TOURO LAW
JACOB D. FUCHSBERG LAW CENTER

Digital Commons @ Touro Law Center

Scholarly Works | Faculty Scholarship

2017

A Reply to the National Conference of Bar Examiners: More Talk, No Answers, so Keep on Shopping

Suzanne Darrow-Kleinhaus
Touro Law Center, sdarrow-kleinhaus@tourolaw.edu

Follow this and additional works at: https://digitalcommons.tourolaw.edu/scholarlyworks

 Part of the Legal Education Commons

Recommended Citation

Darrow-Kleinhaus, Suzanne, "A Reply to the National Conference of Bar Examiners: More Talk, No Answers, so Keep on Shopping" (2017). *Scholarly Works*. 632.
https://digitalcommons.tourolaw.edu/scholarlyworks/632

This Article is brought to you for free and open access by the Faculty Scholarship at Digital Commons @ Touro Law Center. It has been accepted for inclusion in Scholarly Works by an authorized administrator of Digital Commons @ Touro Law Center. For more information, please contact lross@tourolaw.edu.

A Reply to the National Conference of Bar Examiners: More Talk, No Answers, so Keep on Shopping

Suzanne Darrow-Kleinhaus[1]

I. Introduction

In *Let the Games Begin: Jurisdiction-Shopping for the Shopaholics (Good Luck With That)* Mark Albanese defends the National Conference of Bar Examiners' grading practices as essential to assuring reliability given the variability in grading between UBE jurisdictions.[2] In addressing the claim that it is possible to achieve different outcomes on the same test by the same candidate if taken in different UBE jurisdictions, he describes how NCBE monitors jurisdiction variation to ensure grading consistency.[3] Those of us concerned, however, with the possibility that the jurisdiction in which a candidate takes the Uniform Bar Examination (UBE)[4] may make the difference between passing and

---

[1] Professor Darrow-Kleinhaus is the Director of Academic Development and Bar Programs at Touro College, Jacob D. Fuchsberg Law Center. In addition to books on law school learning and the bar exam, including MASTERING THE LAW SCHOOL EXAM, THE BAR EXAM IN A NUTSHELL, ACING THE BAR EXAM, and THE NEW YORK BAR EXAM BY THE ISSUE, she has written law review articles in this area, *A Response to the Society of American Law Teachers' Statement on the Bar Exam* and *Incorporating Bar Pass Strategies into Routine Teachings Practices*. She has also published in the areas of contract law, labor and employment law, the Fourth Amendment Exclusionary Rule, and federal preemption.

Dr. Nancy Johnson, Ph.D., J.D.: As with the writing of *UBE-Shopping: An Unintended Consequence of Portability?,* a special acknowledgement is due to Dr. Johnson for her generosity of time and knowledge. Her expertise in statistical analysis and the bar exam were essential to the writing of this paper. Dr. Johnson is a California attorney, currently in private practice. She is also a licensed psychologist in clinical psychology. She is the author or co-author of over 150 peer-reviewed publications and papers, including more than 20 law study guidebooks and an interdisciplinary book that critically reviewed the 1998 empirical literature on domestic violence. In conjunction with Dennis P. Saccuzzo, she has lectured extensively in law and has taught a variety of supplemental and full Bar programs for first time and repeating Bar candidates. She is co-founder of Applications of Psychology to Law, Inc., a corporation devoted to the application of the psychological sciences to the study of law.

A very special thank you to my friend, Irene Crisci, Interim Director of the Gould Law Library, Head of Public Services, who provides endless enthusiasm and support, both technical and personal.

[2] Mark A. Albanese, *The Testing Colum, Let the Games Begin: Jurisdiction-Shopping For the Shopaholics (Good Luck With That)* THE BAR EXAMINER, Sept. 2016, at 50, 52 [hereinafter Albanese, Let the Games Begin]

[3] *Id*. at 52-53. (See sections entitled "The Reliability of the Written Component Total Score", and "The Correlation of the Written Component Score with the MBE Scaled Score").

[4] *UBE: Uniform Bar Examination*, NATIONAL CONFERENCE OF BAR EXAMINERS, http://www.ncbex.org/exams/ube/ (last visited Jan. 2, 2017). The National Conference of Bar Examiners (NCBE) develops and sells the Multistate Bar Examination (MBE), the Multistate Essay Examination (MEE) and the Multistate Performance Test (MPT), to jurisdictions. The MBE is a multiple-choice exam with 200 questions (only 175 are "live") testing examinees' knowledge of Civil Procedure, Constitutional Law, Contracts and UCC Article 2, Criminal Law and Procedure, Evidence, Real Property, and Torts. The MEE includes essay questions covering these MBE subject and five

failing will not find Mark Albanese's explanations satisfactory. Contradictory and confusing perhaps, but not responsive. Rather than answer the question of whether it is possible for the same person to be found "competent" to practice law in one UBE jurisdiction and "incompetent" in another when it is the same person with the same skill level writing the same exam,[5] NCBE deflects and disguises — and despite a lot of words and numbers — avoids it completely.

The process of deciding in which jurisdiction to take the bar exam — the high-stakes licensing exam that determines whether an examinee will be able to practice law — should not be like shopping around for the best deal on a car, a refrigerator, or a new pair of shoes. While the bargain principle is part of the American economy and culture, it has no place in determining an individual's admission to the practice of law. Instead of saying that it simply isn't so — that where you take the bar exam cannot make a difference between passing and failing — the National Conference of Bar Examiners says, in effect, to give it your best shot.[6]

This is not what one wants to hear from the entity that develops and coordinates the Uniform Bar Examination (UBE) and the licensing tests used by most United States jurisdictions for admission to the bar.[7] 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. This result is fundamentally unfair when passing the exam that controls the gateway to the profession depends not so much on the individual's performance as on the group against whom the individual is evaluated. In this case, the individual may be denied a law license on grounds other than a determination of individual competency.

---

additional areas. The MPT consists of two performance tasks where examinees complete "lawyerly" assignments using the material from the provided Law Library and Client File.

[5] Suzanne Darrow-Kleinhaus, *UBE-Shopping: An Unintended Consequence of Portability?* (March 30, 2016) Touro Law Center Legal Studies Research Paper Series No. 16-14, http://ssrn.com/abstract=2756520.

[6] Albanese, *supra* note 2.

[7] *About NCBE: What NCBE Does*, NATIONAL CONFERENCE OF BAR EXAMINERS, http://www.ncbex.org/about/ (last visited March 13, 2017). NCBE develops the Multistate Bar Examination (MBE), the Multistate Essay Examination (MEE), and the Multistate Performance Test (MPT)

This reply identifies the significant flaws in Dr. Albanese's defense of NCBE's scoring practices. These practices include standardizing the written scores to the subset of MBE scores that come only from that jurisdiction and standardizing written scores to multiple choice scores. The way that the written raw scores are standardized is itself a problem for two reasons and will be addressed. In so doing, the underlying question becomes clear: why would NCBE and Dr. Albanese defend admittedly defective practices — practices that are antithetical to the bar exam's objective of determining an individual's minimum competency for the practice of law?

II. It's the grading practices, not the graders.

A. Relative-grading is antithetical to the individual.

The thorny issue raised by forum shopping is not about the quality of the grading materials or the graders[8] but about particular grading processes, especially the rank ordering of papers. Rank-ordering occurs when graders make grading distinctions among papers where the "top grade does not necessary indicate an excellent paper; it just indicates a paper that is better than the other papers."[9] Graders sort papers into piles or buckets "according to their relative strength…."[10] For example, assuming that the scoring scale 1—6 is used in a jurisdiction, then a score of 6 goes to the best papers among all the answers assigned to that particular grader and they go into the "6 bucket." These papers are "better" than those that go into the "5 bucket," which are, in turn, better than those placed in the "4 bucket" and so forth down the line to the "1 bucket" which contains the weakest papers. ***NCBE defends a practice where it has been shown that "an***

[8] Judith A. Gundersen, *It's All Relative MEE and MPT Grading, That Is*, THE BAR EXAMINER, June 2016, at 37. I have attended information sessions at NCBE headquarters in Madison, Wisconsin for ASP professionals where the process for "calibrating the graders" was explained. It was clear from the presentation that every effort is made to ensure that the assignment of raw scores is conscientious.

[9] Susan M. Case*, The Testing Column, Quality Control for Developing and Grading Written Bar Exam Components,* THE BAR EXAMINER, June 2013 at 36 [hereinafter Case, *Quality Control*] http://ncbex.org/assets/media_files/Bar-Examiner/articles/2013/820213Testing-Column.pdf.

[10] Gundersen, *supra* note 8, at 38.["Relative grading means that in any group of answers, even if no single paper addresses all the points raised in an item, the strongest papers still deserve a 6 (using a 1-6 score scale)."]

essay of average proficiency will be graded lower if it appears in a pool of excellent essays than if it appears in a pool of poor essays. Context matters."[11]

A major problem with rank ordering is that graders change the score that the examinee earned to make it fit "whatever score scale the jurisdiction has in place."[12] Consider the following grading scenario: after reading a set of examinee answers and assessing them according to the grading materials, 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 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.

NCBE's defense of relative grading for a test for individual competency makes little sense. "Context" has no role when it comes to determining whether an individual possesses the minimum competency necessary for the practice of law. One is not competent in relation to another but whether one has demonstrated the requisite mastery of core concepts and skills. Even assuming that grading on a curve is acceptable in law school where norm-referenced tests are "designed to separate out levels of learning within a group",[13] it has no place in a licensing exam. While law school exams are graded on a curve to sort competencies for law review, clerkships, and law firm placements, the bar exam is not about sorting competencies but determining them. The objective of the bar exam is not to rank-order examinees for entrance into the profession but to determine

[11] Susan M. Case, *The Testing Column, Frequently Asked Questions About Scaling Written Test Scores to the MBE,* THE BAR EXAMINER, Nov. 2006, at 43, http://www.ncbex.org/pdfviewer/?file=%2Fassets%2Fmedia_files%2FBar-Examiner%2Farticles%2F2006%2F750406_Testing.pdf.

[12] Gundersen, *supra* note 8, at 38.

[13] Lynn M. Daggett, *All of the Above: Computerized Exam Scoring of Multiple Choice Items Helps To: (A) Show How Exam Items Worked Technically, (B) Maximize Exam Fairness, (C) Justly Assign Letter Grades, and (D) Provide Feedback on Student Learning,* 57 J. OF LEGAL EDUC. 399 (2007).

whether a particular examinee meets the requirements for minimum competency.[14] When an examinee meets the criteria set for minimum competency, then it is irrelevant how other examinees performed. If an examinee is not minimally competent to practice law, then being bumped up into the next bucket because of a need to fill buckets does not make him so — and vice versa. If an examinee is competent, then being downgraded to a lower bucket does not make him incompetent. This system perverts the process and harms the individual.

In March, 2016, Oklahoma changed its scoring model when it recognized that some examinees may have failed the bar exam when they should have passed. Oklahoma Supreme Court Chief Justice John F. Reif explained that the court's decision came after learning that several individuals who took the bar exam in July 2015 may have passed if it were not for the scaling system. According to Judge Reif, "[o]nce the scaling and adjustment took place, they no longer had a passing grade." Further, "it didn't have anything to do with what they had demonstrated in the way of knowledge on the essay portion. It happened to be the scaling of that score brought it below the passing grade."[15] Still, the Oklahoma Board of Bar Examiners remained firm in its commitment to the scaling system. Board member Donna Smith told the Tulsa World that "scaling was meant 'to take out the bumps in the road' when the difficulty of essay questions or skill level of test-takers vary."[16] Then she added: "If you have 10 really good papers and then a paper that's more average, generally that average paper will get a lower score if it's graded among really good papers, and vice-versa."[17] This statement is very telling: it shows that bar examiners are very well aware that examinees fail the bar exam not because they lack the requisite knowledge, but because they appear weaker when in the company of stronger candidates — and that they nonetheless find the practice acceptable.

---

[14] Case, *Quality Control supra* note 9, at 35. Dr. Case states that "the bar exam is developed to assess the extent to which each examinee has the knowledge and skills that are required of newly licensed lawyers."

[15] Arianna Pickard, *High failure rate on Oklahoma bar exam prompts change to state test*, TULSA WORLD (Apr. 8, 2016, 12:00 AM and updated on Apr. 11,2016 at 11:49 AM)http://www.tulsaworld.com/news/education/high-failure-rate-on-oklahoma-bar-exam-prompts-change-to/article_b73f0e5b-0075-531a-a39e-f06e8ecad0ee.html.

[16] *Id.*

[17] *Id.*

Although Oklahoma is not a UBE jurisdiction, it recognized the serious problem with relative grading and took action. The problem is only magnified when it occurs in a UBE jurisdiction because the UBE score is "portable" — it is supposed to mean the same thing from one jurisdiction to another.[18]

B. Scaling the written component to the MBE exacerbates the harm when there is a low correlation between the components and, as a result, undermines scoring reliability.

Compounding the problem with rank-ordering is NCBE's practice of scaling the written component to the MBE.[19] NCBE defends this practice with explanations that not only defy credibility but emphasize the problem.

Let's begin with the practice of standardizing the written scores to the subset of MBE scores that come only from that jurisdiction. According to Dr. Nancy Johnson, this practice could result in the case where an examinee who writes her essays in New York might get a different score than if she had written the same answers in a state with a lower MBE mean. This could happen because the examinee's written "raw scores [are] forced into a distribution that compares them to the [examinee] pool in New York, whereas if [the examinee] had written the same answers in a state with a lower MBE mean, their scaled written score could then be different. That [i]s a problem, because that scaled score is

[18] Susan M. Case, *The Testing Column, The Uniform Bar Examination: What's In It For Me?* THE BAR EXAMINER, Feb. 2010, at 51, [hereinafter Case, *What's In It For Me?*] http://www.ncbex.org/pdfviewer/?file=%2Fassets%2Fmedia_files%2FBar-Examiner%2Farticles%2F2010%2F790110_TestingColumn.pdf.
See also UBE Score Portability, NATIONAL CONFERENCE OF BAR EXAMINERS, http://www.ncbex.org/exams/ube/score-portability/ (last visited Mar. 6, 2017). NCBE advises jurisdictions that because every UBE jurisdiction uses the same essay questions, the same performance tasks, and the same grading guidelines, as long as the candidate sits for all portions of the UBE in the same UBE jurisdiction and in the same administration, a portable UBE score is earned that can then be transferred to other states that have joined the UBE network.

[19] *See* Susan M. Case, The Testing Column, Demystifying *Scaling To the MBE: How'd You Do That?*, THE BAR EXAMINER, May 2005, at 46 [hereinafter Case, *Demystifying Scaling To the MBE*], According to Dr. Susan Case, former Director of Testing for the National Conference of Bar Examiners, "[s]caling the essays to the MBE is an essential step in ensuring that scores have a consistent meaning over time. When essay scores are not scaled to the MBE, they tend to remain about the same: for example, it is common for the average raw July essay score to be similar to the average February score even if the July examinees are known to be more knowledgeable than the February examinees. Using raw essay scores rather than scaled essay scores tends to provide an unintended advantage to some examinees and an unintended disadvantage to others." Dr. Case was the Director of Testing until Nov. 1, 2013.

supposed to be portable — it's supposed to mean the same thing from one [UBE] jurisdiction to the next."[20]

While this is a problem, no one knows exactly how big a problem it is because we don't have the information that is necessary to make a determination. Instead of answering the question whether there can be a difference in the outcome if a candidate takes the July UBE in North Dakota or Missouri, NCBE leaves us in the dark. North Dakota and Missouri have the same cut score of 260 — as does New Mexico and Alabama. What's different, of course, is the size and cohort strength of the test-takers in each jurisdiction.[21] If what NCBE tells us is correct — that all UBE scores "have the same meaning across the country"[22] because the UBE is "uniformly administered, graded, and scored by the jurisdictions that adopt it"[23] — then the same person with the same skill level writing the exam in North Dakota should get the same result in Missouri, New Mexico or Alabama. If our examinee is found "competent" in North Dakota, then she should be found competent in another jurisdiction with the same exam and cut score. It cannot be otherwise or the test is neither reliable nor uniform. But it may indeed be possible for our

[20] E-mail from Nancy E. Johnson to Suzanne Darrow-Kleinhaus, Professor of Law and Director of Academic Development and Bar Programs, Touro Law Center (Aug. 20, 2016, 11:46 a.m. EST) (on file with author).
[21] North Dakota had 78 examinees take the UBE in July 2016; *North Dakota Bar Exam Results*, STATE BAR ASSOCIATION OF NORTH DAKOTA, https://www.sband.org/News/NewsDetail.aspx?NewsId=769 (last visited Mar. 10, 2017). Missouri had 685 examinees take the UBE in July 2016; *July 2016 Bar Examination in Missouri*, SUPREME COURT OF MISSOURI, OFFICE OF THE BOARD OF LAW EXAMINERS, https://www.mble.org/exam-web-stats-july-2016 (last visited Mar. 10, 2016). In New Mexico the exact number of examinees is not available for the July 2016 administration of the UBE; however, the University of New Mexico School of Law reports that there were 114 first time takers for the February and July 2013 bar exams, 110 first time takers for the February and July 2014 bar exams, and 107 first time takers for the February and July 2015 bar exams; *University of New Mexico School of Law – 2016 Standard 509 Information Report,* ABA SECTION OF LEGAL EDUCATION AND ADMISSION TO THE BAR, AMERICAN BAR ASSOCIATION, www.abarequireddisclosures.org (last visited March 10, 2017). Alabama had 483 examinees take the UBE in July 2016. *The Alabama State Bar, Examinee Statistics for July,* https://www.alabar.org/assets/uploads/2016/09/July2016-DetailedStatistics-1.pdf (last visited Mar. 10, 2017).
[22] Case, *What's In It For Me?, supra* note 18, at 51. *See also UBE Score Portability,* NATIONAL CONFERENCE OF BAR EXAMINERS, http://www.ncbex.org/exams/ube/score-portability/ (last visited Mar. 6.2017).
[17] Case, *What's In It For Me?, supra* note 18, at 51. NCBE claims that the UBE provides the consistency essential for comparisons between jurisdictions of examinees' competency because all UBE examinees "will be taking exactly the same exam and receiving scores that will have the same meaning across the country." *See also UBE Score Portability,* NATIONAL CONFERENCE OF BAR EXAMINERS, http://www.ncbex.org/exams/ube/score-portability/ (last visited Mar. 6,, 2017). NCBE advises examinees that because every UBE jurisdiction uses the same essay questions, the same performance tasks, and the same grading guidelines, as long as the candidate sits for all portions of the UBE in the same UBE jurisdiction and in the same administration, a portable UBE score is earned that can then be transferred to other states that have joined the UBE network.

examinee to pass the UBE in North Dakota and not pass in the other jurisdictions — all because of the way that the exam is scaled and scored.

We know that this result may be possible because NCBE has told us so, not directly, but by the "numbers" that slip out when it attempts to defend its practices. For example, NCBE acknowledges that there is a low correlation of the written component score with the MBE scaled score and that this correlation varies widely across the UBE jurisdictions.[24] Nonetheless, NCBE assures us that despite a "low correlation", when "these correlations are adjusted for their less-than-perfect reliability, they are generally above 0.60, indicating that the MBE and written components 'assess some shared aspects of competency, and that each method also assesses some unique aspect of competency.'"[25] How is it even possible that an average of "generally above 0.60" is an acceptable correlation when we are told that 0.90 is "the minimum level normally considered adequate for high-stakes testing purposes"?[26]

It is only reasonable to ask why NCBE would insist on scaling the written component scores to the MBE scaled score to achieve reliability when NCBE admits that the written component score is unreliable and there is a low correlation between the written component score and the MBE scaled score. A low correlation between the written component and the MBE scaled score would seem to undermine a fundamental premise of the scoring of the bar exam. In fact, a low correlation between exam components should be justification to cease the practice as antithetical to a high-stakes licensing exam. Instead, NCBE supports it and makes "adjustments." What does it mean to have correlations "adjusted for their less-than-perfect reliability"? What is the "adjustment" process? Why would it be an acceptable practice for a licensing exam to make "adjustments"?

[24] Albanese, *supra* note 2, at 53. For the July 2015 administration of the UBE, we are informed that "the correlations of the written component score with the MBE scaled score ranged from 0.44 to 0.81 and averaged 0.66 across the 14 UBE jurisdictions." The February 2016 administration of the UBE showed even weaker correlations, "ranging from 0.51 to 0.67 and averaged .60 across the 17 UBE jurisdictions."
[25] *Id.*
[26] *Id.*

Perhaps "reliability adjustments" are what Judith A. Gundersen, NCBE's program director for the Multistate Essay Examination (MEE) and the Multistate Performance Test (MPT), relies upon in finding a "correlation above .80" between the MBE scaled score and the written components and calling it "strongly correlated."[27] In addition to possible "reliability adjustments", the .80 correlation that Ms. Gundersen refers to is "a correlation of scaled MBE score to scaled written score and that is a disattenuated correlation (it represents an estimation of the true scores' correlations).[28]

This reliance on scaled score correlations is not in keeping with NCBE's own past practices in grader training and workshops where Dr. Susan Case, former Director of Testing for the National Conference of Bar Examiners, presented accurate raw score correlations.[29] According to Dr. Case, "the correlation with the MBE is "0.58 for the MEE" and only "0.38 for the MPT."[30] She explains that "[t]his shows a moderate correlation for …the MEE, but a weaker correlation for the MPT, indicating that the MPT is measuring different skills than the MBE, and the MPT skills are less like those measured by the MEE…."[31] On the other hand, "[i]f two components measured exactly the same thing, the correlation would be 1.00 (perfectly related)."[32]

In contrast to Dr. Case who presents raw score correlations by the individual components — 0.58 for MBE with MEE and only 0.38 for MBE with MPT — Ms. Gundersen combines the MEE and MPT scores before scaling to the MBE. Still, a more fundamental flaw infects the resulting .80 correlation: it is the fact that Ms. Gundersen's MEE and MPT are "scaled scores" — scores that result "after they have been forced into

---

[27] Gundersen, *supra* note 8, at 41. Ms. Gundersen claims that "because the data have consistently shown across groups and time that the total MBE scaled score is strongly correlated with overall performance on the written components (correlation above .80 when reliability of the two measures is taken into account), we can use MBE performance information as a proxy indicator of the groups' general ability levels."

[28] E-mail from Nancy E. Johnson to Suzanne Darrow-Kleinhaus, Professor of Law and Director of Academic Development and Bar Programs, Touro Law Center (Aug. 30, 2016, 5:38 p.m. EST) (on file with author).

[29] Case, *The Testing Column, Relationships Among Bar Examination Component Scores: Do They Measure Anything Different?* THE BAR EXAMINER [hereinafter Case, Do they Measure Anything Different?] http://www.ncbex.org/pdfviewer/?file=%2Fassets%2Fmedia_files%2FBar-Examiner%2Farticles%2F2008%2F770308_testing.pdf.

[30] *Id*. at 31.

[31] *Id.*

[32] *Id.*

the same distribution as the MBE, and after estimating what the 'true scores' are (that is, trying to take out the error of measurement inherent in the tests)."[33] Even so, it is a correlation of only .80. While this appears to be a strong correlation, it is not. It would be strong if it were based on raw scores, uncorrected. However, it's not.[34]

It is important to understand why scaling and correlations matter. It has to do with exam validity, a core psychometric concept. "The most important psychometric property of any exam is that it be 'valid,' which means that the exam measures whatever it is supposed to measure. An exam that is not valid is not worth much…."[35] A .80 correlation seems to indicate that the MBE and the written component are largely measuring "the same construct — the same ability — so this kind of equating of the written to the MBE is fine. It is not. With these disattenuated correlations, for the test to [be] measuring the same thing, the correlation must approach unity (1.0). Her [Ms. Gundersen's] reported number of .80, even if this were the disattenuated correlation of the raw scores rather than the standard scores, is too low."[36] Studies in the area indicate that there is cause for concern as to the validity and adequacy of using only multiple choice items as anchors to equate forms of a mixed-format test:

> For mixed-format tests, if the MC and CR [written] portions measure the same construct, in principle we would expect an MC-only anchor...to be sufficient to equate the test forms. ... In the case of an MC-only anchor and a mixed-format test, the anchor can be construct representative of the total test only to the extent that the MC and CR portions measure the same thing (I.e., the test must be unidimensional).[37]

Further, "[w]hen we say the disattenuated correlations must approach unity, that means they must be on the order of .97 - 1.0. A disattenuated correlation of .88 is called

---

[33] E-mail from Nancy E. Johnson to Suzanne Darrow-Kleinhaus, Professor of Law and Director of Academic Development and Bar Programs, Touro Law Center (Aug. 30, 2016, 5:38 p.m. EST) (on file with author).
[34] *Id.*
[35] Daggett, *supra* note 13, at 393..
[36] E-mail from Nancy E. Johnson, to Suzanne Darrow-Kleinhaus, Professor of Law and Director of Academic Development and Bar Programs, Touro Law Center (Aug. 30, 2016, 5:38 p.m. EST) (on file with author).
[37] SOOYEON KIM & MICHAEL E. WALKER, RESEARCH REPORT ETS RR-11-44, DOES LINKING MIXED-FORMAT TESTS USING A MULTIPLE-CHOICE ANCHOR PRODUCE COMPARABLE RESULTS FOR MALE AND FEMALE SUBGROUPS?, (2011) https://www.ets.org/Media/Research/pdf/RR-11-44.pdf [hereinafter KIM & WALKER, LINKING MIXED-FORMAT TESTS].

'much less than unity, casting doubt on the unidimensional of these mixed-format tests.'"[38] The disattenuated correlation reported by Ms. Gundersen 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 — and should be abandoned.

C. The low correlation between the written component and the MBE is attributable to their differences in the skills and knowledge tested.

Once again, it is reasonable to question NCBE's grading practices: why would components be scaled to each other when they are so different from each other in terms of what is tested? While a low and widely varying correlation between the written and MBE scaled score is a major cause for concern as to the validity of final test scores, it isn't the only problem: NCBE reports that there is also a problem with the reliabilities of the individual written component scores for the 14 UBE jurisdictions, since they ranged from "0.62 to 0.82 and averaged 0.73"[39] in July 2015 and from "0.48 to 0.77 and averaged 0.72"[40] for the 17 UBE jurisdictions in February 2016. As if these numbers were not sufficiently alarming, Dr. Albanese reports that "a bigger problem is that even the highest reliability [of the written component total scores] achieved in any [UBE] jurisdiction (0.82) does not reach 0.90, the minimum level normally considered adequate for high-states testing purposes."[41]

Nonetheless, Dr. Albanese asserts that scaling the written score to the MBE will account for the "possible variation in grading across jurisdictions."[42] Although this makes no sense whatsoever, we are assured that "[j]urisdictions that scale the essays to the

[38] E-mail from Nancy E. Johnson to Suzanne Darrow-Kleinhaus, Professor of Law and Director of Academic Development and Bar Programs, Touro Law Center (Aug. 30, 2016, 5:38 p.m. EST) (on file with author) quoting KIM & WALKER, LINKING MIXED-FORMAT TESTS, *supra* note 37. (2011) https://www.ets.org/Media/Research/pdf/RR-11-44.pdf

[39] Albanese, *supra* note 2, at 52.

[40] *Id.*

[41] *Id.* According to Dr. Albanese, "the reliabilities of the written component total scores for the 14 UBE jurisdictions ranged from 0.62 to 0.82 and averaged 0.73" in July 2015. "In February 2016, the reliabilities of the written component total scores for the 17 UBE jurisdictions ranged from 0.48 to 0.77 and averaged 0.72." No basis for calculating these scores is provided. How are they determined?

[42] *Id.*

MBE scores for their jurisdiction, that weight the MBE at least 50%, and that make the pass/fail decision on the total score are assured of a sufficiently high reliability and high decision consistency."[43] In short, 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 and to accept the underlying assumption that the written score is in fact unreliable. Neither premise is supportable.

D. Changes in the number and content of MBE items may have an effect on equating and just because NCBE says "no" doesn't make it so.

1. What is the effect of reducing the number of MBE live test items by about 8%?

NCBE's contends that scaling and equating "unreliable" written scores to the MBE assures reliability. NCBE relies on the MBE and its "anchor" items for equating purposes.[44] While the MBE has a large population for each exam administration so that would help with the accuracy of equating, changing the number of MBE test items from 190 to 175 items commencing with the February 2017 administration of the bar exam must have some effect — even if NCBE categorically denies it. Ms. Moeser advised law school deans that while the MBE will consist of only 175 scored items, "MBE scores will continue to be expressed on a 200-point scale. Because MBE scores are equated and scaled, scores will be comparable to those earned when there were more scored questions."[45] As Nancy Luebbert, Director of Academic Success at the University of Idaho College of Law, observed, "[o]ne doesn't have to be a mathematician to recognize that the effect of one wrong answer is magnified when the number of test items goes down, and that this effect is most pronounced for those near the pass line."[46]

---

[43] Case, *Quality Control*, *supra* note 9, at 34, 36.

[44] *See generally* Erica M. Moeser, *President's Page*, THE BAR EXAMINER, December 2014, at 4, http://www.ncbex.org/pdfviewer/?file=%2Fassets%2Fmedia_files%2FBar-Examiner%2Farticles%2F2014%2F830414-presidentspage.pdf.

[45] Memorandum from Erica Moeser, President, NCBE, to Law School Deans (at all American Bar Association-accredited law schools) (August 31, 2016) http://online.wsj.com/public/resources/documents/2016_0831_moeser_memo.pdf [hereinafter Moeser, Letter to Law School Deans, Aug. 31,2016].

[46] Posting of Nancy Luebbert to asp-1@chicagokent.kentlaw.edu (Aug. 31, 2016, 1:46 p.m. EST) (The subject heading of this email is [ASP-L:6281] *Re: ASP Response to concerns with the NCBE*) (on file with author).

Following NCBE's announcement about the increase in the number of pre-test items, the Association of Academic Support Educations (AASE) wrote to Erica Moeser and Robert A. Chong, Chair of the Board of Trustees, to express concerns on behalf of its membership.[47] AASE questioned the expansion of the number of pre-test items from ten (10) questions to twenty-five (25) questions and raised the following issues: first, why did NCBE provide no explanation for the change and; second, whether the change to 25 pre-test items for an exam of only 200 items raised questions about the MBE's reliability because "as a general matter, long exams tend to have higher reliability than short ones."[48] Thus, a reduction in the number of items used to measure performance on the MBE from 190 graded questions to 175 graded questions "negatively impacts the accuracy of the sampling measurements used within one exam to necessarily generalize from a smaller subset of questions to the larger question of minimum competency."[49] Further, "the decrease in the number of items is likely to distort the reliability of the MBE exam instrument especially given the addition of Federal Civil Procedure material…."[50] Finally, there is an undue burden placed on the examinees who are now compelled "to spend one-eighth of their exam time (forty-five minutes out of their six hours) and considerable effort on unscored work."[51]

The burden placed on examinees is not insignificant. According to Professor Deborah Merritt, an increase from 18 minutes to 45 minutes is "a substantial amount of time, especially when 'volunteered' in the midst of a stressful, tiring experience."[52] It is not simply a matter of adding more time to the test: pre-test questions add stress. They add stress because the new questions "may be more ambiguous or difficult than well-tested

---

[47] Letter from Jamie A. Kleppetsch, President, Association of Academic Support Educators, to Robert A. Chong, Chair of the Board of Trustees, National Conference of Bar Examiners and Erica Moeser, President & Chief Executive Officer, National Conference of Bar Examiners (Sept. 23, 2016)(on file with author)[hereinafter, Kleppetsch, AASE Letter]. AASE is an organization comprised of more than 200 academic support and bar preparation professors representing law schools throughout the country. The concerns were three-fold: public revelation of a change without explanation, failure to solicit comments and advice from law school administrators, faculty, staff, and other stakeholders, and being informed only after the decision had been implemented.

[48] Daggett, *supra* note 13, at 396.

[49] Kleppetsch, AASE Letter, *supra* note 47.

[50] *Id.*

[51] *Id.*

[52] Deborah J. Merritt, *The Latest Change in the MBE*, LAW SCHOOL CAFÉ, (Sept. 5, 2016), http://www.lawschoolcafe.org/2016/09/05/the-latest-change-in-the-mbe/.

ones." Further, and perhaps most important, "exam-takers …can't skip over challenging pre-test questions and focus on the 'real' questions… [because] they don't know which are which. Answering flawed pre-test items can absorb disproportionate amounts of time and raise stress levels."[53]

NCBE's statement that the change in the number of pre-test items will have no effect — either positive or negative — is conclusory and unacceptable without evidentiary support. This is especially relevant in light Dr. Case's own prior statements regarding the inherent sampling, reliability, and validity issues with respect to written portions of the bar exam and equating to the MBE.[54] As Dr. Case repeatedly stated, "the more questions you ask, the higher the reliability."[55] She also stated that "[t]he broader the content domain, the more questions are required."[56] And just to be sure that the concept was clear, Dr. Case added that "[i]f more questions provide greater reliability, it follows that reliability is reduced when fewer questions are used."[57]

Once again, NCBE is caught in a contradiction between what it has represented historically and what it says now. NCBE asks us to accept its statement that although it added a whole other content domain to the MBE — Federal Civil Procedure — and reduced the number of questions in each of the other six content areas, and further reduced the number of "live" questions from ten to twenty-five questions, there is absolutely no effect whatsoever on the entire scaling and equating process.[58] It is simply not possible to accept their word on this — not when a measurement error in its equating

---

[53] *Id.*

[54] Kleppetsch, AASE Letter, *supra* note 47,, referring to Susan M. Case, *The Testing Column, What Everyone Needs To Know About Testing, Whether They Like It Or Not,* THE BAR EXAMINER, June 2012, at 29 [hereinafter Case, *What Everyone Needs To Know*] http://www.ncbex.org/pdfviewer/?file=%2Fassets%2Fmedia_files%2FBar-Examiner%2Farticles%2F2012%2F810212beTestingColumn.pdf.

[55] Case, *What Everyone Needs to Know*, *supra* note 54.

[56] *Id.*

[57] *Id.* at 29-30.

[58] Erica M. Moeser, *President's Page,* THE BAR EXAMINER, December 2015, at 4. In responding to questions about whether the addition of Civil Procedure as the seventh MBE content area was responsible for the poor MBE performance, Moeser responded that "[o]ur research is solidly convincing that the addition of Civil Procedure had no impact on the MBE scores earned on the February and July 2015 MBE administrations." http://www.ncbex.org/pdfviewer/?file=%2Fassets%2Fmedia_files%2FBar-Examiner%2Fissues%2F2015-December%2FBE-Dec2015-PresidentPage.pdf.

and standardization will result in a life-changing error for an examinee who is right around the pass line, whether above or below it.[59]

As AASE explained in its letter to NCBE, "we are concerned that changes in the MBE without accompanying changes in the scaling methodology used by the NCBE (and apparently by most jurisdictions) to adjust written scores to the same scale of the MBE based on jurisdictional mean, standard deviation, and range data, might result in further degradation of the efficacy of the entire exam."[60] The public's interest in a fair and transparent licensing process outweighs the interests of any entity. We need time to have this change studied by a disinterested party to validate NCBE's representations.

2. How many of the 175 scored items are MBE's "anchor items"?

Even assuming that the change in the number of pre-test items has no effect on the reliability of the MBE for equating purposes, there is a separate issue with respect to the MBE's anchor items. The MBE now has 175 scored items: how many of those will be anchor items? Anchor items are the embedded test questions that have appeared on previous test forms and included in the current test form. They are used to compare "the performance of the new group of test takers …. on those questions with the performance of prior test takers on those questions. The embedded items are carefully selected to mirror the content of the overall test and to effectively represent a mini-test within a test."[61]

The content and statistical properties of the anchor questions are critical to the equating process, but we don't know how many anchor items are used and how they have been affected by the change in the number and content of the MBE's scored items. However, we do know that the higher the number, the more accurate the equating process. Everything is related: "the accuracy of the equating of the MBE affects the

[59] Posting of Nancy E. Johnson to asp-1@chicagokent.kentlaw.edu (Aug. 31, 2016, 11:19 p.m. EST) (The subject heading of this email is [ASP-L:6276] *Re: ASP Response to concerns with the NCBE*) (on file with author).
[60] Kleppetsch, AASE Letter, *supra* note 47, referring to Case, *What Everyone Needs To Know, supra* note 54, at 31.
[61] Erica M. Moeser, *President's Page*, THE BAR EXAMINER, December 2014, at 4, http://www.ncbex.org/pdfviewer/?file=%2Fassets%2Fmedia_files%2FBar-Examiner%2Farticles%2F2014%2F830414-presidentspage.pdf.

accuracy of the scaled written score because [NCBE] scales the written to the MBE distribution for the jurisdiction."[62]

3. How has the change in the examinee population affected the equating process?

Lowering the number of live MBE questions for scoring purposes is not the only issue with scaling the entire exam. The other critical issue concerns the changes in the population taking the MBE when NCBE claims that their standardization process means that a 135 on the MBE last year is the same as a 135 now and a 135 ten years ago.[63] According to Dr. Nancy Johnson, NCBE's "standardization of the MBE rests on the necessary assumption that the population today is the same (in terms of the underlying ability they're testing) as was the population who originally answered the anchor items they're using to standardize. To the extent that those two populations differ in that ability the standardization becomes unreliable."[64] And the populations differ in that ability because NCBE President Erica Moeser has told us so. They differ in that all indicators "point to the fact that the group that sat in July 2014 was less able than the group that sat in July 2013."[65]

---

[62] Posting of Nancy E. Johnson, to asp-1@chicagokent.kentlaw.edu (Aug. 31, 2016, 11:19 p.m. EST) (The subject heading of this email is [ASP-L:6276] *Re: ASP Response to concerns with the NCBE*) (on file with author).

[63] Case, *What Everyone Needs To Know*, *supra* note 54, at 31. Dr. Case explains that "[s]caling written-component scores to the MBE involves an algebraic process that places the written-component scores on the same scale as the MBE. This process "equates" the written-component scores and assures that the scores mean the same thing across test administrations." Erica Moeser states that "[t]he result is that a scaled score on the MBE this past summer—say 135—is equivalent to a score of 135 on any MBE in the past or in the future." http://www.ncbex.org/pdfviewer/?file=%2Fassets%2Fmedia_files%2FBar-Examiner%2Farticles%2F2012%2F810212beTestingColumn.pdf; http://www.ncbex.org/pdfviewer/?file=%2Fassets%2Fmedia_files%2FBar-Examiner%2Farticles%2F2014%2F830414-presidentspage.pdf.

[64] E-mail from Nancy E. Johnson to Suzanne Darrow-Kleinhaus, Professor of Law and Director of Academic Development and Bar Programs, Touro Law Center (Feb. 16, 2017, 11:21 a.m. EST) (on file with author).

[65] Memorandum from Erica Moeser, President National Conference of Bar Examiners to Law School Deans on Two Matters (October 23, 2014) [hereinafter, Moeser, Letter to Law School Deans, Oct. 23, 2014] https://www.law.upenn.edu/live/files/3889-multistate-bar-exam-memo-oct-2014pdf. Ms. Moeser defended the MBE scores from the July 2014 test administration, and informed law school deans, that "[b]eyond checking and rechecking our equating, we have looked at other indicators to challenge the results. All point to the fact that the group that sat in July 2014 was less able than the group that sat in July 2013."

According to Erica Moeser, the historic plunges in bar exam pass rates over the past few years are not likely an aberration but the start of a trend.[66] She has acknowledged a convergence of events that has changed the world of legal education and by extension law licensure, a situation she has termed "the new normal."[67] According to Ms. Moeser, "[i]t is telling that between fall 2012 and fall 2013 the law school entering class that emerged in 2016 was reduced from 43,155 to 39,674. That figure dropped to 37,892 first-year students in the fall of 2014, the class that will graduate in 2017 and test that July."[68] Not only are there far fewer candidates sitting for the bar exam, but today's bar candidates are different from previous bar candidates for many reasons, not least of which is that they are "less able" because law schools are admitting less qualified students.[69] The determination of "less qualified" is based on entering class data for scores marking the 25th percentile level of the Law School Admission Test (LSAT). The data for the class that entered law school in fall 2015 and will graduate in 2018 are "still discouraging."[70] Further, Ms. Moeser claims that the downward spiral was "not

---

[66] Erica M. Moeser, *President's Page,* THE BAR EXAMINER, September 2015, at 4, ("Over the course of the past year, [last year's] analysis pointed to the probability that the scores earned in July 2015 would represent the continuation of a downward slide, and that is what we can now confirm. At 139.9, this July's mean MBE score is the lowest July score since 1988, when it was 139.8.")http://www.ncbex.org/pdfviewer/?file=%2Fassets%2FUploads%2FBE-Sept2015-PresidentPage.pdf

[67] Erica M. Moeser, *President's Page,* THE BAR EXAMINER, December 2016, at 4, "http://www.ncbex.org/pdfviewer/?file=%2Fassets%2Fmedia_files%2FBar-Examiner%2Farticles%2F2016%2FBE-850416-PresidentsPage.pdf.

[68] *Id.*

[69] Moeser, Letter to Law School Deans, Oct. 23, 2014, supra note 65.

[70] Moeser, *President's Page,* THE BAR EXAMINER, March 2016, at 5, http://www.ncbex.org/pdfviewer/?file=%2Fassets%2Fmedia_files%2FBar-Examiner%2Farticles%2F2016%2FBE-March2016-PresidentPage.pdf. See charts on pages 11 and 12: Change in Enrollment and LSAT Score at the 25th Percentile from 2010 to 2015 and Changes in First-Year Enrollment and Average LSAT Score at the 25th Percentile, 2010-2015, respectively. Although NCBE provides no total number for how many law schools provided data for the charts, the scatterplot analysis indicates that "[m]ost of the schools appear in the lower left quadrant; this quadrant contains schools that have experienced decreases in both the LSAT score at the 25th percentile and their enrollment numbers." *See also*, Paul L. Caron, TAXPROF BLOG, *Law School Applicants From Top Colleges Increased 1% In 2016 (But Down 48% Since 2010),* March 1, 2017 (noting that "for the first time since 2010, the total number of graduates from the nation's top universities increased instead of continuing to decline." While only a slight increase, it may be a sign that top university students are considering law school once again)*;* http://taxprof.typepad.com/taxprof_blog/2017/03/law-school-applicants-from-top-colleges-increase-1-in-2016-but-down-48-since-2010.html.

unexpected" since "[w]e are in a period where we can expect to see some decline, until the market for going to law school improves."[71]

If today's bar candidates are different from past candidates, then Erica Moeser has exposed a true vulnerability: there is no valid way to standardize the test because the current population is not equivalent to past ones. Still, she stands firmly behind the quality of MBE scoring and its equating process — even while pointing "to the fact that the group that sat in July 2014 was less able than the group that sat in July 2013."[72] Historically, NCBE's defense to changes in candidate populations is that "[t]he MBE is merely the messenger" and each jurisdiction "sets its own standards for admission."[73] However, this argument does not apply to the UBE where the scores are promoted as portable: how can the UBE be "portable" so has to have the same meaning in one UBE jurisdiction as it does in another when that score is based on NCBE's standardization and equating process?

The answer is that it can't — not when the populations differ in ability. NCBE's "standardization of the MBE rests on the necessary assumption that the population [taking the test today] is the same (in terms of the underlying ability they're testing) as was the population who originally answered the anchor items they're using to standardize. To the extent that those two populations differ in that ability, the standardization becomes unreliable."[74]

4. How are varying groups affected by the changes?

---

[71] Natalie Kitroeff, *Bar Exam Scores Drop to Their Lowest Point in Decades: Unprepared Students Can't Handle a Harder Test* BLOOMBERG (Sept. 17, 2015, 2:32 PM) https://www.bloomberg.com/news/articles/2015-09-17/bar-exam-scores-drop-to-their-lowest-point-in-decades.

[72] Memorandum by Erica Moeser dated 2014, *supr*a note 65.

[73] Erica M. Moeser, *President's Page,* THE BAR EXAMINER, September 2015, at 4, http://www.ncbex.org/pdfviewer/?file=%2Fassets%2FUploads%2FBE-Sept2015-PresidentPage.pdf.

[74] E-mail from Nancy E. Johnson to Suzanne Darrow-Kleinhaus, Professor of Law and Director of Academic Development and Bar Programs, Touro Law Center (Feb. 16, 2017, 11:21 a.m. EST) (on file with author).

According to Ms. Moeser, "[a]s to the question of minority performance on the UBE, little information exists."[75] Nonetheless, she concludes that the issue of whether minority examinees "fare worse on the MBE than on other parts of the bar exam (or, for that matter, on any bar examination) … has been laid to rest on more than one occasion and has been reported in the pages of this magazine. Minority performance on the MBE is not materially better or worse than it is on other portions of the bar examination."[76]

However, there is "information" and those who have considered it have raised concern as to whether NCBE's equating method works equivalently for different subpopulations.[77] In a 2011 study, Kim and Walker "looked at linking mixed-format using a multiple-choice anchor and asked whether it would produce comparable results for men and women. They 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."[78] In a recent empirical study, Dennis P. Saccuzzo and Nancy E. Johnson evaluated the likely outcome, by California ABA law school, of upcoming changes in the California Bar Exam. The results show that law school will be affected unequally because the weighting of the MBE will be increased and that of the written component will be decreased. The authors sent this research to all of the ABA deans in California informing them that women and minorities will be hurt by the changes. To the extent that a school's proportion of women relative to men increases, the school's pass rate will be differentially affected by the scoring changes in the California bar exam beginning in July 2017.[79]

---

[75] Erica M. Moeser, *President's Page,* THE BAR EXAMINER, June 2016, at 7, http://www.ncbex.org/pdfviewer/?file=%2Fassets%2Fmedia_files%2FBar-Examiner%2Farticles%2F2016%2F850216-BE-June2016-PresidentsPage.pdf

[76] Erica M. Moeser, President's Page, THE BAR EXAMINER, June 2016, at 5.

[77] Posting of Nancy E. Johnson to asp-1@chicagokent.kentlaw.edu (Apr.17, 2015, 1:44:50 p.m. EST) (The subject heading of this email is [ASP-L:5369] *Re: NCBE Responses RE July 2014 MBE Nationwide Decline*) (on file with author).

[78] *Id. See also* KIM & WALKER, LINKING MIXED-FORMAT TESTS, *supra* note 37.

[79] Dennis P. Saccuzzo & Nancy E. Johnson, *California's New Bar Exam Format and ABA's Proposed 75% Bar Passage Requirement Will Adversely Impact Diversity, Women, and Access to the Legal Profession*, TAXPROF BLOG OP-ED., (Jan. 30, 2017) http://taxprof.typepad.com/files/taxprof-blog-op-ed-adverse-impact-on-diversity.pdf l.

Data is essential to determine the validity of the equating process. Dr. Nancy Johnson explains that "there are two basic assumptions that must be met in order to get equated scores (assuming [NCBE uses] a technique called chained equipercentile method): the relationship between the anchor (the subset of MBE items used as the anchor) and the total scores is invariant across populations, and the same thing is true on the new form."[80] As Dr. Johnson further explains, "[i]f you do not have data on populations (ethnic sub populations, gender, jurisdiction etc.), then you cannot know if those assumptions are met and you therefore do not know if the technique is invariant across populations. Apparently, NCBE doesn't know."[81]

But figuring out whether there is bias across groups is a complex thing to do, and you would need to know not just correlations, but also effect sizes — how much did the test takers differ in average proficiency from one administration to the next? We know they differ, because NCBE repeatedly tells us they do. February takers are less proficient than July takers, and recent takers are much less proficient than previous takers.[82] Equating tends to be more accurate when those differences are very small. The magnitude of the error in equating increases as the correlation between two constructs (for example, MBE versus written) decreases, but the error in equating also increases as the group difference increases. "In general, a higher correlation is needed as the group difference increases to achieve adequate equating."[83]

E. "Obvious differences" between the written components make it inappropriate for them to be scaled to the MBE.

While NCBE admits that there are "obvious differences" between the written components (MPTs and MEEs) and the multiple choice component (MBE), it nonetheless concludes that "the two parts of the exam do fundamentally measure similar abilities"

[80] E-mail from Nancy Johnson to Suzanne Darrow-Kleinhaus, Professor of Law and Director of Academic Development and Bar Programs, Touro Law Center (Aug. 30, 2016, 5:38 p.m. EST) (on file with author).
[81] *Id.*
[82] Gundersen, *supra* note 8, at 41. As an example of the variation in examinee proficiency, Ms. Gundersen states that "in the February administration, examinee proficiency tends to be lower due to a larger proportion of repeat test takers. We see this lower performance reflected on the MBE in February and expect to see lower scores on the MEE and MPT as well."
[83] Center for Advanced Studies in Measurement and Assessment, Monograph Number 2.2, Dec. 2012.

such that one should be scaled to the other. Yet again, NCBE presents an assumption as a conclusion without validation or explanation. Maybe there is none. Or, more likely, the explanation supports an opposite result.

According to psychometricians and legal educators, there are significant differences between written and multiple choice exams — differences "leading to inconsistent pass/fail decisions for low-performing examinees in particular...."[84] Even without data, educators know there is a significant difference between "knowing something, and being able to express it."[85] The two are not the same. According to Professor Krimmel, "the ability to recognize the applicable legal rule when it presents itself in a structured array [is not] the same skill as the ability to summon it forth from a body of facts. The answer to a multiple choice question is quite literally on the page; the answer to an essay question is in the student's mind waiting to be born."[86]

Different testing vehicles can produce different results.[87] The following example from Professor Krimmel explains how a simple spelling test can be constructed in different ways and produce different results:

> Even something as mundane as a spelling test can be constructed in several different ways, and at least for some students, their performance can be strongly affected based on the testing vehicle. Some students apparently will often perform quite differently when asked to spell a word versus identify which words in a list are

[84] Sooyeon Kim & Michael Walker, Determining the Anchor Composition for a Mixed-Format Test: Evaluation of Subpopulation Invariance of Linking Functions, 25 Applied Measurement in Education 178 (2012). Commenting on the Kim & Walker (2012) findings, Dr. Johnson notes that it is stunning in its implications in that "as more of the pool approaches the region we would call low-performing, bias increases." Posting of Nancy E. Johnson to asp-1@chicagokent.kentlaw.edu<mailto:1@chicagokent.kentlaw.edu><mailto:1@chicagokent.kentlaw.edu<mailto:1@chicagokent.kentlaw.edu?>> (Apr.17, 2015, 1:44:50 p.m. EST) (The subject heading of this email is [ASP-L:5369] Re: NCBE Responses RE July 2014 MBE Nationwide Decline) (on file with author). *See also* Kim & Walker, Linking Mixed-Format Tests, *supra* note 37; In 2011, Kim & Walker "looked at linking mixed-format using a multiple-choice anchor and asked whether it would produce comparable results for men and women. They 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."

[85] Herbert T. Krimmel, Dear Professor: Why Do I Ace Essay Exams but Bomb Multiple Choice Ones? 63 J. of Legal Educ.433 (2014).

[86] *Id.*

[87] *Id.*

> misspelled, versus find the misspelled words in a document, versus identify which of two spellings of a word is correct.[88]

Not to be deterred by unreliable scores and the extremely low correlations of a mixed-format exam, Dr. Albanese deflects alarm with a rhetorical question: is the variability between jurisdictions in the correlation of the written score with the MBE scaled score really "a difference that makes a difference?"[89] Well, you'd better believe that it's a difference that makes a difference because for some examinees, it might make all the difference between passing and failing the bar exam.

It makes a difference because with the UBE, the written component is not scaled to a national distribution. Instead, it is scaled to that jurisdiction's MBE distribution by forcing it to have the mean and standard deviation as that of the MBE distribution for that jurisdiction.[90] In other words, the same skill level on the essays and MPT would get a different score in different jurisdictions, depending not only on the relative written skill of the jurisdiction's candidates, but also the relative MBE skill. This can have a significant impact on individual scores, especially in smaller jurisdictions.[91]

---

[88] *Id.*

[89] Albanese, *supra* note 2, at 53.

[90] *Id.* at 52. Scaling to the MBE is supposed to provide a consistent meaning over time because the national distribution of the MBE is equated across time and the raw scores across the country presumably approximate a normal distribution. Dr. Albanese states that the 190-item MBE has a reliability of 0.92 for recent administrations and for the July 2016 administration, it had a reliability of 0.93.

[91] Suzanne Darrow-Kleinhaus, *UBE Shopping: An Unintended Consequence of Portability?* (March 30, 2016) Touro Law Center Legal Studies Research Paper Series No. 16-14, *available at* SSRN, http://ssrn.com/abstract=2756520. The following example shows how this is possible. Consider the following examples:[91]

> Using the same method of scaling that NCBE uses, let's see what would happen with a hypothetical candidate. According to Dr. Nancy Johnson, assuming we have a candidate who scores 125 on the MBE when the national mean is 140 and the standard deviation is 15 (so this candidate is 1 s.d. below the national mean because the MBE is her relative weakness). However, our candidate is good at essays and the MPT so her written score is 1 s.d. above the mean for her jurisdiction. According to the methodology that NCBE uses in scaling MBE scores, our candidate's essay score will be computed to be 140 + 15 = 155 because the jurisdiction's MBE mean is 140 and its s.d. is 15. That would give our candidate a total UBE score of 155 + 125 = 280, which is high enough for admittance in several jurisdictions, including New Mexico, Idaho, Washington and New York.
>
> Let's consider what happens if the jurisdiction's MBE mean is down at 135, with a standard deviation still at 15. If our candidate "scores 1 s.d. above the mean on the written, then her written score will be standardized to 135 + 15 = 150. That means that her total UBE score would be 150 +

F. NCBE's assessment of the reliability of the bar exam's written component may be flawed.

Even assuming we could put aside the absurdity of scaling an unreliable written score to an MBE score and the troubling effect of the variability of the applicant pool on a candidate's written score, there is an inherent defect in how Dr. Albanese measures the MEE and MPT for determining their sample size. Sample size is important because a larger sample size means a larger sample of performance which increases the reliability of the written component total score.[92] Dr. Albanese claims that the reliability of the written component total score — the MEE and MPT — is much lower than that of the 190-item

125 = 275. She would no longer be eligible in Idaho (where the minimum required is 280) simply because of the slightly lower mean but same variance in MBE scores in her jurisdiction. Her skill level did not change: that of the pool of candidates did. Is this what we want to mean when we tout the "portability" of the UBE?"

Now consider that the jurisdiction's MBE mean is at 140 but the standard deviation is not as large — make it 12 rather than 15. The MBE score is still 125 but now our candidate's written score that is 1 s.d. above the mean in her jurisdiction gets scaled to 140 + 12 = 152. Her total score on the UBE is then 152 + 125 = 277 and again she would not be able to transport that score to Idaho for admission.

But those are pretty simplistic examples. If our candidate is really that good at the written component (in the 84th percentile in her jurisdiction if she is 1 s.d. above the mean) and she chooses a jurisdiction where the applicant pool is, for whatever reason, weaker in written performance, then her performance will be more than 1 s.d. higher in that jurisdiction. It can get a bit complicated to estimate this but just say that the MBE mean is down at 135 as in the second example, and relative to the weaker pool her written score winds up being 2.5 s.d. above the mean. Then her written score would scale to 135 + 22.5 = 157.5 and that elevates her total UBE score to 125 + 157.5 = 282.5. This would give her entry into just about any UBE jurisdiction.

It would seem likely that with smaller sample sizes, it would be more likely to see variations from the normal distribution. However, it is not possible to determine how seriously that would distort the standardization because so little information about the national sample and the individual jurisdictions are available. Nonetheless, it is possible to see that the more you "work the numbers" the way that NCBE does, the more you see that the same skill level could result in different UBE scores, depending on where the candidate takes the exam and what that jurisdiction's applicant pool does on that particular exam, in terms of both skill level and also the range or spread of their scores.

[92] Albanese, *supra* note 2, at 52.

MBE[93] because the written portion has "only eight different scores in UBE jurisdictions (one for each of the six MEEs and two MPTs)."[94]

Maybe the written component is not quite as unreliable as NCBE contends — and the MBE no longer quite as reliable since its reduction to 175 items. As is usually the case with numbers, there is another way to look at them. Although there are only six MEEs with one score for each MEE, each MEE typically consists of four questions. Each question requires an issue, a rule paragraph, and an analysis. A rule paragraph typically includes a general rule statement, definitions for legal terms of art (or elements and factors), and identification of the applicable exception(s). This is followed by an analysis of the relevant facts and, where applicable, discussion of the counter-arguments. Therefore, depending on how one defines an "item," the total for six MEEs is certainly more than six items. At the very least, an examinee would write at least 24 separate analyses for the MEEs. Given that each rule, definition, and exception would count as an "item" and that an analysis of each fact would count as well, this number would increase exponentially. Consequently, what NCBE counts as six scores for the MEEs can be seen as representing 200 or more, depending on the complexity of the questions.

For example, I deconstructed the answer to Question 1 of the July 2007 MEE. Following NCBE's Analysis Sheets to identify the issues, rules, and facts that might appear in a complete answer, I counted 26 line items based on four issues.[95] Still, this number did not include further breakdowns as would be appropriate on a scoring sheet for rules and facts where there are individual factors, elements, etc. for rules and several relevant facts for an analysis. By way of illustration, this problem included a statute of frauds defense. A complete rule statement would require identification of the one-year rule[96] and what is required for a writing to satisfy the statute of frauds: a writing that is signed by the party to be charged and reflects the agreement with adequate specificity

[93] Effective with the February 2017 bar exam, the MBE has been reduced to 175 scored questions and 25 unscored pretest questions. *See* Preparing for the MBE, Test Format, NATIONAL CONFERENCE OF BAR EXAMINERS, http://www.ncbex.org/exams/mbe/preparing/ (last visited February 10, 2017).
[94] Albanese, *supra* note 2, at 52.
[95] An evaluation sheet for the July 2007 MEE is on file with the author.
[96] Restatement (Second) of Contracts § 130 (AM. LAW INST. 1979).

(the material terms). The problem also required a multi-leveled analysis of the acceptance issue including recognition of mailbox rule and how to analyze the situation when an acceptance is dispatched after a rejection. Depending on how a grader chooses to identify the items, 40 items would be a conservative estimate.

Assuming, therefore, that six MEEs would yield approximately 240 items, then the MEEs are more reliable than NCBE's data would indicate. The same is true of the MPT. NCBE considers an MPT as "one" item for its computation purposes, but it is hardly that. An MPT is a 90-minute simulation of a lawyerly task and typically presents the examinee with two issues to discuss. Here, too, a sample deconstruction is helpful. In this case, I considered *Miller v. Trapp* which is MPT-2 in the February 2016 bar exam.[97] Examinees were asked to complete two tasks: a demand letter in anticipation of a lawsuit for assault and battery and a brief memo to the law partner that analyzes the compensatory and punitive damages that might be recoverable at trial. Following NCBE's Analyses Sheets, I identified 19 items in the Statement of Facts for the demand letter. This does not include the format and guideline requirements (another 6 items) or the analysis of the assault and battery claims (another 20 items). Even without these items, I counted 45 items for only the demand letter. It is safe to assume that the second task analyzing possible damages would yield a minimum of another 25 items. Assuming, therefore, that a single MPT yields around 70 items, it is necessary to double this number because there are two MPTs in an administration of the UBE. Now we can do the math:

2 MPTs @ 70 items each = 140

6 MEEs @ 40 items each = 240

[97] An evaluation sheet for the February 2016 MPT is on file with the author. NCBE writes that "[t]he MPT Point Sheets describe the factual and legal points encompassed within the lawyering tasks to be completed. They outline the possible issues and points that might be addressed by an examinee. They are provided to the user jurisdictions to assist graders in grading the examination by identifying the issues and suggesting the resolution of the problems contemplated by the drafters. An examinee need not present his/her response in the same way or cover all the points discussed in the grading materials to receive a good grade." *See, The MPT: July 2011 MPTs and Point Sheets,* NATIONAL CONFERENCE OF BAR EXAMINERS, http://www.ncbex.org/pdfviewer/?file=%2Fdmsdocument%2F204/, (last visited Mar.13, 2017).

Total = 380 items

This number makes sense for several reasons. First, the number of items in an MPT are about double that of an MEE which correlates to their respective weights in computing an examinee's score.[98] Second, the MEEs and the MPTs are taken in three-hour sessions, like the MBE, which means that they there should be a rough equivalency in the number of test items.

Even if the written component total score is not as unreliable as NCBE claims, there is still no valid reason to scale it to the MBE and several reasons not to. As discussed in *UBE Shopping*, there is a general absence of information regarding the mean and standard deviation for the MBE and the written component used to determine bar scores in jurisdictions. Without this information, there is no way to replicate and therefore validate the "equating process" followed by NCBE and jurisdictions in arriving at examinee scores. Nor is there any way to assess the "validity and reliability of using only multiple choice items as anchors to equate forms of a mixed-format test."[99]

We could continue to go back and forth on this issue but it wouldn't be productive. Unless and until NCBE is forthcoming regarding the mean and standard deviation for the MBE and the written component used to determine bar scores in jurisdictions, there is no way to verify NCBE's assertions of reliability or claims that it is possible for examinees to shop around to increase the likelihood of bar passage. While Dr. Albanese provides some information, it is not specific enough to be useful. We need to know the MBE mean and the standard deviation from that mean for each jurisdiction because the essays and performance test raw scores are scaled using that number. Instead, he provides the range

[98] UBE jurisdictions agree to weight the MEE at 30%, the MPT at 20%, and the MBE at 50% in determining an examinee's score. *See The Uniform Bar Exam, UBE Scores*, NATIONAL CONFERENCE OF BAR EXAMINERS, http://www.ncbex.org/exams/ube/scores/, (last visited Mar. 13, 2017).

[99] Posting of Nancy E. Johnson to asp-1@chicagokent.kentlaw.edu (Apr.17, 2015, 1:44:50 p.m. EST) (The subject heading of this email is [ASP-L:5369] *Re: NCBE Responses RE July 2014 MBE Nationwide Decline*) (on file with author).

of mean MBE scores for the 14 UBE jurisdictions for the July 2015 bar exam and the range for the standard deviation.[100]

Nonetheless, if the computations for the MEEs and MPTs show anything, it's that the written component total score reflects a much larger performance sample than NCBE would have us believe. This directly affects the reliability of the written component total score and allows us to conclude that there is no need to scale it to the MBE scaled score to "achieve" reliability.

III. Measuring minimum competence must be a determination of individual competency

According to state bar examiners, the primary purpose of the bar examination is "to ensure that all who are ultimately admitted have demonstrated minimum technical competence."[101] Given this objective, the practice of relative grading is antithetical to an assessment of the individual's competency to practice law. This is especially objectionable with the UBE when essays are not scaled to a national distribution but are instead scaled to that jurisdiction's MBE distribution. As we've seen, this may lead to the preposterous result of a different numerical score for the exact same performance depending on where the examinee wrote the test.

Relative grading has no place in a licensing test to determine an individual's minimal competency for the practice of law — nor is it necessary. Criterion-referenced grading is a viable alternative and is used in another high-stakes licensing exam, the Uniform CPA Examination. Developed by the American Institute of CPAs ("AICPA"), the Uniform CPA Examination "supports the profession's commitment to protecting the public interest. Equally important is providing reasonable assurance to boards of accountancy that individuals who pass the Exam possess the minimum level of technical knowledge

[100] Albanese, *supra* note 2, at 53.

[101] FLORIDA BOARD OF BAR EXAMINERS, https://www.floridabarexam.org/web/website.nsf/52286AE9AD5D845185257C07005C3FE1/4185C019FBDF17AC85257C0700649F91 (last visited Mar. 10, 2017).

and skills necessary for initial licensure."[102] In representing the CPA profession, the AICPA shares many of the same goals and objectives as the National and State Boards of Law Examiners.[103]

Unlike the Uniform Bar Exam, "[t]he CPA Examination is NOT curved. Every candidate's score is entirely independent of other candidates' Examination results."[104] Moreover,

> [t]he CPA Examination is a criterion-referenced examination which means that it rests upon pre-determined standards. Every candidate's performance is measured against established standards to determine whether the candidate has demonstrated the level of knowledge and skills that is represented by the passing score. Every candidate is judged against the same standards, and every score is an independent result.[105]

Moreover, where the written component of the UBE is relatively-graded and then scaled to the MBE mean in that jurisdiction to produce a total score, the Uniform CPA Exam arrives at a total score as follows:

> For AUD, FAR, and REG, separate scores are produced for multiple-choice questions and task-based simulations. The two scores are

---

[102] AICPA, *Practice Analysis Final Report: Maintaining the Relevance of the Uniform CPA Examination*, at 2 (April 4, 2016) http://www.aicpa.org/BecomeACPA/CPAExam/nextexam/DownloadableDocuments/2016-practice-analysis-final-report.pdf (last visited February 11, 2017).
The AICPA notes that for "the purpose of identifying the domain of tasks, knowledge and skills necessary to protect the public interest, a newly licensed CPA is defined as an individual who has fulfilled the applicable jurisdiction's educational and experience requirements and has the knowledge and skills typically possessed by a person with two years of experience."

[103] *AICPA Mission and History*, AICPA, http://www.aicpa.org/About/MissionandHistory/Pages/default.aspx (last visited February 11, 2017)."Founded in 1887, the AICPA represents the CPA profession nationally regarding rule-making and standard-setting, and serves as an advocate before legislative bodies, public interest groups and other professional organizations. The AICPA develops standards for audits of private companies and other services by CPAs; provides educational guidance materials to its members; develops and grades the Uniform CPA Examination; and monitors and enforces compliance with the profession's technical and ethical standards. The AICPA's founding established accountancy as a profession distinguished by rigorous educational requirements, high professional standards, a strict code of professional ethics, a licensing status and a commitment to serving the public interest."

[104] *Uniform CPA Examination FAQs- Scoring,* AICPA , published Mar.11, 2017, http://www.aicpa.org/BecomeACPA/CPAExam/ForCandidates/FAQ/Pages/computer_faqs_3.aspx#curve (last visited Mar. 13, 2017).

[105] *Id.*

> then weighted according to the percentage value of each component, and added together to arrive at a total score. For BEC separate scores are produced for multiple-choice questions and written communication tasks and then added together according to the percentage value of each component for the final score. [106]

The example of the Uniform CPA Exam provides an alternative basis for scoring the written component of a high-stakes, uniform licensing exam. If the accounting profession can design and implement a national, uniform licensing exam that uses criterion-based assessment to determine an individual's minimum competency to practice public accounting, then surely the legal profession can do the same. Similarly, if the AICPA can make its testing and scoring process transparent to CPA candidates and the public, then the National Conference of Bar Examiners can make its process of weighting and scaling questions available as well.[107]

IV. Testing the hypothesis

The stakes are far too high to accept NCBE's assertion that a score earned in one UBE jurisdiction has the same meaning as a score earned in another without independent verification by an entity without a stake in the outcome.[108] Are we to simply accept the claim that there is no difference in the outcome whether a candidate takes the UBE in New York as opposed to North Dakota when over 10,000 candidates take the July bar exam in New York and less than 100 do so in North Dakota? There is only one way to be confident that a 266 earned in North Dakota, Missouri, New Mexico, and Alabama represents the same quality work as a 266 earned in New York — and that is to actually

---

[106] AICPA, *How is the CPA Exam Scored?* (Effective January 1, 2011) See section entitled *"Frequently Asked Questions and Answers", Question 6: How do you score the written communications responses?* http://www.aicpa.org/BecomeACPA/CPAExam/PsychometricsandScoring/ScoringInformation/DownloadableDocuments/How_the_CPA_Exam_is_Scored.pdf (last visited February 11, 2017). "AUD" is Auditing and Attestation; "FAR" is Financial Accounting and Reporting; "REG" is Regulation; "BEC" is Business Environment and Concepts. AICPA,

[107] AICPA, *How is the CPA Exam Scored?*, *supra* note 106. AICPA, provides candidates and the public with a "non-technical overview of scoring. It is a jargon-free explanation of the scoring process, providing insight into how MST (Multi-Stage Testing) works and including some basic facts about IRT (Item Response Theory)." This explanation includes how multiple choice questions may be of varying difficulty and how the difficulties are accounted for during scoring.

[108] Posting of Nancy L. Reeves, Director of Academic Success Programs, to asp-1@chicagokent.kentlaw.edu (Aug.25, 2016, 1:26 p.m. EST) (The subject heading of this email is [ASP-L:6226] *Re: Need data to assess the "uniformity" of the UBE*) (on file with author).

cross-grade exams (including scaling) and see if they are the same (or at least roughly the same).[109]

The legal academy and law school deans must act to request that state boards of bar examiners request the collection and analysis of this data by independent psychometric experts. There is no time to waste. More jurisdictions are considering joining the UBE roster[110] and others are contemplating changes in cut scores.

V. Conclusion

The critical question of whether the UBE achieves its primary purpose of assessing whether a candidate is minimally competent to practice law and whether it does so with reliability and validity remains shrouded in doubt and mystery. NCBE's attempts to address the question fail to do so because the entity is not forthcoming with its procedures. Instead, they provide irreconcilable conclusions based on insupportable assumptions. In other instances, NCBE deflects attention from the issue by focusing on peripheral matters, thus distracting and delaying us from the discussion that is essential to the future of legal education and admission to the bar.

We must insist that the law licensing exam be a fair and reliable assessment of an individual's minimum competency to practice law. A criterion-based assessment would be one step toward achievement of that goal. It is possible to do this — and it is being done by another high-stakes licensing exam. Every jurisdiction has a stake in the outcome, even if it is not a UBE jurisdiction.

To determine whether the UBE is really a "uniform" exam, it is necessary to evaluate NCBE's claims that despite differing cut scores for admission set by UBE jurisdictions, despite changes in the populations taking the bar exam, despite changes in the content and number of MBE test items, and despite scaling unreliable scores to each

[109] *Id.*

[110] As of Feb. 1, 2017, the UBE has been adopted by 27 jurisdictions. *See Adoption of the Uniform Bar Examination with NCBE Tests Administered by Non-UBE Jurisdictions,* NATIONAL CONFERENCE OF BAR EXAMINERS, http://www.ncbex.org/pdfviewer/?file=http%3A%2F%2Fwww.ncbex.org%2Fdmsdocument%2F196.

other, what remains "reliable" is the assurance that a UBE score represents an individual's competency for the practice of law. If it doesn't, then we must make sure that it does.