NCBE®

National Conference
of Bar Examiners

Response by NCBE to the NYSBA Task Force Report

July 9, 2020

Executive Summary

July 29, 2020

NCBE®

National Conference
of Bar Examiners

302 South Bedford Street
Madison, WI 53703

ncbex.org
thebarexaminer.org
testingtaskforce.org

/ncbexaminers
company/ncbex
@ncbex
/ncbexaminers

Table of Contents

NCBE®
National Conference
of Bar Examiners

Executive Summary

Introduction

The National Conference of Bar Examiners (NCBE) offers this response to the Report of the New York State Bar Association Task Force on the New York Bar Examination (Task Force Report), which was released in March 2020. The Task Force Report includes numerous criticisms of the Uniform Bar Exam (UBE) and of NCBE, many of which are based upon errors and incorrect assumptions regarding psychometric methods and practices. Our response addresses these criticisms and errors while providing context and clarification regarding the UBE's uniformity, value, and fairness; how the UBE is scored; and the study of New York's adoption of the UBE that was conducted by NCBE at New York's request.

Psychometric Expertise Supporting NCBE's Tests

The UBE, like all NCBE test products, is scored and equated by NCBE's research/psychometric staff. The basic scoring and equating methods used for the UBE were established by internationally renowned psychometricians, each with decades of experience in high-stakes testing and educational measurement. NCBE's research/psychometric staff members all have advanced degrees in psychometrics or closely related fields—most have PhDs in psychometrics—and have been nationally recognized for their technical expertise by peers in the profession. NCBE also receives input on psychometric and technical issues from a Technical Advisory Panel made up of some of the world's leading psychometricians, and frequently engages with the Center for Advanced Studies in Measurement and Assessment (CASMA) at the University of Iowa.

Uniformity, Value, and Fairness of the UBE

The purpose of the UBE, like that of any bar exam or other licensure exam, is to help protect the public by offering a consistent assessment of whether examinees can demonstrate that they possess essential knowledge, skills, and abilities. The UBE offers clear benefits for bar applicants, would-be clients, employers, and law schools via increased mobility and marketability, as well as increased consistency in the subjects tested on the bar exam across jurisdictions.

The UBE includes the same questions, which are given the same weights and graded using the same grading materials with support provided by NCBE, in every jurisdiction. It tests on generally accepted fundamental principles, an understanding of which, combined with the legal skills and abilities also assessed by the UBE, provides the foundation needed to practice competently in any jurisdiction.

The UBE assesses essential knowledge, skills, and abilities in a manner that is fair to all examinees. Research has shown that similarly prepared examinees perform similarly on the bar exam regardless of race, ethnicity, or gender. And although deeply rooted social inequities have contributed to some examinees, particularly those from historically underrepresented populations, lacking the resources and opportunities to be as well prepared to pass the bar exam as those from majority groups, there is no evidence that the UBE creates or worsens a disparate impact. Rather, any performance disparities on the UBE reflect what culminates from a lifetime of inequities in the larger social environment. NCBE takes seriously the need to work to eliminate any aspects of its exams that could contribute to performance disparities among groups. We maintain high standards in developing our test questions through the work of our diverse drafting committees and by conducting a rigorous process of external review, bias review, pretesting, and differential item functioning (DIF) analysis to ensure fairness. We conduct or facilitate studies of predictive bias, and conduct research with jurisdictions—as in the New York study just completed.

NCBE's Equating Method

The MBE, the multiple-choice component of the UBE, uses a statistical procedure known as equating to adjust for potential differences in difficulty between exams. Equating makes it possible to report scaled scores with consistent score interpretations. The Task Force Report criticizes NCBE's equating method based on an oversimplified example that illustrates a different kind of equating than the type NCBE uses. The example used by the Task Force does not provide a fully accurate or fair representation of the actual process used to score NCBE exams.

Impact of Reducing the Number of Scored Items on the MBE

Beginning in February 2017, the number of scored items on the MBE was reduced from 190 to 175 in order to increase the number of unscored items being pretested for future use. The number of equator items remained the same. The Task Force Report claims that this change had a negative impact on the exam. In fact, however, the change had a negligible effect.

In particular, the Report claims that the change caused the reliability of the exam (a measure of the precision of scores) to decline. However, the reduction in the number of scored items was offset by an improved ability to select items that distinguish well between different levels of examinee proficiency, and the reliability of scores has in fact increased with almost every administration since February 2017.

Impact of the Changing Proficiency of Examinees over Time

The Task Force Report questions the comparability of scores given differences in examinee populations from one exam administration to another. However, the purpose of equating is precisely to ensure that scores have the same meaning over time, regardless of differences in examinee proficiency or in the difficulty of the exam.

Relative Grading

The Task Force Report, in criticizing the relative grading method recommended by NCBE for jurisdiction graders of the written portions of the UBE, appears to rely on an inaccurate description of relative grading. Relative grading is a means of providing uniformity to grading practices across different essays, graders, and jurisdictions. Graders should go through a calibration process before beginning their grading, and while grading they are asked to assign rank-ordered grades based on the merit of the answers, while using as much of the score scale as possible in order to limit the effect of grader bias. A relative grading approach that uses rank ordering is one step in a process that also includes scaling the written score to the MBE.

Scaling the Written Scores to the MBE

The Task Force Report claims that the UBE is vulnerable to "forum shopping," in which examinees intentionally try to take the exam in a jurisdiction where they believe they will have a better chance of passing due to differences in examinee populations and grader variability. However, the scaling formula used by NCBE helps compensate for such differences and for variations among graders, which are an unavoidable part of any grading process for essay and performance test components.

The Task Force also expresses concern that scaled written scores are not reliable enough to produce a reliable total exam score. However, this is not the case. While the reliability of written scores is somewhat lower than the reliability of the MBE, the combined score has a reliability well above the required minimum for a high-stakes exam.

Correlations Between MBE and Written Scores

The Task Force Report erroneously states that correlations between MBE scores and written scores are low and uses this claim to argue that written scores should not be scaled to MBE scores. In fact, MBE scores and written scores are strongly correlated. Therefore, it is appropriate to scale the written score to the MBE score.

Equal Weighting of the MBE and the Written Component

Contrary to the Task Force Report's claim that the written component of the UBE is not given significant weight in UBE scoring, the written component is in fact weighted 50% of the total UBE score.

New York UBE Study Included Appropriate and Sufficient Data for Analyses

The Task Force Report criticizes the UBE study that NCBE conducted for New York for including data from a limited number of data collection points (exam administrations). However, the number of examinees within each administration was large, providing sufficient data for analysis, including analysis of subgroups.

NCBE's Objectivity in Conducting New York UBE Study

The Task Force Report calls into question NCBE's objectivity in conducting New York's UBE study. NCBE undertook the study at the request of the New York State Board of Law Examiners (BOLE) as part of its mission as a nonprofit corporation. The New York Court of Appeals, in collaboration with the BOLE, approved the design of the study, and the BOLE provided the data. NCBE's role was to offer advice on study design, analyze the data, and prepare the report. A neutral, objective perspective was maintained throughout the report, which included as much detail as possible about the analysis that was performed so that anyone with questions about the results could examine the data themselves.

NCBE®
National Conference
of Bar Examiners

Response by NCBE to the NYSBA Task Force Report

The New York State Bar Association (NYSBA) charged a Task Force in April 2019 with examining the impact of the adoption of the Uniform Bar Examination (UBE) by New York. The report from the Task Force (hereafter referred to as the Task Force Report), released in March 2020, includes numerous criticisms of the UBE, the New York Law Examination (NYLE), and the National Conference of Bar Examiners (NCBE). NCBE offers this response to the Report's comments related to the UBE and NCBE, leaving the topic of the NYLE to others to address. Our response is organized around the following 11 primary issues raised in the Task Force Report:

I. Psychometric Expertise Supporting NCBE's Tests

II. Uniformity, Value, and Fairness of the UBE

III. NCBE's Equating Method

IV. Impact of Reducing the Number of Scored Items on the MBE

V. Impact of the Changing Proficiency of Examinees over Time

VI. Relative Grading

VII. Scaling the Written Scores to the MBE

VIII. Correlations Between MBE and Written Scores

IX. Equal Weighting of the MBE and the Written Component

X. New York UBE Study Included Appropriate and Sufficient Data for Analyses

XI. NCBE's Objectivity in Conducting New York UBE Study

XII. Conclusion

NCBE's response begins with a synopsis of the qualifications of organizations that contributed to the development of the bar examination over its history and a summary of the qualifications of NCBE's current research/psychometric staff. We then address the Task Force's criticisms of the UBE. Finally, we address the numerous errors and incorrect assumptions about scoring and equating methods used by NCBE and the faulty descriptions of some basic measurement concepts that the Task Force Report includes in its criticisms of NCBE's psychometric methods and practices. Throughout this response, when an article or book is cited, a parenthetical citation is provided in the text indicating the author, publication year, and page number(s) (e.g., Klein & Bolus, 1997, p.12). Full bibliographic information for these citations is provided in the bibliography included at the end of the response.

I. Psychometric Expertise Supporting NCBE's Tests

Much of the critique in the Task Force Report calls into question the expertise and integrity of NCBE staff. A short history of how the bar examination was developed and is maintained, and by whom, is provided here.

The bar examination has evolved over a period of decades.[1] NCBE's role in the process began with the creation of the Multistate Bar Examination (MBE) in 1972. The basic scoring and equating methods were established by internationally renowned psychometricians, each with decades of experience specifically in high-stakes testing and educational measurement and coming from highly respected testing organizations: ETS, ACT, and the Rand Corporation. Many of these methods have been perpetuated by NCBE's psychometric staff for good reason.

All NCBE research/psychometric staff have scientific backgrounds. Most staff currently in the department, and all staff working directly on the operational equating of NCBE's exams, have PhDs in psychometrics. Most staff without PhDs specifically in psychometrics have advanced degrees in closely related fields. Some staff members worked at the National Board of Medical Examiners before joining NCBE, and others were previously on the faculty at R1 universities (major research universities). Collectively, current staff have well over 160 years of directly relevant professional experience, have published over 200 articles in peer-reviewed[2] professional journals, and have made over 300 presentations at peer-reviewed professional meetings. Staff have been nationally recognized for their technical expertise by invitations to serve as reviewers for 35 professional journals, members of the editorial boards for 10 professional journals, and members of more than 45 grant review/technical advisory panels for the National Institutes for Health and other agencies, as well as by numerous awards and invitations to present at professional meetings.

Additionally, NCBE periodically seeks input on psychometric and technical issues from a Technical Advisory Panel[3] made up of some of the world's leading psychometricians, each with decades of experience. Consulting with a panel of outside experts is a common practice among high-stakes testing organizations to ensure a breadth and depth of expertise and ideas. NCBE also frequently engages the Center for Advanced Studies in

[1] For a timeline of key milestones in NCBE's testing program, readers are referred to NCBE Testing Milestones, available at https://testingtaskforce.org/about/ncbe-testing-milestones/.

[2] Peer-reviewed research is accepted for publication or presentation only after being reviewed and approved by individuals with recognized expertise in the profession—psychometrics, in this case.

[3] A list of the measurement experts that compose NCBE's Technical Advisory Panel is available at http://www.ncbex.org/statistics-and-research/tech-advisory-panel/.

Measurement and Assessment[4] (CASMA) at the University of Iowa.

II. Uniformity, Value, and Fairness of the UBE

This section will begin by briefly describing the history and purpose of the Uniform Bar Exam (UBE) for context, followed by a discussion of several areas of disagreement about the UBE between NCBE and the Task Force.

History and Purpose of the UBE

The UBE was first adopted a decade ago, with benefits to law school graduates, the legal profession, and the public squarely in the minds of the individuals and entities working to develop a uniform exam. The UBE has since been endorsed by the Law Student and Young Lawyers Divisions of the ABA, the ABA House of Delegates, the Council of the ABA's Section of Legal Education and Admissions to the Bar, and the Conference of Chief Justices.[5] The UBE offers clear benefits to students via increased mobility and marketability as well as increased consistency in the subjects tested on the bar exam across jurisdictions. For graduates seeking licensure in more than one jurisdiction, it eliminates the costs of preparing for and retaking the bar exam in additional jurisdictions. The value of the score portability offered by the UBE was particularly important in the challenging economic times after the 2008–2009 recession.

The benefits and quality of the UBE are corroborated by the expansion in the participants in the UBE. As of June 12, 2020, 37 jurisdictions have adopted the UBE. Representatives of UBE jurisdictions meet regularly to share their experiences and discuss policies and practices related to uniformity. There have been no significant complaints voiced by the UBE jurisdictions.

UBE jurisdictions retain their autonomy, independently setting their own passing scores and maintaining control over additional policy matters, including how long incoming UBE scores will be accepted, admission eligibility requirements, and policies surrounding all character and fitness decisions, among many others.

To be clear, NCBE supports the UBE because it is the right thing to do for law students and for the profession, not, as some claim, because it enables NCBE to "sell more tests." On the contrary, the UBE results in NCBE selling fewer tests, because examinees who move to another jurisdiction can transfer their UBE score without retaking the exam, whereas they previously would have had to retest.

Testing Core Knowledge and Skills and Generally Applicable Principles of Law

The Task Force Report cites critics of the bar exam who argue that the exam is flawed because it does not test some key skills that lawyers need. No licensure exam could test every skill a professional needs without the exam becoming unreasonably lengthy and expensive. However, this impossibility does not represent a fatal flaw in the bar exam or other licensure exams. The bar exam tests a representative sample of core knowledge and skills. The purpose of the

[4] Information about the Center for Advanced Studies in Measurement and Assessment is available at https://education.uiowa.edu/centers/casma.

[5] Links to these endorsements are provided at http://www.ncbex.org/exams/ube/further-reading/.

bar exam (and other licensure examinations) is to help protect the public by offering a consistent assessment of whether examinees can demonstrate that they possess essential knowledge, skills, and abilities. The examination should be considered within the context of education requirements, character and fitness requirements, jurisdiction-specific law components such as the New York Law Course (NYLC) and NYLE, requirements for continuing education, etc. It would be a practical impossibility to address every topic covered in law school or encountered in practice on an exam.

Professor Suzanne Darrow-Kleinhaus[6] wrote, in a 2004 article in the *Journal of Legal Education*, "But in our world competence matters, as it does in the case of a lawyer's ability to engage critical analysis. The bar examination, by testing competency in the most basic and essential analytical skills required for the practice of law, serves a necessary function" (p. 442). She continues, "I submit that the bar examination

- seeks to measure the analytical skills required for the practice of law, which requires an understanding of the rules and not just the ability to memorize.
- tests the ability to act and not react under pressure.
- requires a sound mastery of legal principles and basic knowledge of core substance for which tricks or techniques cannot be substituted.
- covers the subjects students should have learned in law school in preparation for the general practice of law.
- neither demands nor requires the sacrifice of skills-based courses for substantive courses." (p. 444)

The Task Force Report cites Professor Deborah Merritt's claim that "the UBE requires really extensive memorization of federal rules and what we call 'the law of nowhere,' because the law of nowhere is supposedly majority rules" (p. 30). The UBE in fact tests on generally accepted fundamental legal principles in the 37 jurisdictions that have adopted it—principles that newly licensed lawyers should know. This claim about testing "the law of nowhere" is generally leveled at the Multistate Essay Examination (MEE) component of the examination. MEE questions, whenever possible, focus on testing those issues of minimal competence where the answer will be the same in the majority of jurisdictions. In all questions, the MEE is designed to prompt an examinee to review a set of facts, recognize the legal issues involved, and craft a thoughtful response that demonstrates an ability to engage in legal analysis and to produce a clearly written and well-supported answer—skills that are critical to the practice of law, whatever the jurisdiction one is licensed in.

The legal authorities for topics in Evidence, Civil Procedure, Constitutional Law, Criminal Procedure, and Secured Transactions are the respective federal rules, the United States Constitution, Supreme Court precedent, and Article 9 of the Uniform Commercial Code (adopted in New York). Questions on Business Associations test on common corporate law issues. Partnership questions are drafted in such a way that there will be no significant difference in the result whether the jurisdiction follows the Revised Uniform Partnership Act or the Uniform Partnership Act (1914).

[6] Professor Darrow-Kleinhaus is a member of the NYSBA Task Force on the New York Bar Examination.

In areas in which there is likely the greatest variation in the law among jurisdictions, such as Trusts and Estates, MEE questions will often provide the relevant statute for examinees to apply. One of the most important skills a minimally competent lawyer should have is the ability to read and apply a statute. Again, because the standard is testing for minimal competence, many of the issues tested involve basic concepts where there is a similar result across jurisdictions (e.g., all states have some form of anti-lapse statute), even though the MEE grading materials may cite the Uniform Probate Code or the Uniform Trust Code as legal authority. Another approach was used in a question on the July 2017 MEE—the execution of the will in the problem was described in such a way that it would meet the execution requirements of any state. That said, graders may also give credit even though the legal conclusion reached relies upon a minority rule, if the examinee's legal analysis reflects a solid understanding of the issues.

An understanding of generally applicable legal principles, combined with the legal skills and abilities assessed in the UBE, provides the foundation needed to practice competently in any jurisdiction. Characterizing the UBE as testing "the law of nowhere" is inaccurate. It would be more apt to say that the UBE tests "the law of everywhere."

None of this is to say that there is not important and unique state-specific information lawyers need to know to practice safely and effectively, which is the purpose served by the New York Law Course (NYLC) and NYLE. NCBE believes that it is most fair and efficient for examinees to be tested on the most widely applicable law, and to learn state-specific law through law school coursework or experiential learning, through jurisdiction-specific components like the NYLC and NYLE, through continuing legal education, through their own experience and study, and/or through mentorships or on-the-job training.

Uniformity of Grading Practices Across Jurisdictions

The Task Force Report claims that the UBE is not uniform because the passing score varies from jurisdiction to jurisdiction and the grading is not consistent (p. 11). Differences in passing scores do not, however, make the exam itself non-uniform. Different branches of the US military, for example, have different height and weight requirements, but no one would take those differences in *standards* to imply that the *instruments* (scales or measuring tapes) are not uniform. The UBE includes the same test questions (MBE, MEE, MPT), which are assigned the same weights and graded in accord with the same grading materials, in every jurisdiction that administers it. It is unclear how the test *instrument* itself could be made more uniform.

Grading of the written components of the UBE is performed locally, with each jurisdiction recruiting and monitoring its own graders. However, all UBE jurisdictions are required to use NCBE grading materials, which are extensive and detailed. Additionally, coordinated training and calibration[7] support is provided for UBE jurisdictions via NCBE's MEE/MPT Grading Workshops, which are held the Saturday after each bar exam administration. NCBE has also written extensively on best practices in essay grading and routinely shares these

[7] Calibration is a process designed to ensure that graders agree upon and adhere to consistent grading standards.

best practices with UBE jurisdictions via workshops and articles in the *Bar Examiner,* NCBE's quarterly publication. The grading might not be perfectly uniform in practice, to the extent that human graders may not always be perfectly consistent despite the comprehensive grading materials, training, and calibration support provided by NCBE to all UBE jurisdictions.

The Benefit of Score Portability

The Task Force Report argues that "only a small number of candidates benefit from portability of scores out of New York" (p. 70). As of July 7, 2020, more than 6,900 UBE scores have been transferred out of and more than 2,700 have been transferred into New York since 2016 (6,941 scores in and 2,781 scores out). The transfer of over 9,500 UBE scores in and out of New York represents concrete benefits to both lawyers and the public. It represents 9,500 additional bar exams that did not have to be taken by newly licensed lawyers. To most, 9,500 would not be "a small number of candidates." If New York had not adopted the UBE, the number of candidates benefitting from score portability would have been zero; certainly, 9,500 is a large number compared to zero. The Task Force asserts that "the primary arguments in favor of the UBE center around convenience or lack of hardship to bar applicants," which it views as a shift of focus away from public protection (p. 2). NCBE does not agree with this "either or" mentality; we believe that the UBE meets the necessary public protection purpose while also facilitating new lawyer mobility to the benefit of bar applicants as well as would-be clients, employers, and law schools. Apparently, the 36 jurisdictions aside from New York that have adopted the UBE similarly believe that the UBE serves the important goal of public protection.

Disparate Impact for Women and Minorities

The Task Force Report is highly critical of the UBE for performance differences on the UBE reported for women and minorities in a study conducted by NCBE for the New York State Board of Law Examiners.[8] While the Task Force concedes "[t]hat these troubling statistics existed prior to the adoption of the UBE," it goes on to mischaracterize NCBE's objective, neutral reporting of the results of our research—that performance on the bar examination by racial, ethnic, and gender subgroups was not affected by New York's adoption of the UBE—as "NCBE's bland acceptance of the broken status quo" (p. 37). The Task Force Report states that "NCBE should not take comfort in a finding that disparities exist but are no worse than they always were," and even goes so far as to cite one Task Force member's comment that "[NCBE] is saying there's a gender disparity and there's a race disparity, but we're fine with it" (p. 37).

To be crystal clear: NCBE is *not* "fine with it." It is central to our mission to promote "fairness, integrity, and best practices in admission to the legal profession . . ." and to help foster "[a] competent, ethical, and diverse legal profession."[9] The Task Force unjustifiably and unprofessionally

[8] In adopting the UBE, the New York State Court of Appeals directed the New York State Board of Law Examiners (BOLE) to study the impact of the change to the UBE on bar exam performance. The BOLE requested assistance from NCBE in conducting the study, which NCBE provided as part of its mission as a nonprofit corporation. NCBE's report, released by the Court of Appeals in August 2019, is available at https://www.nybarexam.org/UBEReport.html.

[9] NCBE's mission and vision are set out on our website at http://www.ncbex.org/about/.

insults NCBE's integrity, values, and principles when what is needed to achieve a diverse legal profession are professional and respectful partnerships all along the education and training pipeline.

NCBE takes seriously the need to work to eliminate any aspects of its exams that could contribute to performance disparities among groups (including race and gender, among others such as disability or background). We maintain high standards in developing our test questions through the work of our diverse drafting committees and by conducting a rigorous process of external review, bias review, pretesting, and differential item functioning (DIF) analysis to ensure fairness. We conduct or facilitate studies of predictive bias, and conduct research with jurisdictions—as in the New York study just completed.

Additionally, NCBE prioritizes diversity on our Board of Trustees, and we have an active Diversity Issues Committee whose purpose is to recommend policies and initiatives through which NCBE can enhance the participation and performance of historically disadvantaged groups with respect to legal education and bar admissions, including bar passage. We are launching a series of articles in the *Bar Examiner* dedicated to the topic of why diversity and inclusion matter. The series will debut this summer with the inaugural column by Judge Phyllis Thompson of the DC Court of Appeals, a member of our Board of Trustees. We also continue to partner with the Council on Legal Education Opportunity (CLEO) to promote bar exam success for students from historically disadvantaged groups. Our commitment to CLEO is in its second year, and we've renewed it for another three years. And NCBE's Testing Task Force is in the final year of a comprehensive, empirically based study to design the next generation of the bar examination. As part of that study, we are convening diverse committees of stakeholders to develop recommendations for the future bar exam's blueprint and design.

The Task Force Report says the UBE has a "known disparate impact." This framing makes it seem as though the UBE creates a disparate impact, as opposed to reflecting what culminates from a lifetime of inequities in the larger social environment. The 2012 report by the American Psychological Association (APA) Presidential Task Force on Educational Disparities notes:

> Pervasive ethnic and racial disparities in education follow a pattern in which African American, American Indian, Latino, and Southeast Asian groups underperform academically, relative to Caucasians and other Asian-Americans. These educational disparities (1) mirror ethnic and racial disparities in socioeconomic status as well as health outcomes and healthcare, (2) are evident early in childhood and persist through K-12 education, and (3) are reflected in test scores assessing academic achievement, such as reading and mathematics, percentages of those repeating one or more grades, dropout and graduation rates, proportions of students involved in gifted and talented programs, enrollment in higher education, as well as in behavioral markers of adjustment, including rates of being disciplined, suspended and expelled from schools. (APA, 2012, p. 7)

Research done by others has shown that similarly prepared examinees perform similarly on the bar examination regardless of their race/ethnicity/gender.[10] Studies conducted by Stephen P. Klein, PhD, and Roger Bolus, PhD, on why some groups do better on the bar exam than others explored several factors, including the format of questions, subjects tested, the race/ethnicity of graders, and academic ability of the examinees. They concluded that "[a]pplicants with the same LGPAs [law school grade point averages] from the same law school have about the same probability of passing regardless of their racial/ethnic group. The exam does not favor one group over another" (Klein & Bolus, 1997, p.12). More recently, in a study of the causes of decline in performance on the California bar exam, including the MBE, researchers found that "[c]onsistent with the 1997 findings of Klein and Bolus, this study reconfirmed that racial/ethnic minorities with equivalent credentials to whites will tend to earn the same scores on the CBX [California bar examination] and have the same probability of passing" (Bolus, 2018, p. x).

The deeply rooted societal inequities noted by the APA contribute to some examinees, particularly those from historically underrepresented populations, lacking the resources and opportunities to be similarly prepared in comparison to examinees from majority groups. In other words, the fact that some people are not as well prepared to pass the bar exam as others is the result of a serious and long-standing pipeline problem. While it would be wonderful if bar preparation were able to compensate for a lifetime of inequities, such that performance disparities no longer exist by the time students take the bar exam, this is unfortunately not the current reality. We must all work together to eliminate the disparities so we can achieve the shared goal of a diverse legal profession to better serve society and promote justice for all.

III. NCBE's Equating Method

There are several questions and misunderstandings in the Task Force Report regarding how NCBE equates the MBE. It will help to set the record straight by giving a brief description of the process.

The MBE, like most other large-scale tests for high-stakes decision making, uses a statistical procedure known as *equating* to adjust for potential differences in difficulty between the current exam and past exams. Equating makes it possible to report scaled scores with consistent score interpretations regardless of when an examinee takes the exam.[11]

The Task Force Report offers a description meant to provide readers with some basic understanding of equating. While some sort of analogy or example could have been useful to readers, the description the Task Force provides is not technically accurate and not applicable to NCBE or its

[10] The differences in performance on the bar examination between men and women are small but persistent. NCBE typically finds that the differences between women and men are in opposite directions on multiple-choice and written components of the exam, with women tending to do better on the written portion and men tending to do better on the multiple-choice portion. Overall, these differences have tended to result in men passing at slightly higher rates on average compared to women. The explanation for the small but persistent performance differences between women and men is not entirely clear to us or to other researchers studying these patterns across time and across exams; as we observed in the New York UBE study, however, men tended to have higher average law school GPAs and average LSAT scores than women, so it would not be unexpected to see corresponding differences in bar exam scores based on these other indicators.

[11] For further discussion of conditions conducive to satisfactory equating, see Michael J. Kolen and Robert L. Brennan, *Test Equating, Scaling, and Linking* (second edition), Springer (2014), pp. 312–313.

exams. As a result, it is more confusing or obfuscating than it is helpful. For example, the Task Force Report quotes Dr. Nancy Johnson saying that she assumes NCBE uses a chained equipercentile method (p. 56); however, NCBE does not use this method, nor has it ever been used for equating the MBE. Rather, NCBE uses *item response theory* equating, which is explained in more detail below.[12]

The Task Force Report gives an overview of equating through an example (quoted in its entirety in the footnote below) that involves three groups of test takers with differential performance on equator items (questions) and "unique" (non-equator) items.[13] The example was taken by the Task Force from a post on Professor Derek T. Muller's blog, Excess of Democracy.[14] In his original (2015) post, Muller notes that his equating example is an oversimplification, but useful insofar as it provides at least a basic understanding for individuals unfamiliar with the concept. Professor Muller is to be applauded for his attempt to explain equating in basic terms to individuals unfamiliar with it (and possibly also unfamiliar with most statistical concepts). Unfortunately, the Task Force Report failed to include his caveats. Although Professor Muller's example is a good start at setting out some very basic information about equating, it lacks quite a bit of important nuance and technical detail, as Professor Muller himself acknowledges.

For many audiences and for many purposes, including Professor Muller's, this nuance and detail might not be important to understand. However, it is wholly inappropriate to use such a simplified example as evidence meant to show that NCBE's equating methods are flawed. Most immediately, the example does not include relevant information about the total number of equators and unique questions involved, which would be important to trained psychometricians thinking about

[12] Equating, and more specifically equating the MBE, is a topic that has been addressed several times in the Bar Examiner, NCBE's quarterly publication. Interested readers may refer to the following articles for more accurate information regarding NCBE's design and method for equating the MBE:
NCBE Testing and Research Department, "The Testing Column: Q&A: NCBE Testing and Research Department Staff Members Answer Your Questions," 86(4) *The Bar Examiner* (Winter 2017–2018) 34–39; Mark A. Albanese, PhD, "The Testing Column: Equating the MBE," 84(3) *The Bar Examiner* (September 2015) 29–36; Michael T. Kane, PhD, and Andrew A. Mroch, PhD, "Equating the MBE," 74(3) *The Bar Examiner* (August 2005) 22–27; Deborah J. Harris, "Equating the Multistate Bar Examination," 72(3) *The Bar Examiner* (August 2003) 12–18; and Lee Schroeder, PhD, "Scoring Examinations: Equating and Scaling," 69(1) *The Bar Examiner* (February 2000) 6–9.

[13] "Consider two groups of similarly situated test-takers, Group A and Group B. They each achieve the same score, 15 correct, on a set of the 'equator' questions. But Group A scores 21 correct on the unique questions, while Group B scores just 17 of these questions right. Based on Groups A and B's same score on the equator questions, we can feel fairly certain that Groups A and B are of similar ability. We can also feel fairly certain that Group B had a harder test than Group A. This is because we would expect Group B's scores to look like Group A's scores because they are of a similar capability. Because Group B performed worse on unique questions, it looks like they received a harder group of questions. Now we scale the answers so that Group B's 17 correct answers look like Group A's 21 correct answers, thus accounting for the harder questions. Bar pass rates between Group A and Group B should then look the same. In short, it is irrelevant if Group B's test is harder because the results will be adjusted to account for variances in test difficulty. Group B's pass rate will match Group A's pass rate because the equators establish that they are of similar ability.
Now consider Group C. In the unique questions, Group C did worse than Group A (16 right as opposed to 21 right), much like Group B (17 to 21). But on the equators, the measure for comparing performance across tests, Group C also performed worse, 13 right instead of Group A's 15. We can feel fairly certain, then, that Group C is of lesser ability than Group A. Their performance on the equators shows as much. That also suggests that when Group C performed worse on unique questions than Group A, it was not because the questions were harder; it was because they were of lesser ability" (p. 46).

[14] Although the example from Professor Muller's blog was initially used in the Task Force Report without attribution, the chair of the Task Force subsequently apologized to Professor Muller and the Report was updated to give attribution. See D. T. Muller, "When the Task Force on the New York Bar Examination plagiarizes your work without attribution" (April 3, 2020).

the example, since the number of equators should be a certain percentage of the total number of questions. For instance, in Professor Muller's example, Group A scores 21 correct on the unique questions, but is that out of 21 unique questions or 210? It is important to know the denominator to make sense of the example.

Assuming the numbers used in the example represent the mean number of items correct, the example illustrates what is called mean equating, where the means of groups are adjusted to deal with group differences. NCBE does not use mean equating but, instead, *item response theory* (IRT), which Professor Muller acknowledges. The beauty of IRT results mainly from the fact that it explicitly models the relationship between *individual* items and examinees.[15] Using the IRT statistical parameters and actual responses to those items, examinees' ability levels are estimated more accurately. The estimation process of item parameters and examinee ability levels is complex, both conceptually and computationally. And IRT equating adjusts MBE scores in a more *holistic*, *nonlinear* manner, taking into account characteristics of both items and examinees, rather than simply adding or subtracting some raw score points.

While these points are technical in nature, they are important because the example used by the Task Force to explain equating does not provide a fully accurate or fair (absent the caveats in Professor Muller's original description) representation of the actual process used to score NCBE exams.

IV. Impact of Reducing the Number of Scored Items on the MBE

Impact on Equating

The Task Force Report is highly critical of the reduction of the number of live (scored) items on the MBE beginning in February 2017, arguing that it had a number of negative effects, particularly on equating the exam. While the number of live items was reduced from 190 to 175 in order to increase the number of unscored items being pretested for future use, the number of equator items remained the same. The set of equators embedded in each of the MBE forms may be viewed as a "mini-MBE," because it is constructed to represent the content and statistical characteristics of the whole test. A best practice in psychometrics is to ensure that "[each equator set is] at least 20% of the test for tests of 40 items or more."[16] The number of equators used on the MBE far exceeds this guideline.

In addition to strictly observed and consistent rules about the number of equators included on each MBE, a comprehensive list of criteria is considered when equator items are selected, including content representation, date of most recent use, placement from most recent use, and statistics from most recent use. An extensive review of equators (and, of course, of all the items) is performed before the exam is administered, at both the individual item level and the overall exam level, in order to ensure that the items meet criteria for content, statistics, and testing industry best practices. Once the examinees' responses are received after the exam, additional

[15] For an introduction to IRT, see Mark A. Albanese, PhD, "The Testing Column: Equating the MBE," 84(3) *The Bar Examiner* (September 2015) 29–36.

[16] See Michael J. Kolen and Robert L. Brennan, *Test Equating, Scaling, and Linking* (second edition), Springer (2014) 312–313.

analyses are conducted to evaluate performance on the equators and to verify their function.[17]

Before implementing the change from 190 to 175 live items, NCBE modeled the impact the reduction in the number of scored items would have on scaled scores using prior MBE exams. This modeling was conceptually straightforward: psychometricians went back and rescored old exams as though they had 15 fewer scored non-equator items. NCBE research/psychometric staff did not expect to see more than negligible effects via this modeling on individual scores or on mean scores for the group. Results of the February 2017 and subsequent exam administrations bore out the prediction of the modeling and confirmed that the change had a negligible effect.

Impact on Score Reliability and Scaled Scores

Score reliability is a measure of the precision of scores and indicates the extent to which a group of examinees would be rank-ordered the same across multiple test administrations covering the same content. Reliability estimates can assume values that range from 0 to 1.0, with a value of 0 indicating that scores on repeated assessment give no information about how the examinee will rank order from one administration to the next, while a value of 1.0 indicates that one score will perfectly predict rank order on a future administration. The Task Force Report cites a 2012 *Bar Examiner* article to support its argument that score reliability declines if the number of scored items is reduced. In the article cited, key concepts about sampling, reliability, and validity are introduced, and the effect of test length on the reliability of test scores is explained.

Other things being equal, the longer the exam, the greater the reliability, and vice versa. But the key point here is "other things being equal." Besides test length, score reliability can also be affected by item discrimination (Traub, 1994)—that is, the ability of items to distinguish between different levels of examinee proficiency. Other things being equal, a test consisting of items with higher discrimination values would lead to a higher reliability, so test length and item discrimination are compensatory in nature; increasing one can offset a reduction in the other. This is a very basic concept that anyone versed at any level in psychometrics would be expected to understand, but it was not mentioned in the Task Force Report, again raising questions about the expertise of the psychometric consultation to the Task Force.

As explained above, NCBE psychometricians modeled the expected impact of the reduction in the number of scored items on the reliability of scores and found that no change or maybe even a slight increase could be expected. The reduction in the number of scored items would be offset by an ability to select items that better discriminated between lower and higher scoring examinees because of the added pretest data and an ability to be more selective because of the need for fewer items. This finding has been confirmed in the years since, as the reliability of scores has steadily increased with almost every administration since the number of scored items was reduced to 175.

[17] For detailed information concerning the selection and review of equators, see Mark A. Albanese, PhD, "The Testing Column: Equating the MBE," 84(3) *The Bar Examiner* (September 2015) 30–32.

V. Impact of the Changing Proficiency of Examinees over Time

The Task Force Report calls MBE standardization and equating processes unreliable because the examinee population has changed over the years and because the characteristics of the examinees taking the July exam and February exam are different, the former being about 30% repeat takers while the latter are about 60% repeat takers. In fact, the Report states that "[t]here is no valid way to standardize the test if the current population is not equivalent to past ones" (p. 55).

NCBE uses the term "standardization" to refer to the conditions under which the test is administered, such that they are the same across examinees and across time. NCBE has procedures in place that help to make sure that the bar examination is administered in a standard way across different administration sites. However, the authors of the Task Force Report seem to consider "standardization" to apply to a sample. This is a nuanced distinction, but an important one. Standardized samples are often used in clinical assessments to provide reference information (norms) about how an individual's performance compares to a particular (clinical) sample. This type of "standardization" is not used with licensure exams like the bar exam.

The whole point of equating is to account for potential differences in difficulty across exams containing different questions; but differences in examinee proficiency can also be separated out by using an appropriate equating design (like the common-item nonequivalent groups design used with the MBE).[18] In fact, the explanation of equating provided in the Task Force Report even includes an example of examinees differing in ability (Group C versus Group A—see footnote 13 and discussion in section III). The fact that examinee proficiency may change somewhat over time or across exams does not imply that exam scores are invalid or that the equating of that exam therefore becomes "unreliable," as the Task Force Report states. Reliability indicates the degree of consistency in the quality (precision) of measurement; it does not require that the measurement itself is constant. A thermometer does not become less reliable because the temperature changes. Indeed, equating accounts for potential changes in examinee proficiency, as explained through the Task Force's own example.

VI. Relative Grading

The Task Force Report appears to rely on an inaccurate description by Professor Suzanne Darrow-Kleinhaus of the relative grading method recommended by NCBE. The example Darrow-Kleinhaus provides to criticize relative grading makes one factual error and fails to consider the systems used to minimize grading errors. The example describes a grader using the "bucket system," in which essays are assigned to "buckets" corresponding to each separate score (in this case, 1–6) in the grading system:

> …the grader finds that most of the answers are strong and belong in the 4 and 5 buckets. However, since all the buckets must be filled, distinctions must be made and the papers are redistributed. Unfortunately, these adjustments

[18] "Common-item nonequivalent groups" is a design used for collecting data for equating. As its name suggests, this design involves the use of some common items between the new and old test forms and assumes that candidates taking the new form are not always equivalent in ability when compared with those taking the old form.

> do not have the same effect on all of the papers. Papers at the top end of the bucket list may get a boost up but those in the middle may not fare so well because some papers must be placed in the 1, 2, and 3 buckets. This may well result in an examinee failing the bar exam because he or she was kicked out of the higher bucket on a technicality—in effect, a distinction without a difference as to competency, just bucket placement. (Darrow-Kleinhaus, 2019, p. 176)

The factual error is the claim that distinctions between essays must be forced upon graders so that all buckets are filled. Albanese (2016) states that "essays should get different grades only if quality differences merit different grades, not to hit targets for allocations to grading categories" (p. 35). NCBE has repeatedly recommended that graders use the entire score scale and do their best to spread out scores, but that differences in grades should be based upon merit. The recommendation to use the entire scale is primarily to limit grader bias from impacting the grades awarded. Otherwise, a lenient grader might not use the bottom grades while a harsh grader might not use the top grades when grading the same set of essays. Because written scores effectively involve weighting scores by score variability, the essays graded by a particularly stringent or lenient grader will be given less weight than those read by graders who use the full scale.

The Darrow-Kleinhaus example also fails to consider that NCBE advocates for graders to go through a calibration process in which they must demonstrate consistent grades before they engage in live grading. Additionally, NCBE recommends that there be calibration papers embedded throughout any set of answers to be graded to help ensure that graders are remaining consistent in applying grading standards. Graders making arbitrary allocations to "fill buckets" will not be consistent with the calibration process, and they are likely to be put through recalibration.

Darrow-Kleinhaus also misrepresents the purpose of relative grading, stating that "[t]he objective of the bar exam is not to rank-order examinees for entrance into the profession but to determine whether a particular examinee meets the requirement for minimum competency" (Darrow-Kleinhaus, 2019, p. 177). But rank ordering of examinees is not the end of the grading process; rather, a relative grading approach that uses rank ordering is one step in a process that also includes scaling the written score to the MBE. Scaling the written score to the MBE produces a written score that harnesses the power of the equating done to the MBE. The purpose of relative grading is to make the fairest, most precise, and most stable decisions possible about whether examinees have met the requirement for minimum competency; relative grading is a means of providing uniformity to the grading practices across different essays and jurisdictions. This article from the *Bar Examiner* explains why NCBE recommends relative grading:

> …asking graders to maintain consistent grading standards across administrations, examinees, and items would be extremely difficult, if not impossible. There are simply too many moving parts across test administrations to make such a grading task reasonable for maintaining score meaning across administrations. But relative grading—comparing answers among the current pool of examinees and then scaling those raw scores to the MBE—is manageable for graders and fair to examinees. (Gundersen, 2016, p. 41)

VII. Scaling the Written Scores to the MBE

In producing a UBE score, the MBE is weighted 50% and the written score scaled to the MBE is weighted the other 50%. The Task Force Report's criticism of scaling has two parts. The first argues that there is so much variability across jurisdictions in MBE means that examinees would obtain a different scaled written score for the same performance depending upon which jurisdiction they test in. The second criticism is skepticism that a weighted combination of the scaled written score (which is less reliable than the MBE score) and the MBE score will produce a reliable end result.

Effect of Jurisdiction Autonomy in Grading

The arguments that underpin the Task Force Report's criticism of the scaling process appear to be drawn from the Darrow-Kleinhaus article cited in the Report, which attacks the UBE scoring process for vulnerability to "forum shopping," whereby a savvy candidate, relying on the portable nature of the score, could supposedly game the system by testing in a jurisdiction where the examinee proficiency profile was more favorable than her own and as a consequence would have a different (i.e., higher) score than if she tested in her own original jurisdiction.

It is true that an examinee could get a different raw score on the written portion of the bar exam depending on which jurisdiction she sat in. Given the relatively high correlation between MBE scores and scores on the written portion, if an examinee sits in a jurisdiction with a relatively low mean MBE score, the raw score that examinee receives on the written portion is, indeed, likely to be higher than it would be if she sat in a jurisdiction with a higher MBE mean. However, the fact that different jurisdictions could or would award different raw written scores does not mean that the examinee will ultimately get a different scaled written score and, by extension, a different total UBE score.

Understanding how this can be the case requires careful examination of the scaling formula NCBE uses and a solid working understanding of each component of the formula. However, the Task Force Report, crucially, seems to rely upon an erroneous statement of the formula in Darrow-Kleinhaus's article. To be clear, the formula used by Darrow-Kleinhaus is not the one NCBE uses.

The scaling formula used by NCBE is shown below in two parts: A and B.

$$\begin{matrix} Scaled \\ Written \\ Score \end{matrix} = \overset{\underline{\mathrm{A}}}{\frac{(Written - Mean_{Written})}{SD_{Written}} SD_{MBE}} + \overset{\underline{\mathrm{B}}}{Mean_{MBE}}$$

$Mean_{MBE}$ and SD_{MBE} are, respectively, the mean and standard deviation (SD) of the scaled MBE scores in the jurisdiction. *Written* is the raw written score for a given examinee; $Mean_{Written}$ and $SD_{Written}$ and are, respectively, the raw written score mean and SD in the jurisdiction. In Part A of the formula, the examinee's raw written score is first converted into a *z-score* or standard score, which is a measure of how far the raw written score is from the written score mean. When this number is multiplied by SD_{MBE}, it is given in units of distance from the mean on the MBE scale. Part A thus results in a measure of how far the raw written score is from its mean but expressed on the MBE scale. Adding this result to Part B ($Mean_{MBE}$) then converts the raw written score's distance from its mean to a commensurate distance on the MBE scale.

Darrow-Kleinhaus's erroneous formula is shown below:

$$Scaled\ Written\ Score = SD_{Written}SD_{MBE} + Mean_{MBE}$$

Note that NCBE's formula and Darrow-Kleinhaus's formula differ in the first term in the equation. Most people with statistical training would understand the term $SD_{Written}$ in these formulas to mean the spread of the written scores around the written score mean. In Darrow-Kleinhaus's formula, however, this would not make any sense, because in that case entering the group SD into her formula would result in every single candidate having the same written scaled score. Because this cannot be what she means, and given additional context from the rest of her article, it seems she intends $SD_{Written}$ to describe what NCBE and most quantitative researchers would call a z-score or a standard score.[19]

Even if Darrow-Kleinhaus's general argument were based upon the formula NCBE actually uses, however, it would still fall apart, as illustrated by the following example. An examinee going from a jurisdiction with a high MBE mean to a jurisdiction with a low MBE mean could anticipate receiving a higher raw written score than would be given in the jurisdiction with a high MBE mean and vice versa, even though the $Mean_{Written}$ may be about the same for both jurisdictions. For argument's sake, suppose two jurisdictions with quite different MBE mean scores are willing to separately grade the same set of written materials. They use the same six-point grading scale and produce grades that range from 8 to 48 over the eight parts of the written score (six MEEs, two MPTs). Each jurisdiction has a $Mean_{Written}$ = 24 and an $SD_{Written}$ = 8. The MBE mean for the high jurisdiction is 140 and for the low jurisdiction is 132. Both jurisdictions have SD_{MBE} = 16. When the written materials are graded in the low jurisdiction, a score of 26 is received. In the high jurisdiction, a score of 22 is obtained. The results from the formula are as follows:

$$\text{High Jurisdiction} = 140 + 16\left(\frac{22-24}{8}\right) = 140 + 16\left(-\frac{1}{4}\right) = 140 - 4 = 136$$

$$\text{Low Jurisdiction} = 132 + 16\left(\frac{26-24}{8}\right) = 132 + 16\left(\frac{1}{4}\right) = 132 + 4 = 136$$

As the results from the formula show, the same scaled written score is obtained in both jurisdictions. Clearly, this is a hypothetical example; whether a different scaled score would be obtained if a given set of written materials were graded in different jurisdictions cannot really be tested without having two jurisdictions with different MBE mean scores willing to grade the written material for a group of examinees. But the example illustrates the compensatory systems operating within the scaling formula. It is math, not trickery.

At one point, the Task Force Report says that a truly portable UBE score will not exist until there is pooled grading of the written materials among jurisdictions. Pooled grading could enable more uniformity of the grading process, increasing the reliability of the raw written scores, which would be a good thing. Pooled grading

[19] If this were the only time the term SD appeared in Darrow-Kleinhaus's formula, she could be given the benefit of the doubt. But she uses the term SD_{MBE} properly as the very next term in the formula. This adds to the confusion, however, in that if she intends that SD_{MBE} should be interpreted as an individual z-score or standard score, her formula is again wrong but in a new way. Criticizing a non-statistician's use of technical terms could seem overly critical or harsh, but Darrow-Kleinhaus's misuse of terms is analogous to a criminologist confusing the terms "rehabilitative punishment" and "retributive punishment." To someone unfamiliar with the general concepts, the terms sound similar, and a non-expert could feasibly confuse them. But it would be patently problematic for a court to rely on the expert scholarship of someone who regularly mixed up the terms.

could also deter examinees from doing forum shopping if they think they can game the situation; NCBE cannot say whether some examinees are attempting to forum shop, only note that the math suggests they will not be successful. NCBE agrees with the Task Force that there are good reasons to explore pooled or centralized grading, just not necessarily for the reasons the Task Force raises in their report.

The Task Force Report also makes the argument that because of "anecdotal evidence some applicants are forum shopping … it is *ipso facto* true that the unfair elevation of one or more test takers as a result of the foregoing will result in the failure of that number of test takers that would otherwise have been regarded as having passed the test" (p. 48). Setting aside for the moment that the basis of their "evidence" is anecdotal, the argument rests on the assumption that applicants will be "taking the test with presumptively less able test takers" and "that the system is a closed system and that someone will come up to the line but not cross it" (p. 48). The idea seems to be that these forum shoppers are looking for a jurisdiction of poorly performing examinees so they can let their score on the written portion of the bar examination give them a boost over the locals; and, since they believe it to be a zero-sum game (it is not), that their success will be at the expense of the other examinees. For the shoppers to get the bargain they hope for, they will have to find a jurisdiction of poor performers, for starters. This will not be as easy as it sounds, because bar exam performance is fairly dynamic across jurisdictions at any given administration. Assuming for a minute that the shopper finds such a jurisdiction, though, it is *not* a closed system with one loser for every winner.

If a "rock star" examinee goes to a jurisdiction she expects to be populated with "presumptively less able test takers," as the Task Force puts it, that rock star examinee might raise the MBE average. So, although she might push down the relative performance on the written portion for the "less able test takers," as Dr. Johnson and Ms. Darrow-Kleinhaus anticipate, she will also elevate the whole group in terms of the MBE mean and thus benefit the group as a whole. These two offsets should more or less balance. But the rock star's win will not mean that one of the locals will lose; it is not a zero-sum game. The rock star would have to be a huge outlier with a low MBE score and a high written score to have any possible effect on the locals.

Darrow-Kleinhaus is very critical of NCBE's response to the issue of forum shopping, stating: "What you need to hear, and what you should hear is an emphatic, unequivocal, 'no'—that there is no way that the same person can be found 'competent' to practice law in one UBE jurisdiction and 'incompetent' in another" (Darrow-Kleinhaus, 2019, pp. 174–175). Darrow-Kleinhaus is demonstrating a lack of understanding of measurement principles. Grading is far from an exact science, and the same graders may not be consistent even with themselves over time owing to fatigue, for example. Different graders will also have nuanced differences in their grading criteria, such that one grader may differ from another in terms of what score they would give to answers they grade. Such variation is the price paid for having essay and performance-type components as part of the bar exam. Calibration efforts and re-calibration during grading are intended to keep such variation in grading to a minimum, but such efforts cannot totally make variations go away.

Does that mean that the scores awarded are not uniform? Psychometricians understand that examinees would likely not get the exact same score if they were to take an exam a second time or have their exam graded by a second grader. And some examinees might get a different score such that they would pass with one grader and fail with another grader, particularly borderline examinees. These differences are referred to as measurement error. They aren't errors in the sense that someone made a mistake, but in the sense that there is normally going to be some difference in two instances of scoring the same thing because people are biological entities, not machines. But just because grades awarded by different graders might differ somewhat doesn't mean that the exams are not graded uniformly. As much as possible, the process will be the same in different jurisdictions given the support NCBE provides in the form of grading materials, training and calibration support, and grading workshops. Scaling to the MBE also helps to make the resulting scores as uniform as possible given that grading is done on a jurisdiction-by-jurisdiction basis.

Additionally, because jurisdictions have differing cut scores, it is quite possible that an examinee could pass in one jurisdiction and fail in another even with the same score. However, cut scores do not just define competence, they represent policy decisions that reflect the jurisdiction's relative concern about the risk of passing an incompetent candidate versus failing a competent candidate.

Finally, contrary to the accusation in the Task Force Report that "NCBE is protective of the confidentiality of its scoring practices and appears to consider at least some of its methodologies to be its own intellectual property" (p. 48), NCBE researchers have repeatedly published and presented in lay and professional (educational measurement) settings how NCBE's scaling process works. Albanese (2014) and Case (2005) are two examples where readers can find more information on the scaling process.

Reliability of the Weighted Combination

"NCBE asks us to accept the premise that it is possible to achieve a reliable final score when it is based in part on an unreliable one" (p. 53). This comment from the Task Force Report questions how combining the relatively less reliable scaled written score with the higher-reliability MBE score will result in an overall score that has high reliability. This suggests that the Task Force's experts think reliability is averaged or is the reliability of the lowest score being combined. It isn't.

A better way to think of it is to consider that the written portion of the bar examination is administered in two three-hour sessions, the same time allocated for the MBE. The combination of scores from the MBE and the written portion is essentially equivalent to doubling the testing time. Thinking that total reliability will decrease when adding a less reliable exam component to a highly reliable exam component is akin to worrying that a person's total income will decrease when they add a second job at reduced pay to a better-paid primary job. Of course, adding a second job will not decrease total income; keeping a primary job and adding a second job on top of it should only increase total income, even if the second job is paid at a lower rate than the primary job.

Mathematically, it is a bit more complicated than that, however. The actual formula uses the total scores and not the individual scored units, so unless the MBE is weighted between 60 and 70%, the reliability of the combined score is somewhat lower than the reliability of the MBE alone. NCBE's former Director of Testing showed that with a 50:50 weighting, an MBE reliability of 0.90 and a written score reliability of 0.72 will result in a combined reliability of approximately 0.88 (Case, 2008). The reliability of the MBE since 2018 has exceeded 0.93, so the combined score will almost certainly exceed the 0.90 threshold (Kane and Case, 2004). We refer the Task Force to the following references as examples of the long history of sources that illustrate and explain the well-documented concepts of reliabilities of weighted combinations of scores: Case, 2008; Haertel, 2006; Kane & Case, 2004; Wang & Stanley, 1970; Gulliksen, 1950; Mosier, 1943; Kelley, 1927.

VIII. Correlations Between MBE and Written Scores

The comments on correlations in the Task Force Report show some misconceptions about what a correlation represents, how to interpret a correlation (whether raw or disattenuated), and how the correlation between examination components relates to test reliability.

What a Correlation Represents

The correlation coefficient that NCBE most typically reports represents the relationship between the MBE score and the written score on the bar exam, which are the two variables that are of primary interest in the Task Force Report. The correlation between any two variables has a possible range from -1.0 to 1.0. A correlation of 1.0 means that if examinees were rank ordered by scores on the MBE, their written scores would be in the same order: the examinee with the highest score on the MBE would also have the highest written score, with corresponding values all the way down to the last examinee who would have both the lowest MBE score and the lowest written score. A correlation of -1.0 means that if examinees were rank ordered by their scores on the MBE, their written scores would be ranked in perfect reverse order, with the highest score on the MBE matched to the lowest written score and vice versa. A correlation of 0.0 would indicate that the MBE score had no relationship to the written score.

In reality, the correlation between the MBE score and the written score is not 1.0, but it is closer to 1.0 than 0.0 in most cases. For example, in February 2016, the correlations ranged from 0.51 to 0.67 and averaged 0.60 across UBE jurisdictions. In July 2015, the correlations ranged from 0.44 to 0.81 and averaged 0.66.

(Mis)interpreting Correlations (Raw and Disattenuated)

The Task Force Report erroneously states that correlations between the MBE score and the written score are low, falsely claiming that "NCBE acknowledges that there is a low correlation of the written component score with the MBE scaled score" (p. 50). NCBE made no such acknowledgment in the article referenced by the Task Force. The Report then uses this false information to discredit the validity of scaling the written score to the MBE.

The most widely held criteria in the measurement field for interpreting correlation coefficients are provided in Cohen's (1988) seminal work on computing effect sizes. He describes correlations of 0.1, 0.3, and 0.5 as small, medium, and large associations, respectively. The US Department of Labor Employment and Training Administration has published guidelines for interpreting correlation coefficients in predictive studies in which > 0.35 = "very beneficial," 0.21–0.35 = "likely to be useful," 0.11–0.20 = "depends on circumstances," and < 0.11 = "unlikely to be useful." Even the lowest correlation found for any UBE jurisdiction in February 2016 and July 2015 (0.44) would be considered between a medium and large association by Cohen's criteria and "very beneficial" by the US Department of Labor criterion. The average correlations obtained at those two administrations (0.60 and 0.66) far exceed Cohen's criteria for being a large association and the US Department of Labor standard for being very beneficial.

The Task Force Report also demeans the disattenuated correlations that NCBE has reported both as not being high enough and as being inappropriate to report for a high-stakes examination. (It is common practice in psychometrics to estimate what the correlation would be if one could obtain a perfect measure of the two scores being correlated, a process called disattenuation.) If the MBE is intended to assess examinees' proficiency in understanding core concepts needed for the practice of law, and the written components are intended to assess their understanding of core concepts in a written product, one would expect the underlying constructs to be highly correlated, but not perfectly correlated. If the constructs were perfectly correlated, there would be no need for the written component, and using the MBE alone would be preferable because it would be the most reliable measure of the construct.

Between 2013 and 2019, disattenuated correlations between the MBE scores and the written scores have been at least 0.70 and generally above 0.80. This is almost a "Goldilocks value": high enough, but not too high. It supports the added value of having both the MBE and the written portions as components of the bar exam. Indeed, the abilities measured by the MBE and the written component are not the same, which is why both components are needed; at the same time, they have been shown to be strongly related, which is why it is appropriate to scale the written score to the MBE score.

The Task Force Report was particularly confused about and pointedly harsh in its critique of how correlations have been reported by NCBE over time. The Report raises concerns about whether the written score is in raw or scaled form when correlated with the MBE and is highly suspicious of the practice of disattenuating correlations. The Task Force criticizes NCBE for reporting correlations from the written component after scaling the written raw score to the MBE, saying "[t]his reliance on scaled score correlations is not in keeping with NCBE's own past practices in grader training and workshops where [NCBE's former Director of Testing] presented accurate raw score correlations" (p. 51). A naïve reader could take this to mean that NCBE has more recently reported correlations that are distorted by scaling, while the earlier reported correlations based on raw scores are more accurate. This argument is indicative of the Task Force's pervasive lack of understanding of basic statistics and psychometrics.

This critique is like saying that, globally, some climate scientists are nefariously attempting to mislead the public because they report temperatures in degrees Celsius whereas others use degrees Fahrenheit. Correlation values are primarily an index of the degree to which the two scores rank order examinees the same; correlations are insensitive to linear scaling, which is the form of scaling used within the written score.[20] For the February 2019 bar exam and the July 2019 bar exam, the difference in

[20] An exception might occur if the distributions were wildly different, but in this context they will not be.

correlations across all jurisdictions between the raw and scaled scores was never more than .001. A linear transformation of a variable like a raw written score would have a negligible impact on the correlation with a variable like an MBE score.

The Task Force Report continues to inappropriately criticize NCBE, expressing suspicion around reporting disattenuated correlations that adjust for the unreliability of the two component scores, writing:

> Why would it be an acceptable practice for a "high-stakes" licensing exam to make "adjustments"? Perhaps "reliability adjustments" are what Judith A. Gundersen, NCBE's prior program director for the MEE and the MPT, relies upon in finding a "correlation above .80" between the MBE scaled score and the written components and calling it "strongly correlated." (p. 50)

Again, the naïve reader could take this to mean that NCBE engages in deceptive practices that are unacceptable for a high-stakes licensing examination in order to claim a strong correlation. This, however, is a distortion of the truth. To report a disattenuated correlation without describing it as such would be inappropriate, but the 2016 article by NCBE that is referenced by the Task Force was quite clear about what was reported, as is directly quoted in footnote 259 of the Report. Had the Task Force included a psychometrician, or consulted a psychometrician, this criticism would surely have been dismissed before the Report's release.

How Correlations Relate to Test Reliability

The Task Force Report asks, "How is it possible that an average of 'generally above 0.60' is an acceptable correlation when 0.90 is 'the minimum level normally considered adequate for high-stakes testing purposes'?" (p. 50). The Task Force Report seems to conflate the concepts of reliability—for which 0.90 is the minimum acceptable level, as indicated in the *Bar Examiner* article cited—and correlation. But the criteria used for assessing the utility of correlations and the levels of reliability needed for test scores are two different concepts. As described above, correlations generally quantify the strength of the association between two variables. Often, they are used to support validity arguments for use of a score. For example, the correlation between MBE scores and written scores is used to support the validity argument for the combined score having greater value than either has individually. Reliability, on the other hand, is a measure of the precision of scores. While it can be estimated in various ways in order to determine how much scores are affected by different influences (e.g., internal consistency—effect of different items; stability across time—test-retest, etc.), it is important for scores to have high reliability if they are used to make high-stakes decisions. Whereas correlations are considered useful if they have values above 0.10, the reliability generally deemed necessary for making high-stakes decisions is 0.90 or above.

How Correlations Relate to Equating and Scaling

The Task Force Report states that the disattenuated correlation reported "is lower yet at .80. The evidence indicates that the bar exam's written component and the MBE do not measure the same thing, further supporting the claim that equating written to MBE as the anchor may be a deeply flawed technique" (p. 52). This statement reflects a theme throughout the Report of confusion between what constitutes equating and what constitutes scaling the written score to the MBE. Scaling the written score to the MBE is not equating. Rather, scaling the written score to the MBE can be considered a form of statistically linking the two score scales. MBE scaled scores are equated, and because the correlation between the MBE and the written score is relatively high, linking the two enables the written score to take advantage of MBE equating. Because the written component questions are not reused and all written component grading is done by jurisdictions, it is not feasible to directly equate the written scores; rather than being a "deeply flawed technique," linking the two score scales maximizes fairness and has stood the test of time and professional scrutiny.

The Task Force Report references Kim and Walker (2011), who studied linking mixed-format tests using a multiple-choice anchor and found that "when the correlation between the multiple choice and the written (constructed response items) is relatively low, large differences are seen between groups, and the use of multiple choice anchors is of questionable efficacy" (p. 56). It goes on to explain what the study means with a quote from Dr. Nancy Johnson, who writes that she is "assuming" that NCBE uses the chained equipercentile method for equating. However, the MBE has been equated using IRT since February 2005, and was previously equated using a process called common-item linear equating,[21] neither of which is the same as chained equipercentile equating. While all these methods may be viewed as vehicles for equating, they use entirely different mathematical models that have important practical implications.

Kim and Walker (2011) has limited applicability to what NCBE currently does, because the mixed format test equating that they studied lumps the multiple-choice and constructed response formats together in the IRT calibration, linking, and equating processes. In contrast, equating the MBE and scaling the written score to the MBE are two related yet independent procedures. While MBE scaled scores are obtained through equating, scaled scores on the written component are calculated through a completely independent process. Equating of the MBE is not affected by performance on the written component. Scaling the written score to the MBE does depend on the relationship between the two; the procedure currently used has been justified based on the fact that "the content and concepts assessed on the MBE and written components are aligned and performance on the MBE and the written components is strongly correlated" (NCBE Testing and Research Department, 2017–2018, p. 36).[22]

Despite the fact that the MBE and the written components are scored separately, however, the Kim and Walker study actually provides support for linking the MBE and the written components via scaling as NCBE currently does. Kim and Walker showed that of the different tests studied (mathematics, social studies, science, and English), only one test, English, showed unsatisfactory equating results. The disattenuated correlation between item formats for the English test was in the

[21] For further reading on IRT equating and the equating method used prior to 2005, see Michael T. Kane, PhD, and Andrew A. Mroch, PhD, "Equating the MBE," 74(3) *The Bar Examiner* (August 2005) 22–27.

[22] For more discussion of equating and scaling see, e.g., NCBE Testing and Research Department, "The Testing Column: Q&A: NCBE Testing and Research Department Staff Members Answer Your Questions," 86(4) *The Bar Examiner* (Winter 2017–2018) 34–39.

.30–.40 range; the others, with disattenuated correlations in the .60–.70 range, showed satisfactory results. Correlations between the MBE and the written portion of the bar exam have been found to vary across jurisdictions, but across UBE jurisdictions the disattenuated correlations average 0.79, well above the 0.60–0.70 range that was found to have satisfactory results in the Kim and Walker study.

IX. Equal Weighting of the MBE and the Written Component

The Task Force Report argues that "[a] nswering an essay or a question calling for a short, written answer requires the test taker to contemplate the appropriate answer and then express it clearly and concisely. This is a key skill for lawyers—the knowledge to answer a question and the ability to communicate that answer effectively. The absence of, or lack of weight given to, essay or short answer questions leaves this ability untested" (p. 63). It is unclear why the Task Force seems to think there is a lack of weight given to the written component or why it describes this ability as "untested." The written component is weighted 50%, and it is hard to imagine that examinees might think a full day of testing that produces half their final score would not be worth studying for.

X. New York UBE Study Included Appropriate and Sufficient Data for Analyses

The Task Force Report criticizes the UBE study that NCBE conducted for New York because "the data sample sizes were small: (a) Only two samples were from the prior NYBE [New York bar examination]: July 2015 and February 2016; (b) Only three samples were of the UBE: July 2016 and July 2017; and (c) February results tend to be less stable in general" (pp. 36–37). To clarify, when the Task Force Report talks about data sample size, they are referring to what might also be called data collection points. They were critical of only having data collection at two points prior to the UBE and three points post UBE. The criticism of February results being less stable is a criticism of the nature of the data. Most likely, the instability of the February results is because of the large percentage of repeat takers, usually on the order of 60% or more versus about 30% in July. New York also has many foreign law school graduates who take the bar examination and generally, as a group, have a relatively high rate of failure; thus, they are heavily represented in February as repeat takers, which further contributes to the instability of the results for several of the analyses.

When statisticians (and psychometricians) talk about "sample sizes," they are typically referring to, for example, the number of examinees within the data set for the July 2015 exam administration. The number of examinees within each bar exam administration was large, allowing for adequate analysis of subgroups in most cases (and in cases where sample sizes were small, for example for domestic-educated first-time takers in February for some racial/ethnic groups, care was taken to indicate that the sample sizes were small in the study and caution readers not to overinterpret results). Of course, one can always complain that more data from more years would be better, but there are practical limitations to how much data can be collected, and the study released by the New York Court of Appeals included a huge amount of data from the critical time frame just before and just after the transition to the UBE in New York. Adding more years of data might have placed the results in a larger context but would not necessarily have better answered the question of the impact of UBE adoption in New York.

XI. NCBE's Objectivity in Conducting New York UBE Study

The Task Force Report states, "[w]hile we are cognizant that a three-year study has recently been published as to New York's experience with the UBE, we are concerned by the fact that the study was conducted by NCBE, which is the sponsor of the UBE" (p. 1). The Task Force recommends that "[a]n independent psychometric analysis should be conducted on the grading and scaling of the UBE" (p. 3). NCBE undertook the UBE study at the request of the New York State Board of Law Examiners (BOLE). NCBE staff performed hundreds of hours of work on the study as part of its mission as a nonprofit corporation. NCBE's involvement included providing advice on the design of the study: primarily that there should be data collected pre- and post-UBE, and what data should be collected. The New York Court of Appeals, in collaboration with the BOLE, ultimately approved the actual design of the study. All data from New York were provided by the BOLE. NCBE's role in the study, beyond offering advice on study design, was to analyze the data and prepare the report. NCBE staff put as much of the analysis as possible in the report and appendices for purposes of transparency, but also so that anyone who had questions about the results could likely find data to answer their questions. The report was as clear as possible about all analyses conducted and the reasons why each step was taken in the analysis. A neutral, objective perspective was maintained throughout the study, including the reporting of results such as group differences in average performance. After the report was accepted by New York, all data were sent back to New York and deleted from NCBE's systems.

Considering the inaccuracies pertaining to basic psychometric principles as presented within their critique, the Task Force should have followed its own advice and had an independent assessment of the conclusory statements made by its sources before publishing its report. It is also troubling that one of the experts the Task Force relied upon was a defendant in a lawsuit brought by NCBE for violation of NCBE's copyright of its secure test questions. Dr. Nancy Johnson, whose work is referenced by the Task Force Report, was one of the defendants in a 2003 lawsuit brought by NCBE against a bar preparation company that NCBE accused of obtaining copyrighted materials through student recollections and posting them on their website. The court ruled in favor of NCBE and entered a consent judgment and permanent injunction.[23]

XII. Conclusion

NCBE's typical approach in reporting the results of our research or commenting on the research of others has been one of measured discussion of facts and concepts as free as possible from editorializing. And historically, NCBE's research/psychometric staff have tended to give critics of the bar examination the benefit of the doubt, assuming that perhaps these critics simply did not understand nuances or even key points of what was said or written, despite all the educational efforts NCBE makes each year, including professional meetings,

[23] The court ruling can be found at https://casetext.com/case/national-conference-of-bar-examiners-v-saccuzzo.

workshops, and publications. Why the Task Force did not clarify key points with NCBE or others with relevant professional expertise is baffling.

Fairness, integrity, and respect for others are core values at NCBE. Normally NCBE does not directly respond to criticisms of this nature, in the belief that others are entitled to their own opinions. But in this case, the report of the Task Force is so deeply flawed and so profoundly unfair that it left no choice but to respond directly and bluntly. The BOLE or anyone else who reads the Task Force Report should be skeptical of the information and recommendations contained therein.

In closing, we note that the Task Force Report offers several possible alternatives to the current UBE, including refinements, step exams, and exam models from other countries and states. NCBE has not responded to that portion of the Report but rather encourages readers to review the information and reports regarding the three-year comprehensive study NCBE's Testing Task Force is currently undertaking to redesign the bar examination. As NCBE's Testing Task Force embarks on the final year of its study, it is considering ideas and perspectives from all stakeholders and actively involving stakeholders in the process of developing recommendations for the next generation of the bar examination.

Principal authors:

Mark Albanese, PhD,
Director of Testing and Research

Juan Chen, PhD,
Senior Research Psychometrician

Mark Connally, PhD,
Principal Research Psychometrician

Joanne Kane, PhD,
Associate Director of Testing

Andrew Mroch, PhD,
Senior Research Psychometrician

Mark Raymond, PhD,
Director of Assessment Design and Delivery

Douglas Ripkey, MS,
Deputy Director of Testing

Mengyao Zhang, PhD,
Research Psychometrician

Editorial support provided by:

Kellie Early,
Chief Strategy Officer

Claire Guback,
Editorial Director

Valerie Hickman,
Communications Coordinator

Bibliography

Albanese, M. A. (2014, December). The Testing Column: Scaling: It's Not Just for Fish or Mountains. *The Bar Examiner, 83*(4), pp. 50–56. Retrieved from https://thebarexaminer.org/wp-content/uploads/PDFs/830414-testingcolumn.pdf.

Albanese, M. A. (2015, September). The Testing Column: Equating the MBE. *The Bar Examiner, 84*(3), pp. 29–36. Retrieved from https://thebarexaminer.org/wp-content/uploads/PDFs/BE-Sept2015-TheTestingColumn.pdf.

Albanese, M. A. (2016, December). The Testing Column: Essay and MPT Grading: Does Spread Really Matter. *The Bar Examiner, 85*(4). Retrieved from https://thebarexaminer.org/article/december-2016/the-testing-column-essay-and-mpt-grading-does-spread-really-matter/.

Albanese, M. A. (2016, September). The Testing Column: Let the Games Begin: Jurisdiction-Shopping for the Shopaholics (Good Luck With That). *The Bar Examiner, 85*(3), pp. 51–56. Retrieved from https://thebarexaminer.org/wp-content/uploads/PDFs/BE-TestingColumn-850316.pdf.

ALM Staff. (2020, May). Pay Cuts, Layoffs, and More: How Law firms Are Managing the Pandemic. *Law.com: The American Lawyer*. Retrieved from https://www.law.com/americanlawyer/2020/04/20/pay-cuts-layoffs-and-more-how-law-firms-are-managing-the-pandemic/?slreturn=20200511134113.

American Psychological Association. (2012). *Ethics and Racial Disparities in Education: PSychology's Contribution to Understanding and Reducing Disparities*. Retrieved May 27, 2020, from https://www.apa.org/ed/resources/racial-disparities.

Bolus, R. (2018, December). *Performance Changes on the California Bar Examination, Part 2: New Insights from a Collaborative Study with California Law Schools*. Retrieved from https://www.calbar.ca.gov/Portals/0/documents/admissions/Examinations/Bar-Exam-Report-Final.pdf.

Bosse, D. F. (2016, September). A Uniform Bar Examination: The Journey From Idea to Tipping Point. *The Bar Examiner, 85*(3), pp. 19–23. Retrieved from https://thebarexaminer.org/wp-content/uploads/PDFs/BE-Bosse-850316.pdf.

Brennan, R. L. (1984). Scoring and Combining of MBE and Essay Questions. *NCBE Resource Handbook. Bar Examinations: The State of the Art*, pp. 67–90.

Brennan, R. L. (1986, April). Scoring and Combining of MBE and Essay Questions. *NCBE Seminar Resource Handbook*. Chicago.

California Attorney Practice Analysis Working Group. (2020). *The Practice of Law in California: Findings from the California Attorney Practice Analysis and Implications for the California Bar Exam.* Retrieved from http://board.calbar.ca.gov/Agenda.aspx?id=15573&tid=0&show=100024743#10032701.

Case, S. M. (2005, May). The Testing Column: Demystifying Scaling to the MBE: How'd You Do That? *The Bar Examiner, 74*(2), pp. 45–46. Retrieved from https://thebarexaminer.org/wp-content/uploads/PDFs/740205-testing.pdf.

Case, S. M. (2006, November). The Testing Column: Frequently Asked Questions About Scaling Written Test Scores to the MBE. *The Bar Examiner, 75*(4), pp. 42–44. Retrieved from https://thebarexaminer.org/wp-content/uploads/PDFs/750406-Testing.pdf.

Case, S. M. (2008, February). The Testing Column: Best Practices with Weighting Examination Components. *The Bar Examiner, 77*(1), pp. 43–46. Retrieved from https://thebarexaminer.org/wp-content/uploads/PDFs/770108_testing.pdf.

Case, S. M. (2012, June). The Testing Column: What Everyone Needs to Know About Testing, Whether They Like It or Not. *The Bar Examiner, 81*(2), pp. 29–31. Retrieved from https://thebarexaminer.org/wp-content/uploads/PDFs/810212_be_TestingColumn.pdf.

Case, S. M., & Ripkey, D. R. (2005, April). Accountability in the Licensing of Lawyers: And the Verdict Is . . .? *Annual Meeting of the American Educational Research Association*. Montreal, Quebec, Canada.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Lawrence Erlbaum Associates. Retrieved from http://www.utstat.toronto.edu/~brunner/oldclass/378f16/readings/CohenPower.pdf.

Conference of Chief Justices. (2016). *Resolution 10.* Retrieved from http://www.ncbex.org/pdfviewer/?file=%2Fdmsdocument%2F194.

Darrow-Kleinhaus, S. (2004). A Response to the Society of American Law Teachers Statement on the Bar Exam. *Digital Commons @ Touro Law Center.* Retrieved from https://digitalcommons.tourolaw.edu/cgi/viewcontent.cgi?article=1090&context=scholarlyworks.

Darrow-Kleinhaus, S. (2016, March). UBE-Shopping: An Unintended Consequence of Portability? *Touro Law Center Legal Studies Research Paper Series No. 16-14*. Retrieved from http://ssrn.com/abstract=2756520.

Darrow-Kleinhaus, S. (2019). A Reply to the National Conference of Bar Examiners: More Talk, No Answers, So Keep on Shopping. *Ohio Northern University Law Review, 44*(2), 173–202. Retrieved from https://digitalcommons.onu.edu/onu_law_review/vol44/iss2/1.

Duhl, S. (1980). *The Bar Examiner's Handbook (2nd ed.).* Chicago: The National Conference of Bar Examiners.

Gopnik, A. (2017). *Are Liberals on the Wrong Side of History?* Retrieved May 2020, from New Yorker: https://www.newyorker.com/magazine/2017/03/20/are-liberals-on-the-wrong-side-of-history.

Gulliksen, H. (1950). *Theory of Mental Tests.* John Wiley & Sons Inc. Retrieved from https://doi.org/10.1037/13240-000.

Gundersen, J. A. (2016, June). It's All Relative —MEE and MPT Grading, That is. *The Bar Examiner, 85*(2). Retrieved from https://thebarexaminer.org/article/june-2016/its-all-relative-mee-and-mpt-grading-that-is-2/.

Haertel, E. H. (2006). Reliability. In R. L. Brennan, N. C. Education, & A. C. Education, *Educational Measurement* (Vol. 4, pp. 65–110). Westport, CT: Praeger Publishers.

Harris, D. (2003, August). Equating the Multistate Bar Examination. *The Bar Examiner, 72*(3), pp. 12–18. Retrieved from https://thebarexaminer.org/wp-content/uploads/PDFs/720303-harris.pdf.

Kane, M. T. (2009, November). Reflections on Bar Examining. *The Bar Examiner, 78*(4), pp. 6–20. Retrieved from https://thebarexaminer.org/wp-content/uploads/PDFs/780409_Kane.pdf.

Kane, M. T., & Case, S. M. (2010). The Reliability and Validity of Weighted Composite Scores. *Applied Measurement in Education*, 221–240.

Kane, M. T., & Mroch, A. (2005, August). Equating the MBE. *The Bar Examiner, 74*(3), pp. 22–27. Retrieved from https://thebarexaminer.org/wp-content/uploads/PDFs/740305-Kane-and-Mroch.pdf.

Kane, M. T., Case, S. M., Mroch, A. M., & Ripkey, D. R. (2007, April). The Psychometric Properties of Multi-Component, High-Stakes Assessments Employing Compensatory Scoring; Bar Examinations and Law-School GPAs. *Annual Meeting of the National Council on Measurement in Education*. Chicago, Illinois.

Kane, M. T., Case, S. M., Ripkey, D. R., Mroch, A. A., & Bonner, S. M. (2005, April). Evaluating Potential Threats to the Validity of Licensure Examinations. *Annual Meeting of the National Council on Measurement in Education*. Montreal, Quebec, Canada.

Kelley, T. L. (1927). *Interpretation of Educational Measurements.* World Book Co.

Kim, S., & Walker, M. E. (2011). *Does Linking Mixed-Format Tests Using a Multiple-Choice Anchor Produce Comparable Results for Male and Female Subgroups?* Research Report, ETS. Retrieved from https://files.eric.ed.gov/fulltext/ED528981.pdf.

Klein, S. (1995, November). Options for Combining MBE And Essay Scores. *The Bar Examiner,* 64(4), pp. 38–40.

Klein, S. P. (1979). Are Your Test Scores Only Half Safe? *The Bar Examiner,* 48(2), pp. 137–148.

Klein, S. P., & Bolus, R. (1997, November). The Size and Source of Differences in Bar Exam Passing Rates Among Racial and Ethnic Groups. *The Bar Examiner,* 66(2), pp. 8–16. Retrieved from https://thebarexaminer.org/wp-content/uploads/PDFs/660497-Klein-Bolus.pdf.

Kolen, M. J., & Brennan, R. L. (2014). *Test Equating, Scaling, and Linking.*

Lenel, J. C. (1992, May). Issues in Equating and Combining MBE and Essay Scores. *The Bar Examiner,* 61(2), pp. 6–20.

Miles, V. V. (2016, September). Marketable and Mobile: UBE Recommended. *The Bar Examiner,* 85(3), pp. 27–29. Retrieved from https://thebarexaminer.org/wp-content/uploads/PDFs/BE-Miles-850316.pdf.

Mosier, C. I. (1943). On the Reliability of a Weighted Composite. *Psychometrika,* 161–168.

Muller, D. T. (2020, April 3). *When the Task Force on the New York Bar Examination plagiarizes your work without attribution*. Retrieved from Excess of Democracy: https://excessofdemocracy.com/blog/2020/4/when-the-task-force-on-the-new-york-bar-examination-plagiarizes-your-work-without-attribution.

Muller, D.T. (2015, September 28). *No, the MBE was not "harder" than usual*. Retrieved from Excess of Democracy: https://excessofdemocracy.com/blog/2015/9/no-the-mbe-was-not-harder-than-usual

Murphy, G. G., & Thiem, R. S. (2016, September). Co-Chairing the UBE Committee: A Labor of Love. *The Bar Examiner, 85*(3), pp. 24–26. Retrieved from https://thebarexaminer.org/wp-content/uploads/PDFs/BE-MurphyThiem-850316.pdf.

NCBE (2016, September). The UBE From Early Concept to the Present: A Timeline. *The Bar Examiner,* 85(3). Retrieved from https://thebarexaminer.org/wp-content/uploads/PDFs/BE-UBETimeline-850316.pdf.

NCBE. *The Uniform Bar Exam*. Retrieved from National Conference of Bar Examiners: http://www.ncbex.org/exams/ube/.

NCBE Testing and Research Department. (2017–2018, Winter). The Testing Column: Q&A: NCBE Testing and Research Department Staff Members Answer Your Questions. *The Bar Examiner,* 86(4). Retrieved from https://thebarexaminer.org/article/winter-2017-2018/the-testing-column-qa-ncbe-testing-and-research-department-staff-members-answer-your-questions/.

NYSBA Task Force. (2020). *Report of the NYSBA Task Force on the New York Bar Examination.*

Olson, S. (2019-2020). 13 Best Practices for Grading Essay and Performance Tests. *The Bar Examiner,* 88(4). Retrieved from https://thebarexaminer.org/article/winter-2019-2020/13-best-practices-for-grading-essays-and-performance-tests/.

Ripkey, D. R., & Case, S. M. (2012, April). Assessment Challenges in Creating The Uniform Bar Examination: The Three P's - Politics, Psychometrics, and Practicality. *Annual Meeting of the American Educational Research Association*. Vancouver, British Columbia, Canada.

Ripkey, D. R., & Kane, J. E. (2016). Assessment Challenges in Creating the Uniform Bar Examination: Politics, Practicality and Psychometrics. In P. F. Wimmers, & M. Mentkowski, *Assessing Competence in Professional Performance Across Disciplines and Professions* (pp. 427–446). Springer International.

Schroeder, L. (2000, February). Scoring Examinations: Equating and Scaling. *The Bar Examiner,* 69(1), pp. 6–9. Retrieved from https://thebarexaminer.org/wp-content/uploads/2018/10/690100-Schroeder.pdf.

Smith, J. (2000, February). Testing, Testing: How is Scaling Accomplished? *The Bar Examiner,* 69(1), pp. 52–53.

The American Bar Association (n.d.). Retrieved from http://www.ncbex.org/pdfviewer/?file=%2Fdmsdocument%2F193.

Traub, R. E. (1994). *Reliability for the Social Sciences.* Sage Publications.

U.S. Department of Labor Employment and Training Administration (1999). Testing and Assessment: An Employer's Guide to Good Practices. Retrieved from https://wdr.doleta.gov/opr/FULLTEXT/99-testassess.pdf.

Wang, M. W., & Stanley, J. C. (1970). Differential Weighting: A Review of Methods and Empirical Studies. *Review of Educational Research*, 663–705.

Weiss, D. C. (2020, April). Pay Cut and Furloughs Continue As More Firms Trim Costs to Address COVID-19. *ABA Journal*. Retrieved from https://www.abajournal.com/news/article/pay-cut-and-furlough-juggernaut-continues-as-more-law-firms-trim-costs-to-address-covid-19.

Zaretsky, S. (2020, April). Am Law 200 Firm Puts Its Employees On Ice With Furloughs, Salary Cuts. *Above the Law*. Retrieved from https://abovethelaw.com/2020/04/am-law-200-firm-puts-its-employees-on-ice-with-furloughs-salary-cuts/?rf=1.