above, it is crucial when setting a passing score for an exam that measures multiple dimensions of competence. The broad range of material tested on the bar, as well as the conditions under which examinees answer questions, makes it unrealistic to expect examinees to pass every essay question. Our suggestion that experts might agree that a minimally competent applicant could fail three out of eight questions, moreover, is realistic. Analyses of actual scores from Ohio show that the overwhelming majority of candidates fail some essay questions—and that minimally competent candidates fail multiple essay questions.[91] States that require candidates to pass a minimum number of essays, moreover, tend to require candidates to pass just over half of the essays. Delaware, for example, which ranks with California as one of the toughest bar exams in the nation, allows passing candidates to fail five of its twelve essays.[92] Rhode Island, which also maintains one of the nation's highest passing scores, likewise allows passing examinees to fail five out of twelve essays.[93]

If we assume that a minimally competent candidate could fail three of the eight essays administered in Minnesota, then 88% of the candidates in the simulation described above were competent.[94] If this

---

90. We realize it is somewhat artificial to talk about how many essays a candidate can "fail" while still passing the bar; essays in Minnesota and other states are graded on a multi-point scale rather than a simple pass/fail dichotomy. Bar examiners, moreover, sum a candidate's scores over all essays. It is possible that a candidate who "fails" two essays with scores just below the hypothetical pass/fail line, but passes the third essay with a top score, is better qualified than a candidate who passes all three essays with barely passing grades. The former candidate might also receive a higher total raw score on the essay portion of the exam.

Klein's entire method, however, relies upon the assumption that essays can be classified into passing and failing categories–and that this classification, in turn, can generate a passing score for the entire bar exam. Any artificiality of this assumption, therefore, derives from Klein's method and might be considered another criticism of that method. We simply incorporate Klein's own assumptions to show the fallacy of equating the percentage of failing essays with the percentage of failing exams.

91. Returning to our sample of candidates who took the July 1998 Ohio bar exam, *see supra* note 83, and using a conservative measure of "failing," *see supra* note 88, we found that 42.9% of candidates in the top decile (for overall performance) failed at least one of the essay questions. Among those in the fourth decile—the group that just exceeded Ohio's new, more stringent passing score—all candidates failed at least one essay and 60% failed three or more. Using a more realistic measure of failing essays (one that counts essays with a score of "3" as failing, *see supra* note 88), all candidates in this decile failed at least six essays, 80% failed at least seven, and 10% failed nine.

92. *See* 2000 COMPREHENSIVE GUIDE, *supra* note 11, at Chart VII, supplemental remarks.

93. *See id.*

94. That is, if we accept the expert panelists' assessment of passing quality on each question and use that assessment to compute the percentage of applicants who wrote passing answers to each question, and

simulation mirrors the actual distribution of passing essays among applicants who took Minnesota's July 1997 exam, then 88% of those applicants likewise should have passed that exam—not the 78% that Klein estimated. The estimated percentage of applicants who deserved to pass the exam is critical in Klein's method because he used that percentage to choose a new passing score for the exam. Using his 78% estimate, Klein recommended a passing score of 271.[95] A different method of estimating the passing rate would have yielded a different passing score. The 88% estimate derived from our simulation would have generated a *lower* passing score than the one Klein derived from his 78% estimate.

Varying the distribution of passing essays, as well as the number of essays a candidate can fail while demonstrating minimum competence, would lead to different estimates of the "correct" passing rate on Minnesota's July 1997 bar—and, hence, of the passing score for future editions of that bar. Our simulation emphatically is *not* intended to produce a passing score for Minnesota or any other bar exam. Instead, the simulation shows that information about the *distribution* of passing essays, as well as an expert judgment about the *number of failing essays* a candidate may write while still showing minimal competence, are essential to calculating a bar exam passing score from panelist studies like those Klein conducted.

If the bar measured only one dimension of legal competence, or if every question measured all dimensions of that competence, then Klein's method (using expert judgments to set a passing score for each question and then averaging the percentage of examinees achieving that score) would be a reasonable approach to setting an overall passing score. For an exam measuring multiple dimensions of competence with different questions, however, this approach literally averages apples with oranges, bananas, and plums to yield an overall passing rate that lacks a sound basis. To set a passing rate for a multi-dimensional exam using a method analogous to Klein's, one must (1) use expert judgments to distinguish passing answers on each question; (2) employ those judgments to calculate the percentage of passing answers on each question; (3) gather expert judgments about the number of questions a minimally competent candidate must pass; and (4) collect information about the distribution of passing answers among candidates taking the

---

if we make realistic assumptions about both the distribution of passing essays and the number of essays that a candidate must pass to demonstrate minimal competence, then 88% of the examinees in our simulation wrote five or more passing essays. A spreadsheet showing the distribution of essays, which allows one to count the number of applicants passing five or more essays, is available from the authors.

95. *See* Minnesota Study, *supra* note 34, at 9.

exam. Klein performed the first two of these steps, but not the equally essential steps three and four.

* * *

The flaw discussed in this subsection invalidates the recommended passing scores Klein derived from both his exam reader and panelist studies. In addition to this fundamental defect, several other problems tainted Klein's panelist studies. We briefly discuss those flaws in the following subsections. In discussing them, we draw upon a rich literature of standard-setting methodology. Although there is no perfect method for setting cut scores, experts in that field have identified important safeguards. Klein, unfortunately, omitted several of those protections from his panelist studies.

2. Number of Expert Panelists

When expert panels attempt to classify exam performances into passing and failing categories, the number of judges making each classification decision must be "large enough to achieve an acceptably small standard error of measurement for the resulting passing score."[96] A recent exercise to set standards for statewide student achievement tests, for example, used twelve to eighteen judges on each expert panel.[97] By these standards, the number of judges used to review bar essays in Florida, Minnesota, and Ohio was woefully inadequate. In Ohio, between two and four experts participated on each panel, with four-fifths of the panels having only two or three experts.[98] In Florida and Minnesota, the number of panelists was only slightly higher, with four or five experts on each panel.[99]

In defending his studies, Klein has stressed the total number of experts participating in all the panelist sessions in each state.[100] Focusing on the total number of experts across all panels, however, rests on the

---

96. Cohen, Kane & Crooks, *supra* note 35, at 351. *See also* Gregory J. Cizek, *Setting Passing Scores*, EDUC. MEASUREMENT: ISSUES & PRACTICE 20, 22 (Summer 1996) ("The larger the panel, (usually) the smaller the standard error of the mean recommended standard. . . . Utilize as many participants as practical, given available resources.").

97. *See* Cohen, Kane & Crooks, *supra* note 35, at 351.

98. *See* Ohio Study, *supra* note 34, at 8.

99. *See* Florida Study, *supra* note 34, at 8; Minnesota Study, *supra* note 34, at 7.

100. *See, e.g.*, Reply to Comments, *supra* note 9, app. at 1, 2 (letter from Stephen Klein to Thomas A. Pobjecky, General Counsel, Florida Board of Bar Examiners (May 8, 2000) ("[A]ll told, there were 52 Florida lawyers involved in the standard setting activities."); *id.* at 3-4 (suggesting that "the combination of three dozen panelists and readers" should have been "sufficient").

same erroneous assumption discussed above: that each essay question draws upon the same skills and knowledge. If all essay questions tapped the same dimensions of competence, then the total number of panelists involved in Klein's standard setting might be reassuring. As demonstrated above, however, essay questions reflect different aspects of competence. Under those circumstances, employing only two to five experts to judge minimal competence for each dimension is inadequate.

Given the importance of setting passing scores for the bar exam, especially with respect to the anti-competitive impact and the effect on the admission of minority lawyers, Klein should have employed more experts on each panel. Enlarging the number of panelists, as states have done when setting other important cut scores, would have increased the reliability of judgments based on those panels.

3. Training Judges

Specialists in standard setting agree that "[t]he training of judges is a critical aspect of any standard-setting method."[101] In particular, judges must develop a realistic sense of what examinees can achieve on the test. To accomplish this, in many standard-setting exercises judges themselves take the test (or the portion of it they are assessing).[102] Other standard-setting experts give their panelists data about the actual performance of examinees on the exam being rated.[103] This information helps "reduce the possibility that totally unreasonable standards will be recommended."[104]

Standard-setting experts thus recognize that expert panelists tend to set cut scores too high and that training can combat this tendency.[105] This type of training may have been especially important for the law

---

101. Cohen, *supra* note 35, at 352 (citation omitted); *see also* Cizek, *supra* note 96, at 22 ("[A] second key beginning question in all standard setting is how to train participants so that they acquire common conceptualizations of . . . critical reference points.").

102. *See* John Christian Busch & Richard M. Jaeger, *Influence of Type of Judge, Normative Information, and Discussion on Standards Recommended for the National Teacher Examinations*, 27 J. EDUC. MEASUREMENT 145, 151 (1990); Cizek, *supra* note 96, at 22 ("Frequently, standard-setting participants are administered (and receive scores on) a form of the examination that will be used to make the certification, licensure, or mastery decisions.").

103. *See* Ronald A. Berk, *A Consumer's Guide to Setting Performance Standards on Criterion-Referenced Tests*, 56 REV. OF EDUC. RES. 137, 166 (1986).

104. Busch & Jaeger, *supra* note 102, at 147; *see also* Lorrie A. Shepard et al., *Setting Performance Standards for Student Achievement: A Report of the National Academy of Education Panel on the Evaluation of the NAEP Trial State Assessment* 10 (1993) ("The introduction of data on examinee performance provides a reality check that is designed to keep the passing score within reasonable bounds.").

105. Such training may also reduce variation in the standards recommended by judges. *See* Busch & Jaeger, *supra* note 102, at 147, 153-58. *But see* John J. Norcini et al., *The Effect of Various Factors on Standard Setting*, 25 J. EDUC. MEASUREMENT 57 (1988).

professors who participated in Klein's expert panels.[106] Professors are quite familiar with the process of grading essay exams, but the essays they usually grade are produced under very different conditions than bar exam essays. Students write law school essays immediately after completing an intensive study of the subject. Law school exams may also allow more time for essay answers than the thirty minutes allocated by states like Ohio for bar exam essays. Given these differences, professors might approach bar exam essays with unrealistic expectations. Training would address these misconceptions.[107]

Compared to other standard-setting exercises, however, Klein provided relatively little training to his experts. Although panelists discussed the criteria they believed would be appropriate for each question, they did so with little guidance. They did not have to answer the essay questions themselves, which might have informed their understanding of what performance was possible under bar exam conditions.

Panelists did receive model answers for questions in at least one of Klein's studies,[108] but these models may have further aggravated a tendency to apply unrealistically high standards to candidates' answers. In reading a model answer, panelists would have seen a very high quality answer—one that, if drafted by the examiners, may even have surpassed the best answer written by any candidate. Panelists did not receive any examples of less proficient, but still competent, answers.[109]

---

106. In Ohio, a specific "effort was made to include at least . . . one law professor on each team." Ohio Study, *supra* note 34, at 3. The panelists in Florida and Minnesota likewise included several law professors. *See* Florida Study, *supra* note 34, at 7; Minnesota Study, *supra* note 34, at 6.

107. To check whether law professors (or any other occupational group that participated in the panel sessions) applied higher standards than other panelists, Klein could have compared the average scores assigned by professors to the average scores assigned by other groups. Indeed, because judges with different backgrounds may have very different concepts of competence, experts in standard setting have "urged the examination of differences between the average test standards recommended by different types of judges as a matter of course." Busch & Jaeger, *supra* note 102, at 146; Lorrie A. Shepard, *Standard Setting Issues and Methods*, 4 APPLIED PSYCH. MEASUREMENT 447, 454 (1980).

108. *See* Reply to Comments, *supra* note 9, app. at 9 (reproducing directions given to Florida panelists, which referred to a "model answer" given to the panelists). We have not seen panelist instructions for the other states in which Klein conducted his studies, and his reports do not indicate whether the panelists received model answers for the questions they judged.

109. Klein's instructions to the Florida panelists included some instructions that encouraged panelists to be realistic about the quality of bar exam answers. He pointed out, for example, that "applicants are writing their answers in a high stakes (and therefore stressful) situation," that "applicants do not have a lot of time to reflect on their answers" and "cannot consult with colleagues or conduct legal research," and that examinees "are not expected to be experts in the subject covered by your question." Reply to Comments, *supra* note 9, app. at 9 (reprinting Florida directions).

These moderating comments, however, were followed by concluding remarks that would have encouraged a very strict bar passage standard. "You may also want to keep in mind," Klein told the panelists:

As part of any expert panelist's training, finally, it is important to give panel members an opportunity to discuss their ratings after every batch of ten or so papers.[110] These discussions are essential "to keep the raters focused on using the performance scale . . . as the basis for their ratings," rather than "laps[ing] into a norm-referenced grading system, in which the better papers simply get high ratings and the worse papers get lower ratings."[111] Raters, in other words, find it difficult to focus on the task of rendering absolute judgments of quality. Periodic discussions of their ratings help them concentrate on that task.[112] These periodic discussions occurred on Ohio's expert panels, but it is not clear that they took place in Florida or Minnesota.[113] If they did not, there is a real danger that the Florida and Minnesota panelists lapsed naturally into comparative, rather than absolute, judgments. That tendency, in turn, could have inflated the number of essays they marked as failing.

4. Averaging Values from an Ordinal Scale

In Klein's studies, panelists rated essay answers on an ordinal scale from clear fail (coded 1) to clear pass (coded 4). These four scores rank order answers, but the differences between adjacent scores do not necessarily reflect uniform differences in performance. Panelists most likely perceived the gap between a "clearly failing" exam (which rated 1) and a "marginally failing" exam (rated 2) as greater than the gap between the "marginally failing" exam and a "marginally passing" exam (rated 3). Similarly, the difference between a "marginally passing" and "clearly passing" exam (rated 4) probably was larger than the gap between the two varieties of "marginal" exams. Because of this kind of problem, statisticians agree that it is inappropriate to average ordinal scores. One can use a median or a mode to measure "central tendency," but not an arithmetic mean.[114] Thus, averaging the ratings

---

> that the bar exam is used to license practitioners. Applicants who fail can take the bar exam again, but those who pass can practice. Moreover, passing applicants can practice on their own. No further supervision is required. This is analogous to the distinction between licensing someone to be a pilot rather than a copilot.

Id.

110. *See* Cohen, *supra* note 35, at 355; *see also* Busch & Jaeger, *supra* note 102, at 151-52, 160.

111. Cohen, *supra* note 35, at 355.

112. Klein has recognized the difficulty of making these absolute judgments. *See* Klein, *supra* note 13, at 40 ("Judgments about how the quality of a given answer compares to the quality of other answers written to the same question usually can be made faster and more reliably than evaluations of the extent to which an answer's quality falls above or below some theoretical pass/fail line.").

113. *See* Ohio Study, *supra* note 34, at 3; Florida Study, *supra* note 34, at 2-3; Minnesota Study, *supra* note 34, at 2.

114. *See* IVY LEE & MINAKO MAYKOVICH, STATISTICS: A TOOL FOR UNDERSTANDING SOCIETY

that panelists gave each essay answer can provide a distorted and unacceptable estimate of average performance on an essay.

It is difficult to predict what effect this error had on Klein's calculations, without knowing both the distribution of panelist ratings and the manner in which the panelists conceptualized the ordinal scale. Given the centrality of these averages to Klein's process, however, his use of arithmetic means is troubling. These averages generated cut points for each essay question, which in turn produced the passing rates for those questions. Because the averages may have given distorted estimates of the merit of individual answers, the cut points and passing rates are questionable.

IV. IMPACT ON MINORITY BAR APPLICANTS

Artificially high bar passage standards are of special concern because those standards can have a disproportionate impact on minority applicants to the bar. Several studies have documented lower bar passage rates among minority applicants than white ones.[115] In some studies, racial differences remain even after controlling for LSAT scores and law school grades.[116] Examinations like the bar, therefore, seem to impose special obstacles for minority members.[117]

Under these circumstances, increasing the score needed to pass the bar raises three related concerns. First, even if the change itself does not have a disproportionate impact (*i.e.*, even if the percentage of minority members among those who fail the bar remains constant after the change), it extends a known discrepancy. Bar examiners know that the percentage of minority applicants currently failing the bar exceeds that for white applicants. At best, raising the passing score will maintain that discrepancy while increasing the number of both white and minority applicants who will fail.

Second, raising passing scores will raise the percentage of minority applicants failing the bar to disturbing levels. Klein, for example, has estimated that only 52% of minority applicants will pass the Florida bar if the Florida Supreme Court adopts the new passing score

---

111 (1995); *see also* DAVID S. MOORE, STATISTICS: CONCEPTS AND CONTROVERSIES 177 (4th ed. 1997) ("usual arithmetic is not meaningful" for an ordinal scale).

115. *See supra* note 7.

116. *See* Wightman, *supra* note 7, at 12.

117. A growing body of literature explores why these exams pose such special threats. *See, e.g.*, C.M. Steele & J. Aronson, *Stereotype Threat and the Intellectual Test Performance of African Americans*, 69 J. PERSONALITY & SOCIAL PSYCH. 797 (1995).

recommended by the Board of Bar Examiners.[118] Almost half of all minority applicants, in other words, will fail the Florida bar exam under the new standard. This failure rate is almost certain to chill minority applications to law school, as well as to restrict substantially the number of minority members entering the profession.

This impact on minority applicants is particularly troublesome when one considers that many states have maintained constant passing scores since the 1980s.[119] As explained in section one above, those passing scores are equated and scaled to represent a constant level of competence. If those scores were sufficient to distinguish competent and incompetent lawyers in the early 1980s, when most applicants to the bar were white, why is the score inadequate to make that distinction at a time when increasing numbers of applicants are minorities? If some of today's minority applicants are unqualified, they (along with unqualified white applicants) will fail the standards set in the 1980s. Unless states can point to the type of evidence discussed in section II above, it is not clear why today's bar applicants should have to pass a higher standard of competence than one used when the overwhelming majority of applicants was white.

Finally, there is substantial reason to fear that raising bar passing scores will, in fact, have a disproportionate impact on minority members. In general, increased passing scores on the bar exam affect minority applicants more than white ones. In other words, the gap in passing rates between minority and white applicants is likely to grow as passing scores go up and passing rates fall. As Klein himself has recognized, "[t]he size of the difference in bar passage rates between whites and minority applicants depends on several factors," and one of these factors is "the relative stringency of the state's pass/fail standard."[120] In particular, "[s]tates that have relatively high passing rates tend to have smaller differences among groups than other states (because all groups have high rates when standards are low)."[121] As a general matter, therefore, raising the bar passing score (and decreasing

---

118. *See* Letter from Stephen P. Klein, to Kathryn E. Ressel, Deputy Executive Director, Florida Board of Bar Examiners 1 (June 9, 1994) (on file with the *University of Cincinnati Law Review*). This projection also assumed that Florida would eliminate its "banking policy" that allows test takers to combine scores from two different administrations of the exam. The banking policy probably played a minimal role in the passing rates Klein projected in this letter. If the contemplated change in that policy played any role, however, it probably increased passing rates slightly. As Klein has pointed out elsewhere, banking rules actually "tend to lower rather than increase a repeater's chances of passing." Klein, *supra* note 13, at 42. Retaining the banking policy, therefore, might depress the passing rates of minority applicants below the level reported in Klein's letter.

119. Florida's passing score, for example, has remained the same since 1982.

120. Klein & Bolus, *supra* note 7, at 8.

121. *Id.*

the passing rate) is likely to *increase* the gap between whites' and minorities' success rates.

Historical records of bar exam scores in individual states could provide evidence contrary to this general trend. Most states, however, apparently do not record individual bar scores by race or ethnicity. We are aware of no evidence in Ohio or Minnesota on this matter. In Florida, the only evidence came from a single year of bar examinations and is almost a decade old.[122] In the absence of clear evidence that no disproportionate impact will occur, the general trends cited above suggest that states should assume such an impact. Even without such a disproportionate impact, of course, an increase in the passing score will extend existing discrepancies and force today's minority applicants to surmount a higher hurdle than one in place when almost all bar applicants were white. At a time when the profession has embraced the need for diversity, this result is contrary to public policy.

CONCLUSION

The legal profession, like other professions, protects the public by limiting the right to practice to persons who have demonstrated that they are competent to do so. But in denying licenses to some would-be practitioners, state bar examination boards have a responsibility to set standards at a realistic level—one that will not restrict the supply of qualified practitioners, drive up the cost of legal services, or disproportionately deny people of color the right to practice the law.

Bar examinations in most states serve as stable screening devices. Through equator items and scaling transformations, they are impervious to changes in exam difficulty or grading practices over time. Because of these built-in protections, there is no *a priori* need to recalibrate passing standards. Nor is there any evidence that state bar examinations currently admit incompetent practitioners. Equated exam scores have been higher since 1992 than in the years before that date, suggesting that recent bar applicants are *more* qualified than their predecessors, and states have pointed to no evidence that these licensed practitioners are incompetent. The most common justification offered for raising bar passage scores is to bring states with lower passing marks into line with states maintaining higher hurdles. This reflexive reasoning may restrain

122. *See* Letter from Stephen Klein to Kathryn E. Ressel, *supra* note 118, at 1. That evidence suggested that raising Florida's passing score would not disproportionately affect minority applicants; that is, it would not increase the gap between white and minority test takers. The evidence, however, confirmed the existence of a sizable gap in passing rates between those two groups. It is difficult to put much faith in the lack of disparate impact, moreover, given the dated nature of the data.

competition while placing a particular burden on minorities who wish to practice law. Recent decisions in more than a dozen states to raise bar passing scores thus lacked sufficient evidence that change was needed.

Equally important, one particular process used to produce new passing scores rests on invalid assumptions and employs flawed procedures. By incorrectly equating the percentage of failing test takers to the percentage of failing essays, the process generates an unreliable standard. Evidence about the distribution of actual bar exam scores suggests that the method usually produces an unduly high standard, one that will fail examinees who are fully qualified to practice law. These inflated standards, in turn, may have unintended effects such as a substantial adverse impact on the admission of minority applicants to the bar.

Psychometric methods used to set passing scores are complex, employing advanced statistical techniques and assumptions that are not always readily visible. Bar examiners and judges who rely upon consultants to apply these methods may have difficulty detecting flaws in the processes. Using an objective, controlled method to set passing scores is commendable, but the process must reflect valid assumptions about bar exams. Lawyers and judges who rely upon these methods must probe the processes and their underlying assumptions. We hope that, with greater understanding of Klein's method and its consequences, states will reconsider recent increases to bar passing scores and devise sounder methods for reviewing their passing scores.