Report of the

Committee on Legal Education and Admission to the Bar of the Association of the Bar of the City of New York in Opposition to the Board of Law Examiners' Proposal to Increase the Passing Score on the New York Bar Examination

January 2003

Table of Contents

Introduction

The Committee on Legal Education and Admission to the Bar of the Association of the Bar of the City of New York submits this Report in opposition to the Report and Recommendation of the New York State Board of Law Examiners to the Court of Appeals Regarding the Passing Standard on the New York State Bar Examination (the "Report").

At the outset we emphasize that we join the Board in recognizing that any change in New York's standard for licensure should be rooted in "scientifically based studies informing our policy judgments" (Report at 13). It is *because* we agree with the Board that there should be both careful scientific analysis and a full consideration of competing policy concerns that this Committee urges that no change be made in the passing score until *both* its effectiveness in meeting the goals of the Board *and* its ramifications for those seeking to become lawyers and those seeking legal services have been explored in scientific analyses that go far beyond what may have been learned from the Klein Study discussed in the Report. We believe that the Board's proposal is likely to have a disparate impact on minority candidates, reduce the availability of legal representation to already underserved persons and discriminate among candidates on the basis of financial means -- all without a sound basis for expecting that the proposed change will be useful in screening candidates for entry-level competence to practice law.

Accordingly we submit this report: (i) to express the Committee's opposition to the Board's proposal to increase the passing score on the New York State Bar Examination from 660 to 675, (ii) to comment on the research conducted by Stephen P. Klein, Ph.D., that is extensively relied upon in the Report, (iii) to suggest additional research that should be conducted as part of any review of the passing score, and (iv) to discuss critical policy considerations relating to the proposal that, in our view, go to the heart of the Board's gate-keeping function.

In addition, we join with the fifteen New York State Law School Deans in requesting that public hearings be held on this most important matter.[1] We note that both Florida and Minnesota, the two states most recently to consider a substantially similar proposal, held public hearings enabling interested parties to testify and present evidence. Given the unquestionable public significance of a proposal affecting access to the legal profession and the availability of legal services, it is certainly appropriate for the Board to give members of the public, including representatives of minority organizations, representatives of the organized bar, deans and law professors, an opportunity to express their views.

[1] Although we understand that the Board recently announced its intent to hold hearings on February 7, 12 & 14, 2003, we believe that at this point it would be most effective if the Court of Appeals was directly involved in the hearing process.

The Relevant Standard: Entry-Level Competence

The Board has stated that the "passing standard must accurately reflect the minimum level of competent performance required for admission into the profession" and that the passing score should be both "representative of current norms for minimum competence" and "calibrated to the level which affords the public protection." (Report at 1.)

Each of these is a laudable, important goal. This Committee has previously expressed its view that the bar examination tests only a few of the many skills and competencies new lawyers should possess in order to competently practice law.[2] We repeat our concern that the current examination addresses only some of those "competencies" and tests them in an artificial setting and manner unlike the actual practice of law. Moreover, while there has been considerable writing and debate on the multiple competencies involved in fulfilling lawyering tasks in a competent manner, there is no independent consensus among practitioners as to the substance of "minimum competence" as measured by the current bar examination. We note that the Board did not seek such a consensus from the panel of practitioners and teachers it assembled as part of the research on which its Report is premised.

Nonetheless we recognize that the Board must perform its duties in the examination of potential lawyers in a context of both limited time and limited resources, and we appreciate that the Board is conducting its current review of the passing score with the goals of testing competence and protecting the public. However, we believe that there has been no showing, in the Report or otherwise, that the current passing score has resulted in the licensing of incompetent lawyers or that setting a higher passing score will protect the public from incompetent lawyers.[3]

We agree with the Board that the precise passing score will always have some element of the arbitrary and that New York's passing score of 660 was not set after any scientific study to measure "minimum competence." (Report at 2.) However as the extensive study commissioned by the Court of Appeals commented in 1993, the 660 passing score does have a "rationale" in that it is based on a historical view taken by Boards of Law Examiners regarding the minimum level of performance for a qualified candidate.[4] When the MBE was added to New York's examination, the pre-1979 passing rate was the basis for setting the passing score at 660. Prior to 1979 the "primary

[2] Report on Admission to the Bar in New York in the Twenty First Century – A Blueprint for Reform, 47 THE RECORD OF THE BAR ASSOCIATION OF THE CITY OF NEW YORK 464 (1992) ("Blueprint").

[3] We do agree with the Board that the incidence of disciplinary complaints about competence or malpractice claims is an insufficient measure of whether particular scores measure competence. It might still be useful to know whether bar examination scores correlate with disciplinary proceedings and malpractice claims.

[4] Millman, Mehrens & Sackett, AN EVALUATION OF THE NEW YORK STATE BAR EXAMINATION (1993) ("Millman Evaluation") at 8-1. The Board confirms that in 1979 it made the assumption that the then-existing passing rate was one practical measure for discrimination based on competence (Report at 2).

criterion was candidate performance relative to the content of the questions rather than adherence to a group standard consisting of a fixed percentage of passing candidates."[5]

But saying that *these* measures are imperfect, or that the existing baseline was based on long experience rather than scientific study, does not support the conclusion, *ipse dixit*, that a candidate who scores 675 (MBE 135) is "more competent" to practice law than one scoring 660 (MBE 132), or that the former is "minimally competent" but the latter is not. To make such a judgment, particularly when those scoring 660 (MBE 132) are *currently* passed as minimally competent, requires careful analysis of the score scale, the characteristics of those candidates currently scoring between 660 and 675 (or 685, the Board's longer term target), the explicit and implicit assumptions behind the definition of "minimum competence" the Board chooses to employ, and the connection between differences in test performance and competence in practice.

In these regards we would emphasize that the extensive study commissioned by the Court of Appeals a decade ago recognized that:

> a licensure examination is not an employment examination, nor have any inferences about degrees of success for those who score above the cut score been validated. Thus, there is no evidence that would justify an employer selecting among applicants based on how far they scored above the cut score.[6]

We agree with the Board on "the importance of engaging in standard setting exercises in determining passing standards, rather than relying on assumptions" (Report at 16). We suggest, however, that "standard setting" means more than adjusting the passing score on the existing examination. Rather, a proposed change in the examination passing score must be evaluated in terms of whether it will better screen for minimal competence or simply increase the short-term pressures of exam preparation. That requires express debate, rather than implicit methodological assumptions, about the content of "minimum competence" and careful analysis and thorough discussion of how the examination may measure such competence (or competencies) and most effectively fulfill its screening function.

Deciding what "level of knowledge and skills [is] deemed appropriate for the entry level practice of law" is a judgment rooted in policy -- policies about lawyers, the populations they serve and the structures through which such services are delivered. It appears that the proposal to increase the passing score reflects more than anything else a concern that New York not be perceived as having standards lower than those employed by other states. We believe that New York's exam is among the most rigorous in the nation. Indeed, some commentators have deemed it to be the most difficult. When

[5] Millman Evaluation at 8-1.

[6] Millman Evaluation at 13-1.

comparing New York's passing score to those adopted in other jurisdictions, the Board and the Court of Appeals should bear in mind New York's extensive testing of mastery of distinctive characteristics of New York law and practice in addition to the testing of "multi-state" substantive law. However one may assess the value of an emphasis on New York law, it seems clear that this already requires a level of post-law school preparation (measured in time and money) not necessary in many other jurisdictions. New York should proudly stand by its commitment to a competent, professional Bar that reflects the great diversity of the State and strives to be available to all who need legal services, rather than act on an apparent desire to compete with other states in terms of whose MBE-driven passing score is the highest.

In making the points which follow we are not arguing against setting standards for admission into the profession. We do not think for a moment that there should not be "any standard at all" (compare Report at 18). Nor are we arguing that an examination should have no role in admission to the Bar. But whether a test is "reliable" in terms of its consistency of administration does not bear on the utility of its scaled score as a <u>relative</u> measure of professional competence, and it is not possible to assess the validity of increments in the scaled score without expressed consensus both on what "minimum competence" *means* and on what external measures of such competence for future practice may be used to assess the validity of the examination and its scale.

<u>Impact on Candidates, Racial Diversity, and the Availability of Legal Services</u>

We are concerned that the proposed increase (i) is likely to have a disproportionate impact on minorities, both candidates for admission and those in need of legal services, and (ii) will undoubtedly increase the time and expense involved in preparation for the examination. (The latter point is addressed in the next Section.) While the Report states that "the Board is advised that an increase in the passing score should have no measurable effect on existing differences among minority and non-minority groups in passing rates on the New York bar exam" (Report at 19-20), it provides neither source nor substance for this "advice." Moreover, the Board acknowledges that it has not previously collected demographic information, considering such "obviously irrelevant to the pass/fail decision the Board is called upon to make" (Report at 14).[7]

It is one thing to admit that because information "irrelevant to the pass/fail decision" has not been collected, the potential for disparate impact cannot be assessed from bar examination data. It is simply erroneous to transform that admission of ignorance into "advice" that there *will be* no such impact.

The LSAC National Longitudinal Study starkly illustrates the disparities in passing rates among different racial groups of a cohort of 1991 law school enrollees

[7] The Report admits, "Evidence that would support or refute that proposition [*i.e.*, the potential for disparate impact on minorities] has not been collected." (Report at 14). In fact, the Board *does* have access to data on July 1992 candidate ethnicity from the Millman Evaluation and hence could make a preliminary effort, using such data, to assess the potential disparate impact of its proposal. See Millman Evaluation Chapter 10.

(most of whom would have taken the bar examination for the first time in 1994).[8] The LSAC Study reported eventual passing rates of 77.63% for Blacks and 96.68% for Whites. That Study also established that Black candidates who initially failed the bar examination were somewhat more likely to abandon the process, and not try again.[9] The Millman Evaluation, using New York's July 1992 examination, reported a difference between the mean score for Blacks and Whites of 80 points.[10] The difference in July 1992 New York passing rates was also enormous: 81.6% for Whites, 37.4% for Blacks.[11]

The Report responds to this concern by suggesting that if the proposal to increase the passing score is adopted, a study be commissioned to gather the relevant demographic information in order to assess the impact of the increased score (Report at 14). But surely such a study should be performed before -- and not after -- the proposed increase goes into effect. It would be a relatively simple matter to collect the demographic data during the next several administrations of the exam and then determine precisely the effect on different groups of increasing the passing score from 660 to 675 (or eventually to 685 as the Board contemplates).

The Committee has learned that Florida collected demographic information in connection with the two administrations of its 2000 bar examination. The data clearly showed that there would have been a disparate impact on minority candidates if the passing score had been raised two or five points, because minority candidates disproportionately scored in these ranges.[12] A passing score increase of 2 points would have failed 4.56% of whites who otherwise passed; a five point increase would have failed 12.33% of whites who otherwise passed. By contrast, 7.57% of minority

[8] Wightman, *LSAC National Longitudinal Bar Passage Study* (1998).

[9] Wightman, *LSAC National Longitudinal Bar Passage Study* (1998). The LSAC study reported large differences in attrition rates for examination candidates based on race. *But see* Strickland, *The Persistence Facts*, AALS NEWSLETTER, Nov. 2000, Page 5, recalculating the LSAC data with respect to attrition to use those initially not passing as the denominator (rather than the pool of candidates). The result (dropouts as a percentage of those initially failing) was 28% for African Americans, 24% for whites and *lower* percentages for Asian Americans and Latino candidates.

[10] Millman Evaluation, Table 10.1 at page 10.4 (regression analysis of a mail-in questionnaire supplies a somewhat narrower but still broad "parameter," *id*. at 9-11). The classification labels are those used in the report.

[11] *Id*. at 10-4.

[12] Minority candidates were about 24.7% of total July 2000 candidates and 21.2% of passing candidates. They were 26.9% of those scoring 131-135. This data was submitted to the Florida Supreme Court in matter SC96869 (which was the Court's review of the proposal to change the Florida passing score) in documents dated July 23, 2001, and August 29, 2001. Again, the classification labels are those used in the particular report. Blacks were 50% of the Minority candidates at the February 2000 sitting and about 32% at the July 2000 sitting. The results for Blacks in analysis of the February pass rates would be more extreme than that for Minorities as a group, but the results at the larger July sitting were statistically equivalent.

candidates who passed would have failed if the passing score had been raised two points and 17.1% would have failed if the passing score had been set five points higher.[13]

Our concern about the impact of the proposal is not limited to its effects on candidates; it extends as well to the people who would be deprived of their legal services. There is good reason to expect that the candidates who will not be admitted because of the new passing score would have been more likely to join small firms or enter solo practice, providing legal services to working class and middle class individuals and families. There is evidence that minority lawyers are more likely than other lawyers to provide services to the poor and other underserved groups. With courts increasingly strained by the number of unrepresented litigants, and the pervasive and complex presence of law in the lives of all of us, responsible public policy requires that the supply of legal services to underserved populations not be unnecessarily adversely affected. Further, our concern extends to the negative impact on public trust and confidence in the legal system when those in the courts and the legal profession do not reflect the makeup of the population and also may not be sensitive to issues and concerns of minority groups.

Given New York's longstanding commitments to increasing the diversity of the Bar and to affording legal services to diverse communities, the proposal to increase the passing score should not be implemented in the absence of data *demonstrating* the relative utility of such an increase and assessing its effect on racial minorities and access to justice. New York's law schools have begun to make noticeable progress in their quest for diversity in admissions. A proposal that runs the risk of disproportionately impacting racial minorities in access to the profession runs the concomitant risk of deterring minorities from even seeking admission to law school.

Research Predicates

We doubt that the "standard setting" study conducted for the Board by Dr. Klein ("Klein Study") demonstrates anything more than the views of the study panel members about the particular essay each panel member reviewed. We note that it is the Board's stated policy that there "is no passing or failing on any one portion of the examination. . . . Passing or failing is determined only on the basis of the applicant's total weighted scaled score." Certainly the Klein study provides no statistically sound basis for extrapolating from views about "passing" one essay question to "passing" the essay portion of the examination (weighted at 40%), let alone a basis for setting the appropriate minimum passing score on the examination as a whole. Our review of the Klein Study persuades us that its resulting recommendation is an artifact of Dr. Klein's methodology which makes assumptions about "minimum competence" without even attempting to define its meaning. As noted below, the expert panel was not even asked to discuss the criteria of "minimum competence."

[13] The overall July 2000 pass rates at the existing cut offs were 83% (Whites) and 68% (Minority). Because a substantially lower percentage of minority candidates pass, drops in absolute numbers of candidates (e.g., 4.56% versus 5.14% of all candidates upon a two point shift and 10.2% versus 11.63% at a five point shift) have a larger proportionate impact on the cohorts passing the test.

In fact, the Klein Study calls into question the existing practice of scaling the New York components of the examination to the MBE. Research from data available to the Board could address this, and further study of the data collected might also provide the Board with insight into how its guide for grading individual questions compares to practitioners' assessments of competence. We believe that the Board should systematically address both of these topics before considering the twice-removed question of the appropriate *overall* passing score.

We respectfully suggest that the policy considerations that underlie any decision to raise the passing score *require* that the Board provide a persuasive scientific analysis of the benefit of a change in the score. Because the acknowledged purpose of the examination is to test "minimum competence" for protection of the public, we submit that the Board's scientific review should focus on: (1) developing a substantive standard of "minimum competence" insofar as this involves skills to be assessed through this examination or any proposed alternatives; (2) how any change in passing score will have a meaningful impact on screening for competence as a practicing lawyer; (3) the population of candidates taking the New York examination and the demographic characteristics of the segment of that population most likely to be affected by the proposed change (*i.e.*, those currently scoring between 660 and 685); and (4) the value of requiring law school graduates to devote even more time and money to preparation for the examination, a factor that necessarily discriminates on the basis of means. We believe that research in each of these focal areas would provide the Board with a wealth of information useful not only in assessing the passing score but also in designing the examination and assessing its function as a "standard for licensure."

The Board hypothesizes that "further preparation" may blunt the effect upon the passing rate of raising the passing score (Report at 10, 13). Others argue both that further preparation would make little or no difference in scores for the candidate pool as a whole and that it is unlikely there would be changes in overall preparation, which we believe is already substantial. Solid research based on several years' experience in the states which have recently raised their passing scores may provide partial answers – although the New York examination's heavy emphasis on local law undoubtedly already prompts greater preparation for this examination than is the case in some other states. Even if such additional preparation were to occur in New York as a consequence of a change in the passing score here, it will still be necessary to ask why incremental time spent on examination preparation will benefit the profession or the public and whether the necessity for still-additional post-graduation preparation for an examination would further discriminate among candidates based on financial means.

Although the Board acknowledges that "most candidates for admission eventually pass the bar exam," it expresses its conviction that those who pass on a later attempt "are, at that point, better prepared to enter the profession." (Report at 18.) This Committee noted its disagreement with this assertion some years ago, suggesting it was far more likely that for the later examination, candidates have "improved the narrow range of test-taking skills necessary to pass" without gaining increased competence for practice.[14] Before policy changes are made based on the Board's assertions, follow up studies of those now passing the examination on later attempts and an analysis of such candidates' initial preparation, initial score, subsequent preparation and subsequent score should be undertaken to test scientifically the validity of the Board's assertions. The Board also has the means to conduct follow-up studies about candidates who in past years have scored between 660 and 675 (or 685) on their first attempt to pass the examination. However, we should not lose sight of the fact that the goal must be better preparation for practicing law rather than better preparation for passing an exam.

According to the Board's analysis of July examination results for 1999-2001, use of the 675 cut-off, absent changes in candidate preparation that resulted in higher scores, would have resulted in between 5.4% and 8.5% of candidates failing rather than passing. Some 12,709 candidates took the Bar examination in 2001 (9,194 in July). The proposed change, then, would have resulted in between 600 and 1000 additional candidates failing in that year. If the net effect of the proposed change is merely to require still-additional preparation before the initial test, then scientific study must inform decision-makers as to the true benefits of this additional cost. If in fact the result of such a change is a decline in the initial pass rate, followed by subsequent successful completion, policy makers must consider on which candidates this burden falls and what the costs and benefits are of deferring admission of these candidates.

The Board's Report simply does not establish that those who currently score between 660 and 675 lack the "knowledge and skills deemed appropriate for the entry level practice of law." In what regard does the Board find those candidates lacking? Since we surely do know that raising the passing score will have an impact on the lives of candidates, what measurable benefit will be conferred on the profession or the public?

Comments on Methodology

We appreciate that involving panels of attorneys in a "standard setting exercise" like that conducted by Dr. Klein may provide interesting information about the portion of the examination under study. We do not understand, however, how the results from this exercise can be extrapolated validly into a "passing" grade for the New York essay portion of the examination (were there such a grade, which is not the case), let alone used to set the standard for the examination as a whole.

The Board acknowledges that the methodology followed by Dr. Klein in his study has been seriously challenged when used elsewhere.[15] We appreciate that the Board and Dr. Klein have attempted to address some of these critiques. However, the challenges to Dr. Klein's work are not limited to the kinds of computational alternatives Dr. Klein explores, but go to the fundamental logic of his research protocol.

[14] Blueprint, *supra* note 2 at 477-78.

[15] *See* Merritt, Hargens & Reskin, Raising the Bar: A Social Science Critique of Recent Increases to Passing Scores on the Bar Exam, 69 U. CIN. L.REV. 929 (2001) ("Merritt Article").

Because comments on methodology require a level of expertise in the high-stakes testing area that is not otherwise available to the Committee, we have sought the assistance of a quantitative psychologist in preparing the remarks which follow.[16]

1. From Individual Questions to Overall Essay Assessment

One flaw in the methodology lies in the extrapolation of an ultimate passing rate from averaging the passing rates on individual essay questions. We expressly label this a "flaw" because it represents a methodological tactic which effectively predetermines the result, without first establishing a definition of "minimum competence" that reflects the policy of the Board or even the consensus of the expert panel.

In his study Dr. Klein computes the percentage of applicants who should pass the entire examination by averaging the pass rates for six individual essay questions. More expansively, Dr. Klein's approach was to determine across the six essays analyzed the average percentage of candidates whose performance on an essay fell below the panelists'/readers' opinion of minimum competence for that essay. The overall passing score to be set for the composite of all parts of the examination was then calculated to be the total score that would fail the same percentage of candidates as on average "failed" an essay question. Since Dr. Klein's approach does not rely on an express definition of minimum competence ratified by the Board or the panel, in choosing this methodology, it determines an overall passing score that may or may not satisfy the Board's intention to set the score to screen for minimum competence. Whether a total score that produces an overall passing rate equal to the average of the "passing" rate for essay questions evaluated by the panels *should* constitute the standard for "minimum competence" is a very substantial question which the Board did not discuss explicitly, and one which the study panelists were not asked to address.

That this methodology represents a tactical choice with policy consequences is made clear by an illustration offered by Carol Chomsky[17] in her testimony before the Minnesota Supreme Court:

> Assume I gave an examination consisting of three questions to four students (Alan, Betty, Cindy and Doug) and I decided that they would have to answer two out of three questions correctly to pass. Then assume that the aggregate statistics show that each of the questions was answered correctly 50% of the time. Dr. Klein's

[16] Dr. Jerard Kehoe has a Ph.D. in Quantitative Psychology and has twenty years experience in the development and management of large scale employment testing programs. Among other professional responsibilities, he serves as an Associate Editor of the Journal of Applied Psychology, edited a book on employment testing, and in 2000-2002, served on his professional society's committee to revise its "Principles for the Validation and Use of Employment Selection Procedures".

[17] Past-President of Society of American Law Teachers, Professor of Law, Minnesota Law School, Director, Bush Faculty Development Program on Excellence and Diversity in Teaching. The methodological issue illustrated here is discussed at some length in the Merritt Article at 949-62.

methods would result in the conclusion that 50% of the students should have passed the exam and 50% should have failed the evaluation standard adopted. But consider the following two circumstances:

Case 1: Alan and Betty answered correctly on #1
Betty and Cindy answered correctly on #2
Alan and Cindy answered correctly on #3
Doug answered incorrectly on all 3
In this instance, Alan, Betty and Cindy - 75% of the students - passed.

Case 2: Alan and Betty answered correctly on #1
Alan and Cindy answered correctly on #2
Alan and Doug answered correctly on #3
In this instance, only Alan - 25% of the students - passed.

The above illustration shows just one alternative to Dr. Klein's conclusion that performance on the entire examination can be compared to performance on a single question. To offer a different example drawn from statistics in the Millman Evaluation: for the July 1992 examination approximately 55% of candidates got any particular New York multiple choice question correct. Nonetheless 74.6% passed that examination, and the mean score was 705, far above the 660 cut score. Thus the bulk of the examination cannot be thought of in terms of a "passing score" on particular questions -- performance *must* be measured on an accumulated basis.

The methodology that would have been most consistent with the current approach to scoring the examination would have been to have the panel score the essays and then have panelists review a sufficient sample of *packages* of scored essays (each package constituting a candidate's total essay performance) and evaluate the entire package in terms of whether the candidate had demonstrated minimum competence. This seems more consistent with current practice, which is to use scaled subtotals of scores on various parts of the examination to measure an overall sufficient (on a pass/fail basis) accumulation of knowledge and demonstration of reasoning. Dr. Klein's methodological assumption introduces a different standard, and that assumption drives its results.

From data already available to the Board it should be possible to analyze the overall performance of the test-takers whose essays were graded by the panels and to learn how strongly their scores on the particular questions correlated with the candidate's overall performance, both on the essay portion and on the examination as a whole. How, if at all, did achieving this "passing" score on a particular question *in fact* compare to overall essay performance?[18] How well, if at all, did performance at the margin of the

[18] In light of Dr. Merritt's observations (which are based on statistical analysis of a session of the Minnesota bar examination) we also think it would be valuable to explore the correlation of the quality of applicants' performance on the variety of New York essay questions, particularly for candidates with overall scores in the marginal ranges.

"passing" score on a particular essay *in fact* predict an overall examination score in the 660-685 range?

If the Board were to accept Dr. Klein's rationale for recommending a higher passing score, it would be endorsing a passing score without knowing whether the score is proposed based on the method most consistent with its own policies about the profile of test performance that identifies minimum competence. In such circumstances the Board would be allowing a methodological assumption to *dictate* new policy and would be changing that policy without any expression of consensus within the expert panel, the Board, the Court of Appeals, or the Bar generally that this different policy is appropriate.

We should here note that, in part as a result of questions raised regarding Dr. Klein's methodology (which we continue to discuss below), the states of Florida and Minnesota apparently have deferred implementation of Dr. Klein's proposals.

We are also concerned about the sample of answers reviewed and the use of this sample to set a percentage of answers that "met or exceeded the panelists' standard." The average quality of the essays reviewed by the panel appears to have been considerably above the quality of all of the candidates' essays from that sitting of the 1999 examination as equated through Dr. Klein's procedures. On Question 4, for example, 57.5% of the answers studied received a score of 2.5 (used by Dr. Klein as the floor for minimum competence) or better from its panel, whereas only 46.1% of the full candidate pool answers were estimated to have met this same standard as recalibrated through a "reader score." Assuming the arithmetic of recalibration was correct, it appears to us that the sample studied must have quite substantially failed to represent the population of responses. From what we understand, the use of a representative sample is a procedural standard in research on "high-stakes testing," and the use of a non-representative sample (even if purely by accident) is recognized as a very serious shortcoming.

How the unrepresentative nature of the sample may have affected panel scoring is something we cannot know. We discuss below why the panels' scoring correlations (which measure relative rankings of essays) do not alleviate this concern. Since sampling differences of the size in the above example exceed the effect of the proposed change in score, we respectfully suggest that this portion of the procedure be reviewed by an independent psychometrician with no prior connection to any of the experts employed by the Board.

Our concerns about this aspect of Dr. Klein's methodology are reinforced by comparing the Klein Study to the 1993 Millman Evaluation. The Millman Evaluation included a variety of reviews of the examination by local practitioners, including a somewhat similar assessment of essay answers.[19] While the 1993 evaluation's methodology is open to some of the same criticism as the Klein Study, we are struck by the fact that its panel analysis produced a cut score of 659 -- some 13 years after 660 had

[19] Millman Evaluation, Chapter 8.

been selected based on historical experience. The Millman Evaluation accordingly recommended that there be no change in the passing score at that time.

We recite this not to commend the methodology of the earlier work, but as the occasion to remark that we cannot imagine what changes in the seven years between the July 1992 examination studied by Millman and the July 1999 examination studied by Klein would lead to such different results. Absent a clear and documented explanation, we are left to suppose that differences in panel preparation or panel composition, use of a "pass rate" methodology, or the unrepresentative nature of the sample are all possible explanations.

One critical difference may be that the Millman Evaluation put more emphasis on defining a standard of minimum competence to be applied by its panel. The 1993 evaluation also used only practitioners on its panels (not full time law professors).[20] Millman also focused on setting a minimum competence score for each essay rather than using the two-step approach built on passing rates.

Given the importance of the 1993 Evaluation commissioned by the Court of Appeals, we were surprised to discover that it receives only one passing mention in the Board's Report (and none on this topic), and no mention at all in the Klein Study. This particular aspect of the Millman Evaluation did have an element of informality. Nonetheless the *wide* divergence in results after only seven years warrants a totally independent review of the technical aspects of both studies *and* an explanation of the difference in results reached.

Michael Kane, Director of Research for the National Conference of Bar Examiners (which administers the MBE), has criticized on a number of grounds the study methodology used by Dr. Klein elsewhere and repeated in New York. Dr. Kane has discussed the proper methodology for "examinee-centered standard setting" in a recent article, and we find one set of his comments particularly pertinent. Dr. Kane believes that while it can be valuable to use panels of academics and experienced practitioners in applying "standard setting" to the examination, because of the relative professional maturity of such panel members, they should have familiarity "with the work of newly admitted lawyers . . . to keep the standard realistic."[21] And once the panel is assembled the next step is critical:

> ". . . at the beginning of the standard-setting process, the panelists are expected to reach some level of agreement on a *performance* standard describing the level of competence required for entry-level practice, which is then used to identify an appropriate *passing* score. The initial statement of the performance standard can be refined during the

[20] We do also note that the 1993 exercise was on a smaller scale and reported a margin of error of 18 points because of sample size.

[21] Kane, Clearing the Bar: Setting the Standard, THE BAR EXAMINER (November 2001) 6, 8.

> study, but it is important to start with a clear focus. The performance standards are likely to be most defensible if they are clearly linked to generally accepted standards of practice."[22]

This simply did not happen in Dr. Klein's study. Without in any way disputing the professional qualifications of the panelists, we must point out that the panelists as a group were not asked to reach a consensus on the level of competence required for entry-level practice. Since the panel group included academics as well as current practitioners, the absence of such discussion could only heighten the potential for inconsistency in the scores (as opposed to the relative ranking of essays), which are critical to Dr. Klein's approach.

The five panelists considering a specific question did discuss "appropriate criteria for their question" for a passing grade on that one particular question. However not only was consensus not required, but there is no indication that panelists were focused on "minimum competence" as the standard for "passing." And clearly the group as a whole did not discuss the qualities of "minimum competence." It seems to us that a distinctly better job of preparing the panel to evaluate the essays in terms of minimum competence was done in the Millman Evaluation, which recommended no change in the passing score.

As we understand the literature, for a panel procedure like this to be valid it is also necessary that the scoring panel understand the consequences of the scores being applied. The panelists in Dr. Klein's study were not informed that performance on a particular question would be used as the baseline for constructing a pass/fail measure of competence for the entire two-day test. We suspect that many panelists would at least have questioned, if not rejected outright, this methodology.

Dr. Klein attempts to address this issue by reporting that the panelists' ratings of the same essays were generally highly correlated with each other, and, to a lesser extent, with the readers' results. But these correlations only reflect that the panelists were in high agreement about which essay was best, which was second best and so on. These correlations do not provide any evidence of any agreement among panelists and readers concerning the criteria for the assignment of competency *ratings* or score weights.[23]

One point in gathering practitioners and academics, rather than merely relying on the Board's existing grading team, surely is to elucidate from a consensus of the panelists what "minimum competence" entails. That requires group discussion of the

[22] *Id*. at 8.

[23] If it is also true in New York, as Merritt reports that studies in other states have shown (Merritt Article at 953-54)), that there are relatively low correlations between candidates' scores across essay questions, then Dr. Klein would not be correct in assuming that the work of the various small panels validate each other in some regard.

standard as a predicate to individual grading of sample answers to a single question. Assembly of this group provided a striking opportunity to address the standard critical to the Board's function, the qualities of "minimum competence." That unfortunately did not happen.

2. From Analysis of the Essay Section to Overall Score

Our second group of methodological concerns turns on the relationship between performance on the New York essay section and the examination as a whole. We appreciate that practitioners asked to make judgments about a "passing" score may feel more comfortable evaluating an essay than picking the number of correct answers on the MBE that seems like "enough" to demonstrate "minimum competence." However, it is the Board's present practice to weight the MBE equally with the essays (at 40% each), and to scale each candidate pool's raw scores on the essays *to* the MBE distribution. In this section we address the implications of this policy for Dr. Klein's research and recommendations.

The Millman Evaluation described the essay scoring as follows:

> The raw scores (0 to 10) assigned by each reader are converted to standard scores having a mean of 50 and a standard deviation of 10. Thus, the mean and spread of these standard scores are the same for each reader and for each essay question. The average of a candidate's six essay scores (expressed as standard scores) is computed, and the distribution of these averages is then converted to common-scale values in the same way the NYMC raw scores were converted . . . to common-scale values having the same mean and standard deviation (a measure of variability) as those New York candidates earned on the MBE. For example, if the mean common-scale value for the New York candidates on the MBE is 680, the mean common scale value for these candidates on the [essay] component is also 680.[24]

Because essay performance for the examination candidate pool as a whole is scaled to the MBE distribution, rather than the reverse, there is a logical fallacy in extrapolating from a panel's view of what a passing score might be for the essay portion (even assuming this were the product of a valid research protocol) to an overall passing grade for the examination. As the Board recognizes (Report at 15-16), whether candidates *as a group* perform well or poorly on the New York essays in comparison to standards that might be set by the Examiners or by an outside panel will not change the distribution of *scaled* scores and accordingly will not affect the passing rate. (Of course an individual candidate can affect his or her likelihood of achieving an overall passing

[24] *Id*. at 7-3 and 7-4 (passages combined and order reversed for clarity).

score by better preparation or performance; the point being made here is about the scaling of the full pool's results.)

We appreciate that this analysis of scaling may be both hard to follow and counter-intuitive. The Millman Evaluation acknowledged that the "procedures for arriving at the final examination scores are complicated and not understandable to many people."[25] Scaling to the same mean is used to "force" (i.e., impose) the assumption that each year's NYMC, essays and MBE be regarded as equally difficult in terms of score (an assumption of equivalent rigor) so that higher scores on the essay in a particular year are assumed to be due to an "easier" test (warranting a scaling correction) rather than to an improvement in the candidate pool's competence (which should be recognized, not obliterated by re-scaling).

In other words, if scores increase on the essays, thereby increasing the mean score on that part of the examination, scaling adjusts the essay scores because of the assumption that, to the extent the rise in essay scores was not accompanied by an equivalent rise in the mean MBE score, the increase in the mean essay score must indicate that the essay test has gotten easier, or the grading less rigorous. Scaling the essay scores to the MBE mean necessarily rejects the possibility that candidates have differentially improved their knowledge and skills applicable to the New York essays. Accordingly, the essay scores are deflated, so that the essay mean still matches the MBE mean. Conversely, if candidates as a group were to score higher on the MBE, thereby raising the mean MBE score, essay scores would be scaled higher even if there had been no true change in student performance on the essay portion of the test.

A study of essay performance might suggest that the essay section should be graded more or less rigorously than at present, but as long as the grades are scaled to the MBE distribution, uniformly applied changes in the rigor of essay grading should have *no effect whatever* on the scaled essay score. If the conclusion is that the norm for minimum competence with respect to performance on the essay portion should be adjusted "then the scaling process -- not the passing score -- should be reassessed" (Merritt Article at 937-38). Thus, Dr. Klein's suggestion that further study by candidates could obviate the impact of the change in the passing score is misleading. So long as there is scaling of essay raw scores to the MBE, then for the candidate pool as a whole only study aimed at improved *MBE* scores would be effective even though the panel's work did not assess the sufficiency of performance on the MBE at all, having focused only on the quality of the essays.[26]

The assumption of equivalent rigor forced by scaling the other subtests to the MBE is not the only concern when proposing to reason from acceptable performance

[25] Millman Evaluation at 7-7.

[26] The April 6, 2000, Comments of David M. White submitted to the Supreme Court of Florida in connection with its review of a similar study by Dr. Klein made this point: for the candidate pool as a whole, he commented, "improved essay writing will have no effect on the passing rate unless accompanied by improvement on the MBE." (*id*. at 7).

on the essay section to an overall pass/fail score. Dr. Klein's extrapolation from essay performance to overall examination score depends on the further assumption that the essays and the MBE are measuring competence in a similar way for each year's population of candidates, i.e., it depends on assuming that candidates are in general equally competent in each tested domain.

Dr. Klein's method also requires the assumption that just-minimally competent candidates would achieve approximately the same score on the MBE and on the essay portion. We disagree. The mere fact that the tests are scaled to have the same mean score and standard deviation does not imply that equal scores on the different tests signify equal competence within the different knowledge/skill domains. Whether by reason of better preparation or otherwise, the competence rate in one domain may well be different from the competence rate in the other domains. Alternatively, experts might well conclude that the threshold measure of competence in one area differs from that in another. A particular score on MBE might signify insufficient competence while the same score on the essays might signify sufficient competence. Dr. Klein's method requires that both the scale of competence and the threshold measure be at least approximately the same in all subtest domains in order to accurately apply the competence threshold set for the essay to the test as a whole. But the whole *point* of having multiple sections is the expectation that competent performance may very well differ across domains.

We have no doubt that there is *some* correlation between performance on the essays as a whole (measured by raw score) and performance on other portions of the test.[27] However, the Board must believe that each section of the examination tests a distinct area of competence, in a distinct way -- why else have so extensive an examination or use New York essays in addition to an extensive multi-state multiple choice test? Multiple essay questions are used on bar examinations precisely to assess candidates' knowledge of different bodies of law, and for this reason the applicants' scores on different questions may show relatively low correlations (Merritt Article at 953-54). This means that it is *reasonable* to expect some applicants to score poorly on some questions and yet achieve a passing score on the exam as a whole. In this regard we again note that current Board policy is that there is no "passing" score for subparts of the examination, only for totality of performance.

We support a study of the essay portion of the New York exam that evaluates the essay section as a whole. Such a study might well support the conclusion that scores on the New York essays should be measured against some objective standard and directly added into the total score without scaling to the MBE distribution. This might result in a raising, or a lowering, of cohort scores. A study principally focused on the essays might even be able to conclude that the relative weighting of essays and MBE results should be changed. But even such a different study should not be the basis for changing the overall passing score, until and unless a great deal more was learned about the interplay of performance on the various segments of the examination.

[27] That was the case a decade ago. Millman Evaluation at 9-2 (showing relatively strong correlation).

3. The Need For Analysis Coupled With Standard Setting

To extrapolate from a standard-setting exercise to the overall passing score, the Board should conduct a thorough analysis of performance on the various segments of the examination, both before and after scaling. Such a study should be possible for the Board using data already available to it. The 1993 Millman Evaluation provides a framework for such a project as well as a baseline for determining to what extent, if any, there has been change in candidate performance over the last decade. We believe the Board should also study the fairly dramatic recent fluctuations in the passing *rate* on the examination.[28] Such an analysis might make clear the extent to which such changes are a function of MBE performance rather than performance on the New York essays. They might also indicate whether because the essay score is scaled to the MBE, apparent declines in *MBE* performance are having a double effect by also artificially depressing scores for the New York essays independent of any objective change in the quality of essay responses.

Setting a passing score -- the line measuring "enough" -- also requires analysis of the significance for the measurement of competence of the multiple choice questions (MBE and NYMC), weighted at 50% of total score.[29] If the Board wants to review the usefulness of the overall passing score as a way to screen for minimum competence, it should review the entire test, using appropriate methods. That review should include a re-thinking and testing of the various assumptions discussed above which underlie the current scoring system but remain essentially unstated and certainly unchallenged. Michael Kane has explained that procedures for assessing the use of multiple choice questions to measure competence are fundamentally different from procedures appropriate for assessing essay and performance tests. [30]

Even assuming all reliability and validity issues were generally resolved in favor of the MBE (a topic far beyond the scope of this comment), the validity tests done for the MBE make no claim that MBE scores represent a linear scale of competence. The

[28] Report at 16, footnote 21 states that the passing rate was 71.4% in July 1996, 67.5% in July 2000, and 72% in July 2001.

[29] Although the Board's Report goes on at some length about the "standards" involved, it alternately minimizes the proposal by pointing out that the proposed change, in "raw score terms . . . represents correctly answering an additional three items on the 200 items MBE." (Report at 20). We find this observation ironic, since it is review of the New York essays, not the MBE, which prompts the Board to suggest a change. We also suspect that this statement is incorrect because it has not factored in the *weighting* of the 200 item MBE at 40% of the total score. If we are right, it is probably the case that if a candidate's score on all other phases of the test remained the same, it would take an additional 8 or more correct MBE answers to move from 660 to 675. In any case, there is little reason to think that still more time spent studying for the MBE would improve the overall performance of the pool (see White, *supra* note 26 at 7-8), and hence a shift for the cohort of candidates presently in the 660 range of 8 MBE items (or even 3) seems unlikely.

[30] *See* Kane, *supra* note 21, at 7. We appreciate that an up-to-date correlation analysis might provide a rationale for extrapolation from overall essay performance (*not* performance on an isolated question) to the overall score. However, we do not see why such indirect reasoning should suffice.

NCBE has not determined that any score on the MBE demonstrates "minimum competence." Each state must make that determination for itself. If the Board is going to review the question of what essay performance demonstrates "minimum competence," the MBE, which carries equivalent weight in the overall score, should be assessed as well. And surely the Board must be open to the possibility that performance on the MBE is over-weighted in the overall examination scoring.

4. Research Proposed By the Board

The Board proposes commissioning a study "to gather and assess demographic information, including age, race and gender, and information regarding candidates' LSAT scores, law school GPA and quartile ranking, foreign education, and other information. The study will be used to study the correlation between demographic factors and other indicators and performance on the bar exam." (Report at 14; the statement is repeated at 19.) We applaud that plan and would be eager to participate in the design of such a study. We differ with the Board, however, on the *timing* of such a study.

The Board proposes to embark on such a study only *after* increasing the passing score, "to study the impact of the increased score" (*id.*). The Board's response is not what the public, the profession or the candidates should expect in such a "high stakes" situation. Given the potential impact of such a change, we believe such a study should be conducted to assist in deciding *whether* to increase the passing score.

Although the Board chastises those whose criticisms rest on "assumptions" about the current candidate pool (Report at 16), at this stage it is the Board that is assuming that a change in the passing score will produce a public benefit. If "standards for licensure are [to be] set by means of scientifically based studies informing our policy judgment," the time to conduct those studies is *before* deciding to change the standard.

5. Timing of this Proposal and Studies By Others

The ABA, AALS, the National Conference of Bar Examiners, and the Conference of Chief Justices recently announced the formation of a Joint Working Group on Legal Education and Bar Admissions to study a number of far-ranging concerns related to the administration of the bar examination, including a "concern[] about the current trend toward increasing minimum bar examination passing scores and the methods that various states are using to establish those cut scores."[31] The Joint Working Group is planning to hold a nationwide conference in the winter of 2004 to address these and other concerns and also to consider a revision of the Code of Recommended Standards for Bar Examiners. Given the fact that the Joint Working Group has announced its intent to specifically study the process of setting a passing score in

[31] Letter dated December 3, 2002 from John A. Sebert, Consultant on Legal Education to the American Bar Association to Professor Lawrence Grosberg, Chair, Committee on Legal Education and Admission to the Bar of the Association of the Bar of the City of New York.

individual jurisdictions, we believe it would be prudent to delay the decision on any change in New York's passing score until the work of that group is completed and fully evaluated.

Conclusion

We appreciate that the Board put substantial time, thought and effort into its Report, and we appreciate greatly the Board's intention to have "scientifically based studies informing our policy judgments." In view of the objections and concerns we have discussed, however, we urge the Board to withdraw its proposal to raise the minimum passing score. Should the Board decide not to withdraw its recommendation, we respectfully urge the Court of Appeals to reject the recommendation. At the very least, we request that the Board defer its proposal pending further review, including both (i) completion of a study of the demographic data that will indicate whether and to what extent an increase in the passing score will affect both the diversity of the profession and consequent delivery of legal services to all sectors of the public, and (ii) the completion of other test-centered and competence-focused research discussed in this report.

COMMITTEE ON LEGAL EDUCATION AND ADMISSION TO THE BAR
OF
ASSOCIATION OF THE BAR OF THE CITY OF NEW YORK

Lawrence M. Grosberg, Chair
Margaret M. Flint, Secretary

Carole M. Bass
Hon. Jack Battaglia
James Beha II*
Peretz Berk
Thomas E. Chase
Sherri Eisenpress
David Epstein
Harmon Fields
Margaret M. Flint*
Nick Fortuna
Paula Galowitz
Fabian Gonell
Hon. Sidney Gribetz**
Lawrence Grosberg
Rosemary Halligan
Eileen Kaufman*
Ellen Lieberman
Tim O'Neal Lorah
Joseph Marino
Karen Markey
Michele Molfetta
Beatrice O'Brien
Jonathan Rosenbloom
Gerald Russello

* Principal authors of the report.
** Dissents