Perfect Compromise or Perfectly Compromised Tests: Law School Examinations that Mimic a Bar Examination’s Format?

Charles J. Senger

Available at: https://works.bepress.com/charles_senger/1/
Cramming bar examination type essay and Multistate Bar Examination (MBE) multiple choice questions into a three or four hour law school examination looks like a perfect compromise. Benefits include increased coverage, facial validity due to similarity to many bar examinations, the opportunity for students to practice in an examination format that they will face later on the actual bar exam, and reduced scoring time compared to an all essay exam. What is not to like? In a word, plenty. Unless a number of challenges and potential problems are addressed, a perfect compromise can turn into a perfectly compromised test.

This article will raise, explain, and offer suggestions for six challenges, each of which can seriously undermine an otherwise good measurement instrument. Those six are: 1. Mismatches between the examination’s questions and scoring methods when compared to teaching, curriculum, or institutional goals; 2. Insufficient sample size; 3. Insufficient resources for drafting, pre-testing, and processing; 4. Two or three-choice, multiple choice questions masquerading as four-choice questions; 5. Unintended clues or cross contamination among questions; 6. Failure to maintain the stated scoring weights for each portion of the examination.

---

1 Charles J. Senger is a Senior Professor of Law at Thomas M. Cooley Law School. After his law degree, his Masters was in Measurement and Evaluation and his Ph.D. was in Educational Psychology. The author wishes to thank reference librarians John Michaud and Allissa Raasch Schmidt for their invaluable help and insights, and Professor Eileen Kavanagh, Philip X. Senger, and Sharon L. Senger for their review and comments on prior drafts.
Underestimating the impact of these challenges easily can result in far too much random error in the distribution of grades influencing career changing positions with law review, law firms, and judicial clerking positions. The heart of many of the suggestions offered here will be a sort of back to the future exhortation to apply carefully what is already known from the history of law school examinations. Although any examination can be analyzed in a number of ways by a measurement expert, these particular challenges, and the suggested responses to them, all fall within the range of what law professors should be able to tackle. The goal is producing superior law school examinations even if both essay and multiple choice questions are included in a restricted time format.

I. Mismatches Among Test and Teaching Goals

The history of law school essay examinations includes Dean Langdell’s “specific questions to measure specific knowledge of points of law,” James Barr Ames’s questions where “a greater range for student creativity became possible,” and “[a] modern question [that] often demands the students to work with complex facts, which disguise multitudes of claims and defenses.” As noted by Sheppard, methods of evaluating the responses produce differing views in the literature. Likewise the format of law school multiple choice questions ranges from the true-false questions studied by Wood at Columbia, to the four choice questions of the MBE.

For a given course, with given teaching goals, a set of particular types of questions, with appropriate scoring methods, will be the best match between teaching and

---

3 Id. at 674.
4 Id. at 677.
5 Id. at 679 n.126.
testing. Visiting classrooms one can observe varying degrees of emphasis given to rules,
policy, deductive reasoning, reasoning by analogy, ethics, drafting, and many other
aspects of being a great lawyer. For example, in one class much time might be spent
teaching students to critically parse judicial opinions. Another might emphasize a
problem requiring the sorting of a tangled skein of facts into a legal framework. Yet
another might focus more on reviewing transactional documents with a goal of learning
how to improve or tailor the documents for a particular set of facts. Hopefully the
examinations in those classes would each include questions that would allow the
examinees to directly demonstrate what they had learned, even though those questions
might not fall within the type or format of a typical bar examination. Thus a case, a
problem, or a sample document, could be included as part of a question. If not, it would
be no surprise if students largely ignored the cases, problems, course documents, and
even much of what went on in class. Instead, if the test does not match the teaching, one
might well find students concentrating instead on black letter law from sources like
outlines or summaries prepared by others. Although the resulting limited learning might
get the students through the bar examination, the overall effect likely would be to hinder
their professional growth throughout their careers.

Evaluation of student responses, in the classroom or on an examination, likewise
will vary, not only in what skills and knowledge are targeted, but also in the degree to

---

7 This list is but a small fraction of the skills and values listed by the American Bar Association task force. SECTION OF LEGAL EDUC. & ADMISSION TO THE BAR, AM. BAR ASS’N, LEGAL EDUCATION AND PROFESSIONAL DEVELOPMENT – AN EDUCATIONAL CONTINUUM, REPORT OF THE TASK FORCE ON LAW SCHOOLS AND THE PROFESSION: NARROWING THE GAP (American Bar Association 1992).
which the evaluation method is primarily analytical or holistic. In other words, the evaluation method of some professors will focus more on an explicit checklist of individual points whereas others will focus more on an overall, holistic evaluation of a response. In one class a professor might teach with the assistance of lists of elements or factors with explicitly designated weights. A response then could be evaluated by checking off the items on the list. The same would be true for feedback after an examination. Another professor might instead assign an overall mark or score based on consideration of the response as a whole. Selected responses could provide both anchoring standards for the professor and helpful feedback for the students. Thus a “B” or “C” rating each would have a sample response or responses showing what the professor would expect for that score level. Further explanation of the assigned mark or score then would involve individual discussion of that particular response, either in class, or in a conference with the student after the examination.9

Much the same can be observed at a curriculum or institutional level. Sometimes the differences are explicit, as when an institution offers more than one curriculum for the first year.10 More often an institution will develop along lines favored by principal decision makers. The resulting institutional structure can favor viewpoints including, for example, that of a litigator, a transactional lawyer, a policy maker, or a law and economics analyst. Some would even argue that selecting a given approach can change, not just the curriculum, but also both students and faculty along predictable lines.11

---

10 For example, the choice of an “A” or “B” curriculum for first year students at Georgetown. See Georgetown University, Georgetown Law – Introduction to the JD Program, http://www.law.georgetown.edu/curriculum/jdprog.cfm#First (last visited Apr. 21, 2010).
Since the bar examination format is optimized to make a single decision\textsuperscript{12}, pass or fail, substantial format modifications could be predicted as necessary to produce the full range of law school grades, even if all the additional variations noted above were not present. No matter how appropriate a particular bar examination format may be for a given state, it is extremely unlikely to match all the important goals of a particular law school situation. In addition, using MBE format of four-choice, multiple choice questions raises further challenges that will be addressed shortly. In any case, if a law school examination blindly takes the bar examination format as a model, the mismatch can fail to test important course goals and seriously mislead students to their detriment. Since examinations often are made public after the test\textsuperscript{13}, the longer term effect again can be that students will ignore desired learning outcomes and focus instead on the more limited goals of the test.

This challenge leads the list because the ease in stating it contrasts dramatically with the difficult in handling it. Those new to academia can take guidance from the political scientist Wallace S. Sayre who said: “Academic politics is the most vicious and bitter form of politics, because the stakes are so low\textsuperscript{14}.” Those with experience in the field can only smile when thinking of all the different turf battles implicit in even discussing the problem. Despite that reality, the suggestion still must be that mismatches can be reduced if professors, and where appropriate administrators, being aware of the potential pitfalls, engage in thoughtful discussions seeking the best possible compromises amid conflicting and strongly held positions. The result of that process should be much

\textsuperscript{13} Sheppard, supra note 2, at 681-82.
\textsuperscript{14} THE YALE BOOK OF QUOTATIONS 670 (Fred R. Shapiro ed., Yale University Press 2006).
better than what would follow from the blind adoption of a particular examination type, or format, even if that examination type, or format, is as well established as that for bar examinations.

II. Insufficient Sample Size

As important as the previous challenge is, the next one is even more important, albeit not as easily stated. Even the mere mention of sample size might conjure up visions of nightmare statistics classes complete with mathematical formulae only a few law professors could love. Fortunately law professors need not dive deep into statistical theory to appreciate both the need for adequate sampling and the impact of too small a sample size.15 This section will begin with a brief explanation of the importance of sample size in measurement, and then take on its application to law school examinations.

As a start, unlike a law school examination, some measurement problems lend themselves to relatively simple solutions. For example, a professor might want to determine if first term life style changes lead to a weight gain by law students. Determining changes in weight over a term could be done fairly quickly in short measurement sessions. Each session could include multiple measurements of each student that, when averaged, would yield a good estimate both of the student’s weight and of any gain over a term. True, it would still be an estimate because the measurement sessions might have disproportionately fallen on days when a student’s weight had varied more than usual due to an unusual eating or drinking pattern the prior day. Still, a good weight scale of known reliability could be used, and repeated measurements with it would neither wear out the scale nor have a serious impact on its reliability. Thus

15 For an introduction, with an example, of how to use a formula to determine the optimum number of items on a test, see Linda R. Crane, Grading Law School Examinations: Making a Case for Objective Exams to Cure What Ails “Objectified” Exams, 34 NEW ENG. L. REV. 785, 795-97 (2000).
repeating the measurements over a number of days would reduce the risk of measurement errors and increase the precision of the estimate of the target quantity being measured, the weight change, if any, of the students.

Law professors, of course, have a much more difficult measurement task on law school examinations. They seek to measure a combination of knowledge and skill as demonstrated by persons who are affected by a number of possible confounding factors like fatigue, stress, motivation level, less than optimal examination environments, physical and mental challenges, and relationship issues. Obviously a law professor cannot average out errors by simply giving the same test two or three times to a student. Not only are multiple testing times typically not available, but repeated tests may result in an examinee learning the answers to questions. The other confounding factors also may change with repeated measures of a law student. These changes again can affect the very performance of the examinee that one is attempting to measure. A student’s score may drop over time, for example, due to fatigue from repeated testing.

Some professors, and attorneys in practice, might respond that confounding factors like fatigue, stress, relationship issues, etc., equally affect the practice of law. Thus, it might be argued, those confounding factors on an examination merely reflect normal challenges in the life of a service professional. At its heart, this is an argument by analogy. Its basic thrust carries force especially for those with experience in the practice of law. Also to be acknowledged, however, will be the possible devastating effect of a single, high-stakes test administered once in just a few hours. Yes, that can happen also in practice. For example, some judges or partners never forget an error, no matter what the actual facts or extenuating circumstances might have been. Viewed in another light,
however, these examples from practice can be taken to urge even greater care in law school testing. Professors, like judges and partners, share the human condition, with its associated risks of flawed data collection and interpretation. No matter what might happen in practice, therefore, professors should strive to limit the unfair effects of confounding factors in the one measurement context, examinations, which they do control.

Professors thus must place a high premium on getting as good a measurement sample as possible when they test. On the one hand they know that they would have great confidence in the measurement if they could somehow repeatedly test each and every thing that the student needs to know. On the other hand, they know that a very short test could elicit a perfect performance by a marginal student even though it was based totally on guessing or on matching exactly the student’s minimal knowledge. In between the two extremes, professors make difficult sampling decisions regarding what to test, and how many questions to ask, often to fit within testing time constraints imposed by the institution rather than by the professor.

At first glance, adopting a bar examination format for a law school examination would appear to improve the sampling of student competencies. Using shorter essay questions, and including multiple choice questions which can be answered in a pencil stroke, allows more measurement points in a given amount of time and thus a larger sample of measurements\(^\text{16}\). Ironically, when the two question types are combined in one examination, the shorter time frame of many law school examinations leaves little time

\(^{16}\) This article focuses on the typical bar examination format. In the more general context, Redlich and Friedland present arguments for including both essay and objective questions on law school exams: Norman Redlich & Steve Friedland, Challenging Tradition: Using Objective Questions in Law School Examinations, 41 DePaul L. Rev. 143, 149-51 (1991).
for either type of question, and thus emphasizes the weakness of each type rather than its strengths. A closer comparison of the typical bar examination and law school examinations will make this clearer.

The MBE contains two sections of one hundred questions each, administered in two three-hour blocks, with an average of 1.8 minutes available for each four-choice multiple-choice question. If no performance questions are used, a second day typically is devoted to answering ten to fifteen essay questions. If the day again consists of two six-hour blocks, this allows twenty-four to thirty-six minutes to respond to each essay question.

If law school examinations are given in time blocks tracking with those of bar examinations, then the same amount of time will be available for sampling using both essay and multiple-choice questions. Difficult sampling decisions then can be guided by the experience and research of bar examiners. Practically speaking, however, those testing time blocks often are not available in law schools, especially if examinations are proctored rather than being given as take-home examinations.

If a bar examination’s twelve-hour design is crammed into something like a three-hour, law school examination period, far fewer essay and multiple choice questions can be asked. A three-hour examination provides, at best, one hundred eighty minutes of examination time. Allowing for classroom turnover between exams, the actual testing time may be only one hundred fifty to one hundred sixty five minutes. With question timing similar to the bar examination, in a 165 minute time period, only four, twenty-four

---

minute essay questions, and about thirty-eight, 1.8 minute multiple-choice questions could be administered.

The primary weakness of essay questions is that the time necessary to draft a response limits the number of questions that can be asked within given time limits. Being able to test certain aspects of a student’s knowledge, but not others, introduces the possibility that the student may do well or poorly based upon which aspects by chance happen to be tested on a given test. In response to the time pressure, students sometimes prepare generic paragraphs that can be used to speed the writing process and even to bluff their way through parts of a question. The primary weakness of multiple choice questions is that random error can be introduced from guessing. For both essay and multiple choice questions, increasing the number of questions asked, within reasonable limits, increases the sample size and reduces the average amount of random error.

Since both essay and multiple choice questions are included in the same examination, the assumption, correct or not, appears to be that the two types of questions do not measure the same things. If they did, using both question types would reduce measurement error by allowing a larger sample size of measurement points. If they measure different things, however, then including both types, within a limited law school exam time period, can increase random error by preventing enough of either type of question to be asked. Perhaps an analogy could be drawn to a pairing of pliers and

---

scissors: each does a different task well, but a combination formed by half of each does a far worse job on those tasks than either taken alone.

One obvious way to increase the number of questions asked, and thus the sample size, is to add a midterm examination. Adding a midterm, however, raises other concerns that merit brief discussion. The author’s first use of a formal midterm was in 1999. Challenges included lining up the required administrative support and dealing with objections from colleagues who felt that the examination adversely affected the preparation of students for other classes during the week of the midterm. Adjusting the length of breaks between class periods and adding a make up session restored the lost teaching time, but it was not easy. The midterm consisted of all multiple choice questions, and the final exam later was part multiple choice and part essay.

For the author, midterms yielded both some problems and a number of benefits. Scheduling a midterm obviously impacted student preparation for other classes during the week of the test. Other professors made their feelings about that well known. Midterms also required scheduling of proctors, modifying score reporting for record keeping, and making changes in how appeals could be handled at the end of the term. But the midterm dramatically increased the sample that could be taken whether in reference to total testing time available or to the number of questions that could be asked. For first term students, a midterm also served as a wake-up call alerting the students to the demands of law school testing. In the author’s opinion, this wake-up was especially important for first generation law students. Thus it partially leveled the playing field with those students.

This harkens back to the early days of law school when quizzes were frequent, as, for example, at Harvard under Asahel Stearns. Sheppard, supra note 2, at 666; Steve H. Nickles, Examining and Grading in American Law Schools, 30 ARK. L. REV. 411, 460-62 (1977). Midterms also play a key role in the grading process, as advocated in: Charles B. Sheppard, The Grading Process: Taking a Multidimensional, “Non-Curved” Approach to the Measurement of a First-Year Law Student’s Level of Proficiency, 30 W. ST. U. L. REV. 177, 182 (2003).
who came from backgrounds steeped in the law. Finally, a number of students told the
author that they appreciated the chance to get part of the examination behind them rather
than having their whole grade riding on their performance on just one day. Intuitively the
students were recognizing that taking two samples, on two different days, increased the
reliability of the testing process by reducing the impact of random factors like a head cold
on the wrong day. An even stronger example is that of beginning students experiencing
their first-ever panic attacks on a midterm. Not only is the partial score of the midterm
able to lessen the impact of the unexpected attacks, but it also allows time to address the
underlying condition before the whole block of examinations at the end of the term.

Other suggestions on increasing sample size will follow in the context of
discussing four-choice multiple choice questions that really are functioning as three or
even two choice multiple choice questions. Those suggestions will be in section IV infra.

III. Insufficient Resources for Drafting, Pre-testing, and Processing

Even if mismatches of teaching and testing goals are minimized, and even if
sample size concerns are addressed, a major problem remains. In a nutshell, drafting a
bar examination demonstrably consumes vast resources not available to law school
professors. If law school professors assume that they will be able to come even close to
matching the challenges faced by the bar examination drafters, then those professors will
need all the help that they can get.

The difference in resources available to bar examiners and law professors is eye-
popping. Over twenty years ago, a court found that it costs “hundreds of thousands of
dollars to develop each 200-question MBE test form.”\footnote{National Conference of Bar Examiners v. Saccuzo, No. 03CV0737BTM(NLS), 2003 WL 21467772, at *2 (S.D. Cal. June 10, 2003).} Likely the costs today only
would be higher. A law professor could only dream of such a testing budget. In that
dream, a professor suddenly would have the support staff, not just for research and
drafting assistance, but also to conduct pre-testing of questions, followed by extensive
quantitative and qualitative study to refine questions and insure that examination scores
would be consistently interpreted from one term to the next. Budgets of that size, in other
words, would enable expert assistance allowing classroom tests to make a quantum leap
upward in measurement sophistication.

In the real world of law school education, however, those dreams might have to stay dreams. But the suggestion here is not to give up so easily. Although what is done on the MBE cannot automatically be transferred by analogy to the law school context, various intermediate steps have support in professional education. For example, faced with similar problems, some medical schools have created special departments with positions staffed, in part, by educational psychologists who assist medical educators in a number of ways, including expert advice on measurement and evaluation issues. This does not replace the expertise available from an associated university community, but it provides in house expertise tailored to the needs of medical educators. The always difficult budget challenge becomes somewhat easier since the costs can be divided over several different fractional positions. For example, both a Director of Legal Education and a Director of Assessment and Evaluation might not be possible, but a half position for each might be.

23 The high standard of excellence achieved by the MBE will be discussed further in section IV, infra pp. 18-26.
Even lesser budget demands are made if a position is carved out of an existing staff position. An engineer doing computer system support, for example, or a professor doing some types of research, may have a strong background in statistics. Hiring decisions can be influenced by such qualifications, and a job description can be built to allow the person to devote some time to assisting with statistical issues. This could help not only with testing, but also with faculty whose research would benefit from like assistance.

Faculty committees are another possibility. Over thirty years ago, a survey of law schools asked: “Does your law school have something in the nature of a committee on examination and grading techniques?”

69 percent answered no. Likely today the percentage would be similar, especially if one looked at what a committee, despite its title, actually did. Such a committee could be formed more easily if some professors had the required background.

On a still more economical level, experts could be brought in on a regular basis for both training and individual consultations. Over time, faculty expertise would increase, the opportunity for peer assistance would develop, and recurring problems gradually would be minimized.

One possible argument on behalf of law school professors is that they can produce high quality, four-choice items by gradually refining the questions over the course of a number of examinations. This is possible, just as it is for the MBE. Often overlooked,

25 Nickles, supra note 20, at 423.
however, is the need for tight examination security for this to work. Although this is the subject of a whole article, the main idea is that students remember questions, talk about them, and thus create unequal access to information for future examinees. The result is that a given question may test, in part, whether a student had heard something, even innocently, rather than whether the student is competent.

In any case, recalling just the size of the MBE budget helps law professors maintain a reasonable perspective when they attempt to use bar examinations, including the MBE, as models for their own examination practices. Enormous resources and expertise are devoted to the production of the four-choice questions on the MBE. Professors might well proceed cautiously if they assume that they can produce the same quality of questions needed to fulfill their testing responsibilities. Although the concern extends to any type of multiple choice question, it becomes especially acute when sample size is reduced by combining essay and four choice multiple choice questions in the time allotted for typical law school examinations.

IV. Two or Three-choice, Multiple Choice Questions Masquerading as Four-choice Questions

The advantages of four choice multiple choice questions, like the MBE, come easily to mind. Following a well-established format lends credibility to the testing. Juggling multiple possibilities also invokes a desired characteristic of legal thinking. Increasing the number of choices reduces the probability of guessing the correct response, especially for lower ability students. As Lord stated it: “The effect of decreasing the number of choices per item while lengthening the test proportionately is to

increase the efficiency of the test for high-level examinees and to decrease its efficiency for low-level examinees.\textsuperscript{28}

Unfortunately the advantages of four choice questions also have offsetting disadvantages. One often overlooked challenge is that of drafting incorrect choices, distractors, such that each of the choices carries its data gathering responsibility of discriminating\textsuperscript{29} among examinees on the basis of their competence. In other words, if a choice is selected by almost no one, then precious testing time has been used that could have been devoted to more direct discrimination among those who are and are not competent. Naturally the unselected choice still could be performing some testing purpose. A choice selected by no examinee, however, does not help in the vital task of discriminating, and thus is of less value than choices that do discriminate.

Professionally prepared examinations, like the MBE, minimize the number of nonfunctional choices to an extent that would be the envy of classroom examiners. For example, in the six MBE examinations given between February 1989 and July 1991, only one examination had as many as five questions with nonfunctioning choices\textsuperscript{30}. In the February 1991 examination, there was only one such question on the 200-question examination\textsuperscript{31}! It is an understatement to say that those MBE examinations represent a quantum jump in quality over what is normally achieved in examinations reviewed by the author, including the author’s own examinations.

\textsuperscript{28} \textsc{Frederic M. Lord}, \textit{Applications of Item Response Theory to Practical Testing Problems} 110-12 (Lawrence Erlbaum Assoc., 1980).
\textsuperscript{29} Lynn M. Daggett, \textit{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}, \textit{57 J. Legal Educ.} 391, 409-412 (2007).
\textsuperscript{31} Id.
If the current excellence of the MBE is kept in mind, then it can be instructive to look at early MBE questions even though the questions do not meet the current MBE standards. The point is to illustrate the challenge of getting all the distractors to do their job, and indirectly to illustrate problems with sample size. The following two questions come from publicly released portions of an early Multistate Bar Examination. The National Conference of Bar Examiners has generously allowed them to be reprinted here. The examples are two parts of a three-part contracts law question:

Questions 18-20 are based on the following fact situation.

Johnson bought 100 bolts of standard blue wool, No. 1 quality, from McHugh. The sale contract provided that Johnston would make payment prior to inspection. The 100 bolts were shipped, and Johnston paid McHugh. Upon inspection, however, Johnston discovered that the wool was No. 2 quality. Johnston thereupon tendered back the wool to McHugh and demanded return of his payment. McHugh refused on the ground that there is no difference between No. 1 quality wool and No. 2 quality wool.

18. Which of the following statements regarding the contract provision for preinspection payment is correct?

(A) It constitutes an acceptance of the goods.

(B) It constitutes a waiver of the buyer’s remedy of private sale in the case of nonconforming goods.

---

34 E-mail from Judith A. Gundersen, Deputy Director of Testing, National Conference of Bar Examiners, to Charles J. Senger, Senior Professor of Law, Thomas M. Cooley Law School (June 17, 2008) (on file with author).
(C) It does not impair a buyer’s right of inspection or his remedies.

(D) It is invalid.

19. What is Johnston’s remedy because the wool was nonconforming?

(A) Specific performance.

(B) Damages measured by the difference between the value of the goods delivered and the value of conforming goods.

(C) Damages measured by the price paid plus the difference between the contract price and the cost of buying substitute goods.

(D) None, since he waived his remedies by agreeing to pay before inspection.

Over a decade ago, the author used these two questions in a two hundred question, practice multistate examination given to one hundred seventy students. Seeing how these two questions fared with those students is instructive. One way to start the analysis of how the questions are performing is by dividing students into those who obtained top scores on the test, those who scored at the bottom, and those in between. 46 of the 170 students formed the top group, 46 the bottom, and 78 remained in the middle group.

The students mainly were in their last year of law school but some were second year students who were getting a head start on preparing for the bar exam. This meant that some students did not feel as much urgency about the exam and some did not have as much preparation as others. Normally, this potentially confounding factor would complicate the analysis. Here, however, it might well strengthen the inferences being drawn. If anything, it would seem that an unmotivated student would be more likely to make careless mistakes. Rather than reducing the number selecting nonfunctional distractors, that carelessness would seem to make those selections more likely. Thus the low number of students selecting a distractor continues to support the inference that the distractor is seriously flawed.

27% was the percentage of the total group for both the top and bottom groups. “If the total test scores are normally distributed, it is optimal to use the 27% of the examinees with the highest total test scores as the upper range and the 27% of the examinees with the lowest total test scores as the lower range.” MARY J. ALLEN & WENDY M. YEN, INTRODUCTION TO MEASUREMENT THEORY 122 (Lawrence S. Wrightsman, Todd Lueders & Cece Munson eds., Brooks/Cole Publishing Co. 1979) (1979). Sometimes simpler methods have been used in legal education literature. For example, 200 examinations were sorted into five groups of forty each based on each examination’s total score: Paul E. Spayde, Frank R. Strong, & W. R.
If these groups are used to analyze how each question choice performs, the following is revealed. The correct response for question 18 was (C). This choice was selected by 100 percent of the top group, 95 percent of the middle group, and 83 percent of the bottom group. Distractor (A) was selected by no one from the top group, only 1 percent from the middle group, and 13 percent from the bottom group. For distractor (B) the respective numbers were 0, 4, and 2 while for distractor (D) they were 0, 0, 2.

How well are question 18’s, incorrect responses functioning? At a minimum, each is contributing some discrimination information because no top student selected any of the distractors but at least some lower performing students did. But given the precious testing time consumed by each distractor, are these distractors carrying their weight? Rounded off, two percent of the 46 students in the bottom group would be one student. If only one student out of the 170 selects distractor (D), is that enough to keep it instead of using that time more productively? Dropping distractor (D) would make question 18 a three-choice item. Although more students chose distractor (B), is it enough to keep it, or would the test be stronger and more efficient if it were a two choice item and more questions were asked?

Turning to question 19, the correct response, (C) was selected by 46 percent of the top group, 28 percent for the middle, and 17 percent of the bottom group. For distractor (A), the numbers were 2, 14, and 24. For (B), they were 52, 54, and 54. Finally, for (D), they were 0, 4, 4.

Again, looking at the pattern of responses, might not this question perform better as a three-choice question, or maybe even two, two-choice questions? For the MBE,

---

Flesher, The Validity of an Objective Examination in Constitutional Law, 1 J. LEGAL EDUC. 251, 253 (1948).
extensive test drafting resources and expertise allow questions like this to be answered by completely redrafting, or even dropping, the question. But if these problems happen even rarely on the MBE, despite their expertise and intensive pre-testing of questions, might not a law professor be better off drafting questions with fewer distractors? From what is known based on current research, the shorter questions would be more efficient in gathering the needed competency data. Reducing the amount of reading for each question at least allows a larger sample size because more questions could be asked in a given amount of time.

As often happens in history, questions with fewer distractors have been used in the past to increase sample size. Wood, in his classic study at Columbia, states that the examinations averaged 200 true-false questions. Over twenty-five years ago, Dr. Robert Ebel suggested a further refinement to the author. Dr. Ebel had extensive experience with standardized testing and had both researched and published in support of the use of true-false questions. However, in one of his measurement classes he suggested that it would be even better to use two choice multiple choice questions. These also are called alternate choice items. The advantage was that the two choices would focus on the precise point being tested. The examinee thus would not have to guess at

---

37 Adams, supra note 30, at 5.
38 A good start on this research is: Michael C. Rodriguez, Three Options Are Optimal for Multiple-Choice Items: A Meta-Analysis of 80 Years of Research, EDUC. MEASUREMENT: ISSUES AND PRAC., Vol. 24, Issue 2, Summer 2005, at 3. In the future, further research may lead even the bar examiners to use three choice items in order to gain greater efficiency. Adams, supra note 30, at 8.
40 For more on Dr. Ebel and his place in testing, see Gregory J. Cizek, Linda Crocker, David A. Frisbie, William A. Mehrens & Richard J. Stiggins, A Tribute to Robert L. Ebel: Scholar, Teacher, Mentor, and Statesman, EDUC. MEASUREMENT: ISSUES AND PRAC., Vol. 25, Issue 1, Spring 2006, at 23. On a personal note, if Dr. Ebel had not died an untimely death in 1982, my program in educational psychology would have been very different.
41 Id.
what portion of a statement was being tested as being true or false. The example he gave was:

In a sealed, insulated small room, a refrigerator is running with its door open. The temperature in the small room will go:

A. Up.

B. Down.\textsuperscript{43}

He said that he chose the example, not just to illustrate the format, but also to show that the two choice question could test more than just the recall of knowledge.

If one minute is assigned for responding to each question, 200 true-false questions could be answered in three hours and twenty minutes instead of the six hours required by the MBE. Reducing the number of choices therefore can substantially increase the sample size obtained in the amount of time often allotted for a law school examination. Of course the key is still good items of whatever type. If a question’s subject matter facilitates drafting four choices, then a four choice question obviously will have strong measurement advantages. If fewer distractors are available, the suggestion here is to break out of the bar examination format and use a question with fewer, but good, distractors\textsuperscript{44}. As an added advantage, what the student sees, and thinks is being scored, is consistent with what is actually being scored. If it looks like a four-choice question, or a two-choice question, then that is what it really is.

V. Unintended Clues or Cross Contamination among Questions

\textsuperscript{43} A four-choice version of this question is presented in: ROBERT L. EBEL, ESSENTIALS OF EDUCATIONAL MEASUREMENT 194 (Prentice-Hall, Inc. 1972) (1965).

\textsuperscript{44} Mixing into a test items with different numbers of distractors now is more feasible due to the increased power and flexibility of available computer resources. MEHRENS & LEHMANN, supra note 9, at 287.
Stating this challenge goes a long way toward explaining it. Producing four attractive choices often means that clues to the correct analysis, and near alternatives, are outlined in the choices. Four choice questions are fertile ground for unintended clues. One type of clue can flow from the common tendency to position the correct response in a middle position. Attali and Bar-Hillel explain this problem, as well as the corresponding tendency of test-savvy examinees, when they have to guess, to demonstrate middle bias or edge aversion. In other words, this potential clue can be avoided by assuring that correct responses are equally distributed across the possible choice positions.

Unintended clues also can be avoided by eliminating certain types of four choice questions. A good example would be what are called “double multiple-choice questions” or tripartite questions. The three parts of the question are: facts; conclusions, usually listed with Roman numerals; and choices that combine various conclusions. For example, choice “B” could be: “II and III,” whereas choice “C” could be “II only,” and choices “A” and “D” would not mention “II” but would instead include some combination of “I” and “III”. The key is the four choices of the question cannot list all the possible combinations. This opens the door to problems. As explained by Gensler, persons with the same amount of partial knowledge could score very differently on the question. Assume, for example, that two students each know only one

---

46 Id. at 111-16.
49 Id. at 28-29.
piece of relevant knowledge. The first student knows only that choice “I” is wrong; the
second knows that “II” is wrong. Assume further that “I” is mentioned only in choice
“A.” The first student can use the one piece of knowledge only to eliminate choice “A”
whereas the second student’s single piece of knowledge can eliminate both “B” and “C.”
The first student thus has a one in three chance in guessing the correct response while the
second student’s chances are one in two. The incomplete listing of possible alternatives
has given the second student a valuable clue that is not available to the first student. The
result is that similarly situated students end up with very different chances on the test.
This type of question no longer appears on the MBE\textsuperscript{50}.

Unintended clues can slip in due to the way that the distractors are drafted in four
choice questions. Beginning with a correct answer, a professor may systematically
modify that answer to produce a set of three distractors. Great caution must accompany
the effort or the set of responses can converge toward the correct answer. Overlaps in the
distractors, created by the systematic modifications, can lead to the convergence. The
end result is that two of the responses will be much more plausible, at least to a test-wise
student. Again this means that an item that looks like a four choice item really is
performing as, at most, a two choice item. Also it means that the item is testing how test-
wise a student is, rather than what it is intended to test and reward\textsuperscript{51}.

Avoiding the tendency to put correct responses in the middle, to use tripartite
questions, or to converge distractors toward the correct response will produce better
multiple choice questions. In addition, care in drafting can minimize the chance that the

\textsuperscript{50} Donahue, supra note 32, at 26.
\textsuperscript{51} Jeffry K. Smith, Converging on Correct Answers: A Peculiarity of Multiple Choice Items, 19 J. EDUC.
MEASUREMENT 211, 211 (1982).
distractors of multiple choice questions will give unintended clues to other questions, especially other essay questions.

At first glance, it would seem that the most serious unintended clues to other essay questions could be avoided by examining different topics on the multiple choice and essay questions. Although this can help, some problems remain. For example, the fact that a topic does not appear in the multiple choice questions could itself be a clue to expected coverage in an essay question. In addition, one reason to use a composite examination with both multiple choice and essay questions is to increase sampling on the key topics of the course. Testing a topic with both essay and multiple choice questions allows one to compare the data obtained from each question type when deciding on the examinee’s true level of competence. Excluding a topic from either section of an examination automatically eliminates the chance to compare data types on that topic and thus undermines one good reason for giving a composite examination.

Separating the administration of multiple choice and essay questions more directly limits unintended cross-over between the question types. On the MBE, for example, examinees do not have possession of the multiple choice questions when they do the essay questions. The same can be done on a law school examination by collecting one part of the examination before distributing the other. Separate administration provides the added advantage of controlling how long students spend on each portion of the examination. If, for example, speed reading is not a grading goal, then controlling the time for each examination section helps equalize differences in how fast students read. This is especially true for MBE type questions with their longer fact patterns and four possible choices.
Time constraints and staffing levels may make it impractical to pick up one portion of the examination before another is started. Another possibility is to use two tests booklets of different colors. At the end of one section of the test, students can then be told to close that test booklet and only look at the other one. Proctoring then carries an additional burden but it can be done.

At the very least, sensitivity to the challenge can reduce the impact of unintended clues from one multiple choice question to other questions on the examination. Thus it is one more way to improve an examination while working with available resources.

VI. Maintaining the Stated Scoring Weights for Each Portion of the Examination

In law school examinations that combine essays and MBE type multiple choice questions, the examinations state how much each question will count toward the final score. The challenge then is to assure that those percentages or weights in fact are being properly maintained.

For over 30 years, the author has seen numerous examinations that fail to deliver on the promise to compute scores and grades based on the stated points or weights. Experience indicates that meeting the challenge requires greater care than some accord it. For example, consider a composite examination where the multiple choice questions count a total of sixty points and the two essays each count sixty points. If the multiple choice questions are four-choice questions, then, over the long run, pure random guessing would yield a score of 15 out of 60. For sake of discussion, let us assume that the actual

scores range from 15 to 60. These multiple-choice point totals crank out automatically in
the scoring process based on a scoring key prepared by the professor. In the author’s
experience, the responses each are scored 0 or 1 without any additional weighting or
exercise of discretion by the professor.

Scoring the essays varies much more across professors due to increased discretion
embedded in the process. Even a professor using a check-the-box scoring checklist still
must decide whether what is in the bluebook is sufficient to receive credit for a particular
checklist box or not. In addition, if a student makes more than a cursory response, a “0”
is unlikely to be given as a score. Thus, practically speaking, a thirty-point essay
question seldom scores over the full range of zero to thirty.

The total point score ranges for a thirty-point essay question thus varies across
professors. Some might consider 70%, or 21 on a thirty-point question, as the cutoff for a
passing score. Their scale might then run from 18 to 30. Another might divide the
question into six parts and score each part on a scale of one to five. Their scores then
would range from six to thirty. Still another professor might score most of the
examinations across a narrow range, perhaps from 20 to 25, and then award only a very
few high or low scores outside that range.

It is not necessary to probe the decision of which essay scoring method to use in
order to see that the variation in scoring methods will raise challenges when those essay
scores are combined with the multiple-choice scores. In brief, the problem is that the
stated point values for each type of question may mislead a student as to the actual weight
that the question type will carry in producing a final total score and grade.
The most extreme case arises when the essay scores predominately fall into a narrow scoring range. For sake of discussion, let us assume that the few high and low scores may track with the multiple-choice scores, and thus there is a good fit between the multiple-choice and essay scores. So far, so good. For the rest of the scores, however, there may be a problem. In effect, lumping so many essay scores together reduces the ability of the essay test to discriminate among all the grade levels the test is intended to measure. For a top “A”, or a bottom “F,” the essay test is doing fine. The points available to discriminate among the rest of the grades, however, are limited. If the multiple-choice questions are discriminating appropriately, then their wider score range will swamp the small score differences among the tightly banded together essay scores. More generally, a problem can arise whenever the multiple-choice and essay scores are simply added together even if their distributions are very different.

The mischief created is twofold. First, the scores on the questions are impacting grades differently than what a student would expect given the point allocations specified on the examination. Second, students can draw incorrect conclusions about what abilities they have to improve in order to become better lawyers. If, for example, multiple-choice scores overly determine grades, students may mistakenly conclude that their problem is in how to take multiple-choice tests. Too much of their time then might be spent on test strategies rather than on learning how to think and problem solve like a lawyer.

Although technical procedures are available to adjust essay and multiple choice scores, typically it is difficult to use them in a law school setting. Raw scores can be converted to scaled scores\(^5\), and the originally assigned weights can be preserved for

\(^5\) A good starting point is: ALLEN & YEN, supra note 36, at 148-174. If one wishes to delve deeper and explore situations where advanced procedures may make little difference, see James S. Terwilliger &
both the essay and multiple choice questions. Students, and even many professors, however, have no experience with the procedures and therefore may not trust them. On the MBE, scaling and equating methods are a given\textsuperscript{54}. In law schools, the same culture usually has not developed, and professors therefore do not have the same level of support if they seek to implement similar procedures. If a professor nevertheless adopts procedures similar to those used on the MBE, the professor probably better be ready to endure some frustrating challenges.

In the end, careful preparation when drafting the examination is the best way in law school to prepare a test using both essay and multiple choice questions while still maintaining the assigned weights for each type of question. In particular, essay scoring protocols should be coordinated with the expected multiple choice scoring range. If a set of essays is expected to count the same as a set of multiple choice questions, then it should produce a similar range and spread of scores. If that is not feasible, then the weight of the essay questions should be adjusted to reflect how they actually will perform. Although it may be impractical to do anything after an examination has been given, monitoring score distributions, over time, will assist a professor in constructing better tests in the future.

**VII. Conclusion**

Bar examinations divide examinees into two groups: those who are competent and those who are not. Law school examinations divide examinees into multiple reported categories of competence. The different measurement goals urge caution in assuming

that what works for the bar examination automatically will work equally well for a law school exam. This article isolates six challenges where that caution is especially appropriate. None of the six challenges prevent the bar examination format from being used on a law school examination. None of the challenges exceed the grasp of a law school professor seeking to improve examinations. If careful attention is not paid to the challenges, however, they do explain why the resulting law school examination can be a measurement disaster.

The effect of flawed examinations on law students should not be underestimated; poor examinations improperly fail or misclassify students. In one study of medical school examinations that violated common drafting principles, the author stated:

One can conclude that some students – perhaps as high as 10-15% of students tested – were incorrectly classified as failed when they should have been classified as passed, due solely to flawed item formats and the ineptitude of test item writers. A false negative rate this high seems unreasonable, given the relative ease and low costs associated with rewriting flawed questions…”

If this happens in medical school, it is difficult to say that it would be any different in law schools.

The damage from seriously flawed examinations can be lessened by addressing each of the six challenges presented here. The six vary in difficulty and in the degree to which success can be achieved in meeting them. The natural starting point was to look more closely at the differing goals of bar and law school examinations. This matching of teaching and testing goals asks no more than must be done in preparing any good law school class or examination. Sensitivity to the potential problems, and care in addressing

---

55 Steven M. Downing, The Effects of Violating Standard Item Writing Principles on Tests and Students: The Consequences of Using Flawed Test Items on Achievement Examinations in Medical Education, 10 ADVANCES IN HEALTH SCI EDUC. 133, 141 (2005).
them, likewise go a long way toward addressing two of the other challenges. Thus
monitoring the performance of questions can lead to the substitution of two or three-
choice, multiple choice questions for the bar examination’s four-choice format. Careful
review of questions, and their results, can minimize unintended clues or cross
contamination among questions.

Meeting two of the challenges may require specialized help from institutional or
outside expertise. Law professors never will have the testing resources and support like
those available to the National Conference of Bar Examiners. Hiring professors with
special skills, university resources, and faculty committee efforts can help. Targeted
training or consultation by experts can offer another economically feasible method of
tailoring the bar examination format to the requirements of a particular situation.
Likewise many law professors will need at least initial guidance in assuring that the
questions on the examination indeed are weighted as the examinees are led to believe.

The last challenge, sample size, always requires continuous attention and effort.
Reducing the number of choices on the multiple choice questions, and adding a midterm
examination, can increase sample size and reduce the effect of random measurement
error. Given the nature of the challenge, all examination drafters struggle with sampling
issues, and always will. The bar examination format offers the hope of easing the
severely limited sampling size of essay question examinations when contrasted with the
use of multiple choice questions. Including both essay and multiple choice questions
potentially leverages the strengths of each, especially in regard to sample size. The key
point here is that, as attractive as that compromise combination may be, it also carries the
risk of producing an examination so compromised that the cure is worse than the disease.
Reviewing what has been written here about these challenges, in the context of the work of Wood, and those cited in Sheppard, evokes a sense that law school examination progress, like learning itself, is cyclical, or even traces a spiral path on which one continually returns to the same concepts, albeit on a higher level. The rate of improvement, however, still appears too low. Thus, unfortunately, calls for improvement, like those of Nichols, still ring true today. The stakes for the public, our profession, and our students have not lessened. The purpose of this article is to assist improvement in our testing. It is a worthy goal to produce a good compromise using the format of a bar examination to improve law school examinations. But producing a seriously flawed examination is unreasonable and professors must strive to meet the necessary challenges to avoid that result. All the stakeholders, especially our students, deserve no less.

56 Wood, supra note 39.
57 Sheppard, supra note 2.
59 Nickles, supra note 20, at 411-12.