New Modes of Assessment

GREG SERGIENKO*

TABLE OF CONTENTS

* Associate Professor of Law, Western State University College of Law. J.D. 1985, Harvard Law School; A.B. 1980, Harvard College. Thanks are due to Gerry Hess and Paula Prather, of the Institute for Law School Teaching, for their comments on assessing skills with multiple-choice questions; to Laura Rovner for her comments on self-analysis, especially from a clinical perspective; to Mike Schwartz and Paula Prather for reviewing an entire draft; and to Terry Roberts for help with the California Bar Exam's procedures. Paul Wangerin graciously permitted review of his presentation at the Institute for Law School Teaching; his work and that of Steve Friedland have been an inspiration to me, even when at times I have taken the liberty of disagreeing with what I understand to be their positions. Errors remaining are, of course, my own.

I. THE IMPORTANCE OF ASSESSMENT

The traditional and dominant mode of formal assessment in law schools is an essay examination administered at the end of the semester.[1] Unfortunately, the essay exam is prone to inaccuracies, some of which can be balanced by other forms of assessment. In addition, essay exams are extremely burdensome to grade.

The purpose of this Article is to call attention to a variety of alternatives to this traditional format that are more accurate and less burdensome than traditional essay exams.[2] Increasing accuracy makes it

1. Robert C. Downs & Nancy Levit, *If It Can't Be Lake Woebegon . . . A Nationwide Survey of Law School Grading and Grade Normalization Practices*, 65 UMKC L. REV. 819, 822–23 (1997); Paul T. Wangerin, *"Alternative" Grading in Large Section Law School Classes*, 6 U. FLA. J.L. & PUB. POL'Y 53, 53 (1993).

2. In describing only less burdensome alternatives, the author does not mean to suggest that these are the only sorts of alternatives to consider. He has described elsewhere the need for performance exams in law school. Greg Sergienko, *Practicing What We Preach and Testing What We Teach*, *in* TECHNIQUES FOR TEACHING LAW 292,

possible to determine whether the instruction has been effective, allowing the instructor to address areas of weakness before the course ends and to improve future classes. Decreasing the burden of assessment of student learning allows for faster feedback, which is more effective.[3] Faster assessment also makes possible frequent assessment, and frequent assessment provides students with the information they need to improve, promoting student learning.[4] Some of these alternatives are formal—that is, used as a basis for assigning a grade—others are not. Others, although informal, can be important educational tools themselves.

This Article starts by discussing ways in which the quality of assessment can be evaluated. Because essay exams are the predominant mode of examination in law school, this Article then turns to a discussion of their strengths and limitations. It then deals with non-instructor assessment and multiple-choice questions as alternatives to essay exams.

II. EVALUATING MEANS OF ASSESSMENT

A. Formative and Summative Assessment

Assessment is of two kinds, formative and summative.[5] Summative evaluation is given at the end of the course and examines how well students have achieved the course goals.[6] Formative evaluation takes place during the course and provides the students and instructors with feedback on how well students are learning.[7]

B. Validity, Reliability, and Practicality

The central concepts in evaluating examinations are validity, reliability, and practicality.[8] Validity is the ability of the test to

292–93 (Gerald F. Hess & Steven Friedland eds., 1999). In fact, a practical ideal of grading involves using a variety of assessment methods to minimize the deficiencies of each. *See* Gerald F. Hess, *Listening to Our Students: Obstructing and Enhancing Learning in Law School*, 31 U.S.F. L. REV. 941, 944 (1997); discussion *infra* part III.B.

3. Wangerin, *supra* note 1, at 65.

4. Hess, *supra* note 2, at 944; Downs & Levit, *supra* note 1, at 823.

5. LUCY CHESER JACOBS & CLINTON I. CHASE, DEVELOPING AND USING TESTS EFFECTIVELY: A GUIDE FOR FACULTY 13 (1992).

6. *Id.*

7. *Id.*

8. PATRICIA L. SMITH & TILLMAN J. RAGAN, INSTRUCTIONAL DESIGN 95 (2d ed.

correspond to the items the test is meant to address.[9] Validity has two aspects, congruence and completeness.[10] Congruence exists if the items on a test agree with the goals of the instruction.[11] Completeness also has two aspects: that the items on the test are representative of the range of items that one could develop for that objective, and that the objectives for the instruction are adequately sampled.[12]

Reliability exists if a test delivers consistent results.[13] If the same test were to be administered on a second occasion, and no learning has taken place as a result of the first administration or of intervening events, a reliable test will produce the same result.[14] Lack of reliability can result from difficulties in grading objectively, guessing on the exam, a short exam (which makes successful guesses on a large portion of the questions more likely), and unclear directions or questions.[15]

Practicality is also an important consideration.[16] A test is practical if it is relatively easy to administer.[17] Impracticality can arise because a form of examination requires considerable time to administer or to grade or because an ideal format involves unacceptable risks. For example, having a student act as sole defense counsel in a felony trial is likely to be impracticable because it would require many days to administer, an equal amount of observation time, and grading by highly qualified and experienced counsel, who are likely to be scarce. It is also impracticable because it involves unacceptable risks to the accused.

To some extent, there are trade-offs among reliability, validity, and practicality.[18] A high degree of reliability and validity may be required on a final exam (summative assessment). Less reliability may be acceptable in providing students with formative assessment during the course, especially if requiring high reliability would mean precluding formative assessment entirely.

C. Norm- and Criterion-Based Tests

Tests are frequently described as being norm-based or criterion-based.[19] Norm-based tests compare students' performance with one

---

1999).

9. *Id.*
10. *Id.*
11. *Id.*
12. *Id.*
13. *Id.* at 97.
14. *Id.*
15. *Id.* at 97–98.
16. *Id.* at 98.
17. *See id.*
18. *Id.*
19. *Id.* at 93.

another.[20] The goal of norm-based tests is to differentiate test takers.[21] Norm-based tests purport to measure general abilities, such as aptitude to learn.[22]

Criterion-based tests (also called objective-referenced or domain-referenced tests) measure students' competence in particular subject matters.[23] Subject-matter knowledge has different aspects. For example, a knowledge of tort law does not imply a knowledge of contract law and a knowledge of strict products liability does not imply a knowledge of the rules on battery. Thus, test results that show takers do better in one subject than another are consistent with good assessment.[24]

Law school assessment usually examines a mix of things; norm-based tests could be applied for some elements, while criterion-based tests could be applied to others. Most law school courses contain a substantial amount of knowledge particular to the course, as do individual subject matters within the course. For these potential items of examination, the standards for criterion-based tests appear to be most appropriate.

Law school courses also attempt to develop skills that would be substantially the same in different courses and in different subject matters within a course. These might include skills in reading a question, in imagining facts to be investigated, and in understanding and making general policy arguments that apply across law school subjects. To these portions of a law school exam, the standard for norm-based tests would appear to be relevant, because all test items would be testing essentially the same skills.

In addition, although an instructor could choose to be guided by norm-referenced standards in testing skills, an instructor could also decide that the assessment should be for ability to achieve a set standard in deploying skills. Such a test should focus on whether the student achieves the standard, rather than on differentiating among students. Because part of any law school course should be criterion-referenced, standards for criterion tests should be considered in assessing the accuracy of testing.

---

20. *Id.*
21. *Id.*
22. *Id.* at 98.
23. *Id.* at 93.
24. *See id.* at 98.

III. LIMITATIONS OF ESSAY EXAMINATIONS

Although essay examinations are the standard method of formal assessment in law school, they have their own strengths and limitations. Teachers are most aware of the strengths of essay examinations. These include the ability to assess writing skills and provide limited clues about the desired answers, thereby requiring students to recall material and generate an answer on their own. An additional advantage of essay exams is that the opportunity to explain an answer may identify an ambiguity in the question that would remain concealed with forced-choice exams.

The limitations of essay examinations result in some way from the strengths of essay examinations. Because students must write out their answers, essay exams are time-consuming to take. Because they are open-ended, they are time-consuming to grade, which causes other problems. Their open-ended nature also makes them difficult to grade reliably.

When considering whether and to what extent to adopt alternative methods of assessment, law teachers need to compare the alternative methods to how essay examinations actually work, rather than comparing them to an inaccurate and idealized version of how they work. This section explores in more detail the limitations of essay examinations.

A. Incongruence with Subjects Taught

1. Sampling Error as to Substantive Legal Knowledge

Essay exams, by requiring knowledge to be organized and written down, are a comparatively time-consuming way of testing knowledge. Because the time for taking exams is limited, an essay examination can test only a relatively small sampling of the course.[25] This means that essay examinations are especially prone to sampling error in the items tested, thereby misrepresenting students' knowledge. This reduces the reliability of the exam.[26]

There is no fully satisfactory way to eliminate sampling error from essay exams. Multiple administrations of essay exams so as to test fully all parts of the course represent a practical impossibility because of the burdens of grading.

Cautious students will minimize this effect by devoting to each subject

25. JACOBS & CHASE, *supra* note 5, at 109.

26. *See supra* notes 11–12 and accompanying text (describing representative sampling as a component of validity).

an appropriate amount of time, but determining the appropriate amount of time is not always easy. A cautious professor would randomly select the subject areas to be tested, so that each area's likelihood of being selected would represent the portion of the course devoted to the area. Few professors do this.

Sampling error cannot be defended as an appropriate way of "punishing" student misallocation of resources. Because professors seldom select exam material randomly, the occurrence of sampling error cannot be blamed exclusively on student misallocation of resources. Moreover, if students misallocate their resources, the appropriate penalty for misallocation is by comparison with the coverage of the subject matter in the course. Sampling error will reward some students—those who "misallocated" their time in favor of the subjects that happened to appear on the exam—and excessively punish others.

2. *Failure to Test for the Skills Taught*

Almost all teachers believe that their goals include teaching the ability to learn new legal material. In fact, most classes devote substantial time to teaching legal skills; skills in interpreting and applying cases constitute a major goal of first-year courses, and skills in dealing with statutory and regulatory materials constitute a major goal of many upper-class courses.

A traditional essay exam calls for the application of recalled law to a factual situation. This tests the ability to read, to identify facts as triggering the application of legal rules, and to write analysis. Thus, the traditional law school exam does not test the ability to interpret and apply unfamiliar legal materials.

As a result of this discrepancy, there is a substantial lack of congruence between the subjects taught and the subjects tested.[27] This removes a motivation for students to learn critical skills and makes tests unrepresentative of students' abilities.[28]

Using performance exams solves this problem by testing for a full range of relevant skills, including reading cases, statutes, and rules.[29] However, the time necessary to apply such legal skills means that performance exams achieve narrower coverage of the substantive law

27. *See* Sergienko, *supra* note 2, at 292.
28. *See id.* at 292–93.
29. *See id.* at 293.

than conventional essay exams.

Even writing, the skill most often cited as a justification for an essay exam, is not well tested by essay exams.[30] Time limits impose artificial pressure on writing, and students are deprived of the tools that they would have as lawyers, such as dictionaries and thesauruses.[31]

B. Limitations in Frequency and Variety of Testing

The traditional essay examination is extremely time-consuming to grade. Because law professors customarily grade their own examinations, rather than delegating the work to graduate students or other readers, professors are likely to regard reducing the burden of grading as very important.[32]

The time-consuming grading process creates several problems. To mitigate the burden of grading, professors tend to assign a single examination at the end of the semester. One examination is less work than several, and grading examinations at the end of the semester, miserable though it is, avoids conflicts with teaching, committee work, and other duties.

This format, the end-of-semester essay examination, has several difficulties. Using a single occasion for graded assessment decreases reliability by making it possible for random factors, such as the students' or professors' personal crises, to have a substantial effect on grading.[33] Using a single occasion for graded assessment also reduces feedback during the semester. "Research shows that frequent evaluation improves student performance on the final exam."[34] Using a single type of examination for graded assessment decreases validity and student satisfaction, because it means that the outcome is substantially influenced by student ability or skill in that particular exam format.[35]

Even if essay examinations were given several times during the semester, the time it takes to grade an essay examination frequently prevents them from being returned quickly. Rewards and feedback are most effective when they quickly follow the work. Returning midterm

30. JACOBS & CHASE, *supra* note 5, at 109–10.

31. *Id.* at 110.

32. *See* Wangerin, *supra* note 1, at 54 ("[M]ost law school teachers will not give the slightest consideration to the use of teaching/grading techniques that call for a larger commitment of grading time than that already spent.").

33. *See, e.g., id.* at 54 n.4 (mentioning students' crises as a factor); Deborah Waire Post, *Power and the Morality of Grading—A Case Study and a Few Critical Thoughts on Grade Normalization*, 65 UMKC L. REV. 777, 802 (1997) (describing faculty's personal crises as influencing grading).

34. Hess, *supra* note 2, at 944.

35. *See id.* at 944.

examinations in two weeks is quite a good rate of turnaround for a professor in a large course, but the reward comes too late to provide instant gratification.

Feedback after a two-week delay is less effective than more prompt feedback. The feedback can correct a student's errors on specific principles of substantive law, because such errors are usually obvious from the answer the student wrote. Feedback meant to improve a student's ability to address legal problems will almost certainly be useless except in the rare occasion when the professor can reconstruct the student's thought processes. This is because students will have encountered numerous legal issues between writing the midterm exam and receiving its evaluation, therefore reducing recall of their strategy in approaching the exam and the possibility of improving that strategy the next time. Students may encounter less legal material between final exams and the results on those exams, but the time delay is likely to be much greater.[36]

C. Inconsistency in Grading

The grading of essay examinations is likely to be highly inconsistent. Greg Munro writes:

> Contrast the common teacher's belief that "I know a 'D' when I see one" with [a] California study's finding that when the same examiner graded an exam answer twice he or she had only a seventy-five percent chance of being consistent in deciding whether an answer passed or failed. [T]his evidence of lack of reliability occurred under a system that has sophisticated techniques for promoting reliability, techniques which are absent in law school grading. Most teachers have probably experienced anxiety about reliability when grading for a long period, grading under fatigue, grading after reading a particularly galling paper, or grading after experiencing anything that changes the assessor's "frame

36. *See* Jon M. Garon, *The Seven Principles of Effective Feedback*, L. TCHR., Spring 2000, at 3–4 (stating that feedback must be prompt enough so that students can apply it). The author's experience supports this conclusion. In Remedies, a course customarily taken in the third year at Western State School of Law, six quizzes and a final were administered. In the initial questions, about forty percent of the students failed to address the call of the question, with the exact percentage varying on the section. Although some have pointed out similar failings on the final exams that the students had already taken, this information apparently had no effect on student behavior. By giving quick feedback—less than a week—and by prominently marking "CALL" on the front of exams that failed to address the call of the question, a reduction in the number of exams with failures to respond to the call of the question from more than forty percent to less than three percent was achieved.

of reference."[37]

Munro also calls attention to the even greater problems occurring in grading by multiple instructors.

> Lack of reliability by reason of scoring inconsistency may be a grave problem in law schools which have no internal coordination among faculty members in scoring exams even when teachers are teaching sections of the same course. Michael Josephson cites a disturbing study involving scoring reliability of the California Bar Exam which showed that a candidate had only a sixty-seven percent chance that two different examiners would agree on whether the answer was a pass or a fail.[38]

The grading of law school examinations differs from grading the bar in ways that make law school exams even less likely to be accurate than bar exams. The California Bar Exams that Josephson studied were graded by assessors who prepare a model answer from their answers to the exam, perform independent research and review a sample of applicant answers, and review dozens of exams before beginning grading.[39] These techniques avoid the problem of an exam question that does not effectively convey the intended meaning and the problem of undue reliance on one person's perspective.[40]

In other respects, though, Munro may be overstating the hazard of bad grading. Law school grading is less likely to err on the pass–fail decision than the Bar Examination in California, because the mean examination in California is close to the dividing line between passing or failing, so that a large number of examinations will cluster near that line.[41] The equivalent problem in law school is deciding whether an examination merits "B-" (in a school where that is the mean) or a "C+" or "B." Although most instructors would agree that this is a difficult problem, with a number of examinations, errors are likely to cancel out

---

37. GREGORY S. MUNRO, OUTCOMES ASSESSMENT FOR LAW SCHOOLS 108–09 (2000) (citations omitted).

38. MUNRO, *supra* note 37, at 108 (citing 1 MICHAEL JOSEPHSON, LEARNING & EVALUATION IN LAW SCHOOL 19 (1984)).

39. The author has participated in two calibrations sessions for the California Bar Exam, which involve reviewing the drafting history of the particular performance exam administered and participating in a tentative, preliminary grading session. *See also* Edward C. Stark, *Dispelling Myths About the California Bar Exam*, L.A. DAILY J., May 4, 2000, at 6 (discussing experience of one member of the California Committee of Bar Examiners).

40. Compare this with the problem of ineffectively conveyed meanings as an obstacle to self-assessment. *See infra* note 89 and accompanying text.

41. *See* State Bar of California, *July 1999 Bar Examination Statistics*, *at http://www.calbar.org/shared/2adms799.htm* (last visited Sept. 21, 2000) (reporting pass rate of 50.9%); State Bar of California, *State Bar Announces Results for February 2000 Bar Exam* (May 30, 2000), *at* http://www.calbar.org/2rel/nw2000/newsrelemay30.html (reporting pass rate of 40.0%).

for most students.[42] This is not an entirely satisfying solution, but it is less unsatisfying than the conclusion that large numbers of law school examinations that are judged to be passing could as easily be judged to be failing, and vice versa.

Paul Wangerin has found similar difficulties in analyzing tests given by his colleagues.[43] Wangerin examined the correlation between objective and essay questions on exams and found it strikingly low, generally much lower than would be considered tolerable on standardized tests using multiple-choice questions.[44]

Many scholars would say that Professor Wangerin's analysis overstates the difficulty, because he applies the standards common to norm-based tests, rather than criterion-based tests.[45] Norm-based tests generally purport to be measuring one thing, so there should be a high correlation between different pairs of questions.[46] Criterion-based tests may examine different aspects of a subject, and knowledge as to one part of a subject does not logically imply knowledge as to another part of the subject.[47] Hence, with criterion-based tests, lower intraexamination correlation is acceptable.[48]

Indeed, high-quality teaching may result in exam statistics that seem to indicate a low-quality exam. One instructor in instructional methodology, at the beginning of his teaching career, obtained reliability coefficients of 0.85 for his exams, a remarkably high figure.[49] Using the same exams while the course became "more effective,"[50] the reliability coefficients successively dropped from 0.85 to 0.60 to 0.40 to 0.25.[51] These reduced coefficients resulted from more effective instruction, so that "almost everyone started to do well," thereby reducing the range of performances.[52] Where students are generally capable of the work, and instructional effort is devoted to addressing weakness in student

42. Paul T. Wangerin, Grade Conferences from Hell: Measurement Error in Law School Grading, 34–35 (June 14–15, 1994) (unpublished manuscript from the summer conference at the Institute for Law School Teaching, on file with author).
43. *Id.* at 27–28.
44. *See id.*
45. *See id.*
46. *See* SMITH & RAGAN, *supra* note 8, at 98.
47. *See id.*
48. *Id.*
49. W. JAMES POPHAM, EDUCATIONAL EVALUATION 129 (1975).
50. *Id.*
51. *Id.*
52. *Id.*

performance on prior exams, students may not know one piece of knowledge any more than any other, so correlations may approach zero.[53]

Although Wangerin's analysis may be overcritical of law school assessment practices, it may be undercritical of law school teaching practices. If students perform substantially better on one part of a subject than another, there is reason to suspect that the teaching needs improvement. Even if one topic is intrinsically more difficult than another, this should be reflected in the allocation of teaching resources.

D. Limited Ability to Repeat Testing

Professors relatively rarely repeat essay exams. This seems a wise approach. Essay exams are memorable, making it relatively easy for students to remember the subject of the exams and pass it on to other students. Moreover, because essay exams can represent only a portion of the subject on which testing is possible,[54] the benefit to those favored over other students is high.

Unfortunately, the limited ability to repeat essay exams means that it is difficult to use essay exams to evaluate the quality of instruction. This means that less reliable methods for assessing the quality of instruction must be used.

E. Conclusion on Essay Exams

Although essay exams have advantages, they have many disadvantages. They are time-consuming to grade and are unreliable. The disadvantages related to unreliability are concealed by the forced normalization of grades.[55] Grade normalization ensures that students have the same chance to get good grades regardless of who teaches their courses,[56] but grade normalization also makes it impossible to determine whether students are learning or whether teaching is effective.[57] Thus, grades demonstrate neither what knowledge students have acquired nor the strength or weakness in teaching.

---

53. Popham even reports negative reliability coefficients. *Id.* It is unlikely, however, that law school exams that do more than test recall of legal knowledge should approach zero because substantial skill components would be common to all testing. So long as some students remain better than others at these skills, there will be a positive correlation between performance on question items and on the test as a whole.

54. *See* JACOBS & CHASE, *supra* note 5, at 109.

55. *See* Downs & Levit, *supra* note 1, at 831.

56. *Id.*

57. MUNRO, *supra* note 37, at 33.

IV. NONINSTRUCTOR EVALUATION

A. The Use of Noninstructor Evaluation

The norm for law school examinations is instructor-based assessment. However, this is not an invariable rule in other areas of higher education. On most college campuses, instruction is delivered by the professor and other media, such as books, while formal and informal assessment is delivered by graduate students to smaller sections. Formal assessments are often based on problem sets, papers, midterm examinations, and final examinations, which are graded by graduate students. Informal noninstructor assessment often happens through feedback in small sections, which informs students about the quality of their oral comments.

The reliance of many respected colleges on noninstructor-based assessment should cause law schools to reexamine their practices. In colleges, noninstructor evaluation by graduate students who do not prepare the teaching materials is sometimes the sole method of assessment. Law school students and professors are unlikely to accept having an entire grade depend on an evaluation by someone other than the instructor. Thus, alternative assessment in law schools will likely be used to supplement instructor assessment.

Although law schools lack a pool of graduate students to provide formal, noninstructor assessment, other alternatives exist. These alternatives allow for more frequent assessment. In addition, they broaden students' perspective from believing that the professor's answer is the only right answer to believing that there are many ways to approach a problem.

With noninstructor grading, it becomes all the more important for the instructor to articulate goals and standards for the graders to apply.[58] Even though articulated goals and standards provide important instructional benefits for students,[59] many instructors do not develop them for their students. Thus, developing goals and standards for noninstructors will add to the instructors' burden, although the burden will have an independent benefit of improvement in student learning and may reduce the overall burden by allowing others to evaluate students.

Where the goal is to have students learn and be able to apply legal

58. *See* MUNRO, *supra* note 37, at 239–40.
59. *See* Hess, *supra* note 2, at 944; MUNRO, *supra* note 37, at 239–41.

rules, providing standards is easy. In each unit of material, instructors can say what students should be able to do after the unit is completed and provide sample problems with answers. Those students who cannot demonstrate to themselves the knowledge or skills then know to seek help. This alternative does not require any specific assessment method, so it may be easier for students to accept than frequent exams.

Providing standards in outcome can be accomplished by providing multiple-choice exams as teaching tools. Students receive a multiple-choice exam, in which they must select an option and explain the reasoning for their choice. The forced answer means that they have to commit to a definite choice. This makes it difficult for them erroneously to think in their self-assessment that they have obtained the right answer, but simply used different language than the instructor used. Once the option they chose has been identified as incorrect, the discussion of their thinking process can begin.

Other ways to provide definite self-assessment standards include making partial answers to questions available and making answers to part of a set of questions available, so that students have an exemplar of a reasoning process. With some problems it will not be possible to specify a uniquely correct answer. For example, in simulations testing skills in negotiation, drafting, or trying cases, the best course of action and the best practical outcome depend on the actions of the opposing party in the simulation. For such problems, the instructor can provide her procedures for addressing a problem, question, or self-assessment issue.

An example of providing a process in early law school education is giving the students a list of legal theories to consider and asking them whether, in writing their exam answers, they considered each of those legal theories and their constituent elements. Assessment based on the process a learner goes through in deciding how to act becomes even more important in more advanced courses, such as negotiations, drafting, trial practice, and clinical courses, because there may not be a clear connection between the students' decisions and the result the student obtains. For example, a student could make an excellent argument that was doomed to failure because of unfavorable precedent. Similarly, a student might not achieve the best imaginable result in a negotiation because the other side was too well or too poorly prepared.

Where a student's action is not clearly connected to an outcome, the difficulties of bias are exacerbated, and clear procedures for assessment are especially needed.[60] In such situations, students will often be assessed on whether they considered the factors in the same way as an

60. *See infra* Part IV.C.2.b.

experienced lawyer would. For example, in deciding where to file a case, a lawyer would consider the judges who might hear the case, the size of verdicts rendered in similar cases in the relevant jurisdictions, the time to trial in the various potential jurisdictions, the possibilities for removal and transfer, and so forth. No given outcome from balancing these considerations may be an unambiguously right answer, and a lawyer or student may lose a case despite doing a good job. Conversely, the failure to consider relevant factors can be an unambiguous mistake even if a good result occurs. The test of the students' ability is her thoughtfulness in considering the relevant issues.

A defined assessment process can add the outcome and any external inputs to a task process. For example, if the task was client counseling or a negotiation, the assessment process can incorporate the outcome by asking how people felt about the result achieved.[61] Unexpected questions from a judge, unexpected concerns from a client, and unexpected arguments from an opposing party are all external inputs that can be considered in an assessment process that, because of their unexpectedness, cannot be incorporated into the task.

B. Outside Assessment

One possibility is outside assessment by judges, practicing lawyers, and members of the community.[62] Outside assessment provides a variety of perspectives, helping to counteract the belief that there is a single right way—the instructor's way—to practice a skill.[63] Greg Munro reports that outside assessment provides increased support for a school in the community and greater credibility for the school's programs among students.[64]

Although outside assessment is an important resource, it is not fundamentally different from instructor assessment. In both cases, the assessor is presumably experienced and impartial. However, some characteristics associated with outside assessment may require more care in assessment practices. The University of Montana School of Law,

61. Richard K. Neumann, Jr., *A Preliminary Inquiry into the Art of Critique*, 40 HASTINGS L.J. 725, 765 (1989).
62. MUNRO, *supra* note 37, at 125.
63. Ralph Cagle, *Critiques of Students' Lawyering Skills*, *in* TECHNIQUES FOR TEACHING Law 310-12 (Gerald F. Hess & Steven Friedland eds., 1999); Ralph M. Cagle, *Critique Is Critical in Teaching Lawyering Skills*, L. TCHR., Fall 1995, at 10–11.
64. MUNRO, *supra* note 37, at 125.

which uses outside assessment, has a list of guidelines to ensure that outside assessment is used effectively.[65]

Second, outside assessment may more often require multiple assessors. If course grades are based on outside assessments, there may be a need to coordinate their grades.[66] This can be done by providing exemplars of performance at a certain standard[67] or, if each assessor has a sufficiently large group, by assuming that the average quality in each group is the same.[68]

C. Student-Based Assessment

1. Benefits of Student-Based Assessment

Student-based assessment offers important advantages to the learning process that do not occur with instructor-based assessment. Lawyering involves complex strategies, and law school cannot possibly anticipate the legal issues or even the skills that students will need in their decades-long careers. The ability to self-assess one's work is, therefore, critical to lifelong learning in practice.[69] Thus, some of the most valuable knowledge we can give students is how to monitor and learn from their responses to novel situations.

Currently, much student assessment in law school takes place informally and spontaneously in classes.

> Because most law students are formally evaluated only at the end of each semester, students are prone to seek out other opportunities to assess their learning. In effect, every classroom exchange becomes an opportunity for self-assessment. Aware, or simply imagining, that she is being evaluated (by the

65. *Id.* at 239–41.

66. *See supra* note 38 and accompanying text. Observations of the author and those of others on assessment suggest that different assessors, even though highly competent, will assign different numerical scores to the same performance, even though they rank the performances similarly. *See* Max Young, *The Multiple-Choice Essay*, *in* TECHNIQUES FOR TEACHING LAW 298, 299 (Gerald F. Hess & Steven Friedland eds., 1999). This was so even with score sheets comparable in detail to those described in Munro's book. MUNRO, *supra* note 37, at 209–17. This means that numerical scores provided by different judges do not accurately compare groups of people.

67. MUNRO, *supra* note 37, at 114 (discussing having assessors grade a single student performance).

68. Daniel Keating, *Ten Myths About Law School Grading*, 76 WASH. U.L.Q. 171, 188 (1998) (arguing for the use of mandatory means as a way of dealing with the lack of information about differences in performance between different sections).

69. MUNRO, *supra* note 37, at 124; *see also* DONALD A. SCHÖN, EDUCATING THE REFLECTIVE PRACTITIONER: TOWARD A NEW DESIGN FOR TEACHING AND LEARNING IN THE PROFESSIONS 317–25 (1987) [hereinafter SCHÖN, EDUCATING THE REFLECTIVE PRACTIONER] (discussing self-assessment as part of the process of revising one's knowledge); *see also* DONALD A. SCHÖN, THE REFLECTIVE PRACTITIONER: HOW PROFESSIONALS THINK IN ACTION (1983) [hereinafter SCHÖN, HOW PROFESSIONALS THINK].

> professor, her classmates, and herself), the student naturally wants to use each interchange to demonstrate knowledge and understanding.[70]

So, using student assessment is not new to law school education. What is underdeveloped is classroom work that builds student assessment into the process and finds appropriate ways to let peer-assessment be conveyed, making self- and peer-assessment as effective as possible.

Thus far, the most progress in integrating self-assessment has taken place in nonclassroom courses. Writers have described the value of self-assessment in externships and placement programs.[71] Such writers have observed that the widespread use of journal writing in undergraduate education and the claim of its proponents that journal writing promotes the development of independent thinking and writing skills create a basis for using similar principles in law school.[72]

> Many students easily fall into patterns of self-diagnosis, moving from the concrete experience, to the initial reflection, to reacting to the reflection, to acquiring confidence in their capacity to observe critically. My favorite example of this type of journal entry is one in which the student rereads an earlier entry and then continues or revises an earlier view independently, without even the benefit of an intervening reaction from me. This mode of expression is almost epistolary ("You remember how I described my frustration about my research project—I just couldn't get a handle on it and no one was around to discuss it. Well, this week, it's fine. I had a long talk with the law clerk and she straightened things out.") Even though the journal entry is addressed to me, the writer is autonomously interacting with her experiences, and solving her own problem.[73]

Other writers, focusing on the professions, believe that self-reflection is a critical component of successful professional work,[74] so that educating people in self-reflection should be a component of education for the professions.[75] In colleges, some professors use journaling to help

---

70. Peggy Cooper Davis & Elizabeth Ehrenfest Steinglass, *A Dialogue About Socratic Teaching*, 23 N.Y.U. REV. L. & SOC. CHANGE 249, 272 (1997).

71. Stacy Caplow, *From Courtroom to Classroom: Creating an Academic Component to Enhance the Skills and Values Learned in a Student Judicial Clerkship Clinic*, 75 NEB. L. REV. 872, 896 (1996) (describing student journals in a judicial clerkship as a critical part of self-assessment).

72. *Id.* at 899.

73. *Id.* Interestingly, this entry shows no ability to generalize from a past experience to learn new approaches.

74. SCHÖN, HOW PROFESSIONALS THINK, *supra* note 69, at 282.

75. SCHÖN, EDUCATING THE REFLECTIVE PRACTIONER, supra note 69, at 32–33 (describing the importance of reflection-in-action in the professions); *id.* at 39 (advocating instruction in reflection-in action).

students define their own course goals.[76] In college, peer review of journals is apparently quite common.[77]

Although self-assessment and peer-assessment have been most prominent in clinical courses, they also have a role in ordinary classroom courses. Without student self-monitoring, students will know when they are not learning only after receiving instructor-based assessment. This is unduly burdensome to the instructor, so structuring curricula to enable good self-assessment offers important benefits for teacher and student.

2. The Obstacles to Student Assessment

The obstacles to good peer or self-assessment are formidable. They need to be carefully considered and minimized. First, those learning a subject area may not be very good at assessing how well they are doing in the subject, which, by assumption, they have not mastered. Second, there are specific problems with self-assessment that peer-assessment can mitigate. Third, there are specific problems with peer-assessment that self-assessment can mitigate. These obstacles will be considered in order.

a. Unreliability of Student Assessment

The chief difficulty of assessment in the learning process is that the same deficiencies that make people poor performers often make them poor judges.[78] A study of Cornell undergraduates led two researchers to conclude that "incompetent" people are likely to overrate their performance and give their performance higher ratings than competent people give their own performance.[79] Thus, the lowest performing subjects on tests of logic and English grammar were most likely to overassess their own performance.[80] Asked to evaluate their performance on the test of logical reasoning, subjects in the bottom quartile scored only in the 12th percentile, but they believed that they

---

76. JACOBS & CHASE, *supra* note 5, at 132.

77. *Id.* at 131.

78. This is so where the skills needed for assessment are the same as the skills needed for performance. This will generally be true for cognitive skills. For psychomotor skills, assessment skills will often be different. Thus, one can assess that Michael Jordan has a great dunk shot without being able to dunk a basketball himself and can assess that his golf drive went into the woods without being able to drive a golf ball effectively. Perhaps one reason athletics gives pleasure to so many is that it allows one to critique and compare performances which he is not capable of performing.

79. Justin Kruger & David Dunning, *Unskilled and Unaware of It: How Difficulties in Recognizing One's Own Incompetence Lead to Inflated Self-Assessments*, 77 J. PERSONALITY & SOC. PSYCHOL. 1121, 1121 (1999).

80. *Id.* at 1125–26.

had scored in the 62nd percentile on the test and that their overall skill at logical reasoning was at the 68th percentile.[81] Similarly, subjects in the bottom quartile scored at the 10th percentile on the grammar test, but ranked themselves at the 67th percentile in the ability to identify grammatically correct standard English, and estimated their test scores to be at the 61st percentile.[82]

Exposure to other people's performance eliminated high-performers' underestimation of their own performance, but did not alter low-performers' overestimation of their own performance.[83] In fact, poor performers even increased their self-assessment as a result of exposure to others' good performance, although not to a statistically significant degree.[84]

At least anecdotally, such misestimation is linked to poor performance on exams. According to an interview with one of the authors of the Cornell study, the impetus for the article was the author's experience that "college students . . . after doing badly on a test . . . spend hours in his office, explaining why the answers he suggests for the test questions are wrong."[85]

It should be observed that the authors use incompetence as a relative term.[86] Because Cornell undergraduates, the subjects of the study,[87] are a relatively qualified group, there is reason to fear that the study's conclusions apply to law students, even though they are generally more highly qualified than typical college students.

---

81. *Id.* at 1125.

82. *Id.* at 1126. Interestingly, in all quartiles in these two sets of tests, students believed that their tests scores underestimated their actual ability. *See id.* at 1125 fig. 2, 1126 fig. 3. Kruger and Dunning did not report on this discrepancy, but it may provide an additional reason why poor test results do not cause individuals to reassess their performance: they believe the test is not representative of their actual abilities.

83. *Id.* at 1127.

84. *Id.*

85. *See* Erica Goode, *Among the Inept, Researchers Discover, Ignorance Is Bliss*, N.Y. TIMES, Jan. 18, 2000, at F7. In the author's experience, the same result is common for poor performers at middle-tier law schools, and for a few of the poor performers even at the very top law schools.

86. Kruger & Dunning, *supra* note 79, at 1122 n.1.

87. *Id.* at 1124 (describing selection of students for logical reasoning from introductory psychology class), 1125 (describing selection of students who received extra credit in an unspecified course).

b. Bias in Self-Assessment

Assessment of one's own work is difficult because most of us have a bias in favor of our own work, even when no grade turns on the assessment.[88] Even when work is ungraded, many students want to defend their work instead of learning from its limitations.

Even without a desire to favor one's own work, people are likely to confuse what they said with what they meant to say. This is so even for highly experienced lawyers.[89]

The difficulties of accurate assessment are exacerbated by the nature of lawyer's tasks. Many chance elements affect an outcome. For example, a lawyer may lose a case despite having done the best she could have done with a poor argument and a client who stubbornly refused to settle. A lawyer may have done enough to win a case, but for the random assignment of the case to the one judge on the court who did not like her sort of client.

The role of chance often makes it impossible to use outcome-based assessment, which relies on the comparison of the actual result with a desired outcome.[90] Even if chance does not make it impossible to say whether the lawyer (or law student) pursued the best course, it makes it easy for someone who does not want to admit to less-than-perfect work to deny his or her contribution to a less-than-perfect outcome.

D. Overcoming the Obstacles

1. Peer-Assessment

Peer-assessment reduces several disadvantages of self-assessment. The most obvious benefit of peer-assessment is that it diminishes bias.[91] Even peer-assessment may be biased, because students may wish to assess their peers leniently out of sympathy for their fellow students or

---

88. *Id.* at 1121.

89. *Cf.* Hilder v. Dexter, A.C. 474, 477 (H.L. 1902) (Lord Halsbury).

> My Lords, I have more than once had occasion to say that in construing a statute I believe the worst person to construe it is the person who is responsible for its drafting. He is very much disposed to confuse what he intended to do with the effect of the language which in fact has been employed.

Id. LON L. FULLER, ANATOMY OF THE LAW 18 (1968) ("The issue ought to be not what the legislature meant to say, but what it succeeded in saying. . . . [T]his is a question that can be tried more objectively by a court than by those who had a hand in drafting the statute.").

90. Mary-Lynne Fisher & Arnold I. Siegel, *Evaluating Negotiation Behavior and Results: Can We Identify What We Say We Know?*, 36 CATH. U. L. REV. 395, 395–96 (1987).

91. Studies show that peer-assessment has a greater agreement with instructor assessment than does self-assessment. JACOBS & CHASE, *supra* note 5, at 212.

with the hope of future lenient assessment in return.[92] Conversely, peers may grade harshly to obtain a competitive advantage in courses graded on a curve[93] or to increase their own sense of ability by deprecating their peers.[94]

A second advantage of peer-assessment is that the peer assessor does not know what the person being assessed was trying to say or do.[95] This contributes to a more impartial review. Also, a student reviewer who identified deficiencies in another student's work may be better able to see them in her own subsequent work.

Peer-assessment requires little or no extra work by professors beyond defining standards to be used in grading. Formal training in interpersonal skills may not be essential. Where graded assignments require students to work in groups, group members have an incentive to convey feedback tactfully, to maintain harmonious relations with their fellow group members, and to maximize their chances of being adopted and thereby achieve a better grade.[96]

Even in ungraded work, the same incentive to provide effective criticism exists if success on the exercise has significance to the students. For example, in a Civil Procedure exercise in which the students seek to discover features of the final exam,[97] students work together and share results with other groups. As a result, they have an incentive to evaluate the strategies of both groups correctly and to convey feedback tactfully.[98] Throughout the semester following such an exercise, the students display considerable professionalism both in the substance of their comments and in how they are conveyed.[99]

For better or worse, students are more comfortable being assessed by

92. *Id.* at 212 (mentioning professors' belief that peer-assessment is too lenient).
93. Most law schools restrict in some fashion the professor's ability to award unusually low or high grades. Downs & Levit, *supra* note 1, at 820.
94. The latter motivation is not unknown among professors.
95. *See supra* note 89 and accompanying text (discussing the difficulty in self-assessing successful communication).
96. *See* Greg Sergienko, *Procedure Students 'Discover' Exams*, L. TCHR., Spring 1997, at 10 [hereinafter Sergienko, *Procedure Students*]; Greg Sergienko, *Teaching Discovery Through Small-Group Discovery About the Final Exam*, *in* TECHNIQUES FOR TEACHING LAW 146, 146 (Gerald F. Hess & Steven Friedland eds., 1999) [hereinafter Sergienko, *Teaching Discovery*].
97. *See infra* Part IV.C.2.
98. Sergienko, *Procedure Students*, *supra* note 96, at 10; Sergienko, *Teaching Discovery*, *supra* note 96, at 292–93.
99. *See also* MUNRO, *supra* note 37, at 124 (discussing peer-assessment in small group work).

their professor than their peers.[100] One positive effect of this is that the thought of prospective peer review can cause students to do better work in the first place. An unfortunate effect is that the thought of another student reading one's work may deter some communications. For example, an extern might be more reluctant to confess a mistake in a journal read by a peer than in a journal read only by a professor. Adding peer review might deter self-assessment. In addition, peer-assessment is limited by the need for people to have good interpersonal skills to convey peer-assessment successfully.[101]

2. *Providing Incentives for Accurate Self-Assessment*

Bias can be minimized by creating incentives for students to provide honest self-assessment. One incentive is made apparent when the instructor explains to students the role of self-assessment in their future careers and the lifelong benefits of doing it.[102]

Another incentive for accurate self-assessment is for the instructor to grade the student's self-assessment separately from the assignment. This is frequently done in clinic courses and negotiations, in which a self-analysis follows the performance, and for which separate points are available.[103] In large classroom courses, grading of frequent self-assessment is likely to be as impossible as frequent grading of essay exams. However, occasional random review of self-assessment exercises provides an incentive for students to evaluate themselves accurately.

Another way of promoting accurate self-assessment is to make students want to succeed on the assignment they are to assess. Consider, for example, a final exam discovery exercise where students use interrogatories, deposition questions, and requests for production of documents to find out about the final exam. The desire to learn about

---

100. When instructed to post their papers on the class's electronic bulletin board, students in the author's Environmental Law class reacted by saying that the posting would make them work harder. This peer review was conducted without any comments being conveyed to the students. If negative reactions become serious, a teacher could point out that most lawyer's work is public and that a lawyer's peers can and do judge a lawyer's competence on the basis of court memoranda and negotiating positions.

101. *See* MUNRO, *supra* note 37, at 241 (making interpersonal skills a consideration in the selection of an outside assessor).

102. Explaining that the relevance of instruction to future goals should accompany every lesson because it motivates learning. SMITH & RAGAN, *supra* note 8, at 261.

103. *See* Don Peters, *Forever Jung: Psychological Type Theory, the Myers-Briggs Type Indicator and Learning Negotiation*, 42 DRAKE L. REV 1, 5–6 & n.11 (1993) (discussing self-analysis of papers in a negotiation class); Laura Rovner, PILF I Midterm Self-Evaluation (Oct. 2000) (unpublished manuscript, on file with author) (describing self-evaluation process in a clinic course).

the exam provides an incentive to learn the rules and use them effectively.[104] Thus, they have a strong incentive to accurately self-assess their tentative approaches to the problem.

Paul Wangerin formerly used students to deliver peer grades that were incorporated into course scores.[105] There is certainly support for some of this process, such as third-party grading, which may have advantages over having the exam writer grade her own exams.[106] In addition, as Wangerin observed, student grading provides powerful feedback on the success of student work.[107]

However, other aspects of using student grading receive less support.[108] The work on incompetence in assessment, although not testing a situation in which detailed model answers are provided, raises questions about the ability of students to grade exam answers that do not conform well to a model answer. Wangerin has stopped using student assessment because of the controversy it engendered.[109]

V. EXPEDITED INSTRUCTOR ASSESSMENT WITH MULTIPLE-CHOICE EXAMINATIONS

A. The Advantages and Disadvantages of Multiple-Choice Questions

Multiple-choice questions can be used as a vehicle for either formal or informal assessment. For informal assessment, multiple-choice

104. *See* Sergienko, *Procedure Students*, *supra* note 96, at 10; Sergienko, *Teaching Discovery*, *supra* note 96, at 146.

105. Wangerin, *supra* note 1, at 65–66 (outlining Wangerin's process).

106. *See supra* notes 39–40 and accompanying text (describing the California Bar Exam grading process and the advantages it provides for grading examinations with concealed or inadequately conveyed meanings). Wangerin, however, uses his own outline as a guide for student grading of essay exams and regrades all such exams. Wangerin, *supra* note 1, at 68. This alternative reduces the risk of subject-matter incompetence on the part of students, but does not address the problem of hidden meaning.

107. *Id.* Based on the author's observations of peer editing in a legal writing course, peer editing strongly improves writing, in part because articulating criticism of other papers made the students conscious of similar deficiencies in their own work in a way that reviewing their own work did not. *Cf.* Kathleen Magone, *Peer Editing*, *in* TECHNIQUES FOR TEACHING LAW 245 (Gerald F. Hess & Steven Friedland eds., 1999) (noting the many benefits of peer editing). However, work on student competence suggests that providing feedback may be less successful in promoting learning for the feedback provider than some would like to believe. *See supra* Part IV.C.2.

108. Wangerin cites "anecdotal" evidence in favor of the accuracy of student grading. Wangerin, *supra* note 1, at 68.

109. *Id.* at 65 & n.21.

questions can be used in class without requiring students to turn in answers. Students' answers provide vehicles for discussing the rules. The focus provided by the possible answers can restrain more wide-ranging discussions and supply objective feedback for students so that they can better assess their work.[110] In-class discussions also allow the instructor to detect problems with questions and fix them before using them on an examination with another class.

The advantages of using multiple-choice questions for formal assessment include broader coverage than essay or performance examinations[111] and a reduction in grading burdens.[112] The reduced grading burdens make possible quick, frequent, and low-stake examinations during the semester. These encourage students to keep up with work and provide them with frequent feedback on their progress. On final examinations, speed of feedback is not usually as important, but the ability of multiple-choice questions to provide broad coverage while reducing grading burdens allows professors to use more complex performance questions for other portions of a final exam.

Because it is easier to calibrate and reuse multiple-choice questions than essay or performance questions, multiple-choice questions can be used to evaluate different teaching methods. Because they are used easily on midterms, they even allow middle-of-the-course correction on points with which students have experienced difficulty.[113]

An additional advantage of objective tests is that students are more likely to accept the results and work to improve performance.[114] Students sometimes dismiss poor results on essay examinations as the result of the professor's arbitrary stylistic preferences.

Multiple-choice questions have limitations. They require students only to evaluate arguments, not to construct arguments on their own. Thus, although multiple-choice questions can evaluate knowledge of grammatical rules and the ability to organize an argument, they are poor in testing the ability to express oneself in writing.[115] In addition, multiple-choice questions allow a student to guess an answer. Where successful guessing occurs, the student need not have independent knowledge of the correct answer.

Even the limits of multiple-choice questions can sometimes be turned

---

110. *See supra* Part III.B.
111. *See* JACOBS & CHASE, *supra* note 5, at 51.
112. *Id.* at 51–52.
113. *See infra* Part V.E.5.
114. *See supra* Part IV.C.2.b. (discussing bias in self-assessment); *supra* Part IV.A. (discussing clearly defined criteria as a way to overcome bias in self-assessment).
115. However, essay exams also have severe limitations in this regard. *See* JACOBS & CHASE, *supra* note 5, at 109–10.

into advantages. Essay examinations require students simultaneously to demonstrate several skills. While more realistic than multiple-choice questions, essay examinations make it difficult to assess where an error is occurring. Multiple-choice questions can break that complex task down into the many elements of effective legal performance and isolate them.

B. Reducing Successful Guessing

A student's guessing a correct answer can come either from sheer luck or from clues relating the facts or the stem to the answer choices in a way not contemplated by the drafter.[116] Obviously, it is desirable to avoid questions that allow test-savvy students to score well even without having the knowledge or skills supposedly tested. Indeed, while successful guessing on exams that supposedly test general intelligence might be defended as simply an alternative way of demonstrating general intelligence, law school exams should test specific knowledge and skills. Tests that allow successful guessing unambiguously mismeasure the relevant ability.

There are three ways to reduce correct guessing: improve drafting to eliminate inadvertent clues, increase the number of distractors, and increase test length. There is a trade off between these alternatives. The more distractors, and the more successful the distractors are, the shorter the test can be.

Increasing the number of questions will reduce the benefits of guessing. Although guessing an answer is a serious risk for an individual question, even if well-drafted, the likelihood that a student will be able to guess on a series of well-drafted questions becomes infinitesimal as the number of questions grows.[117] Unfortunately, the

116. Multiple-choice questions consist of three parts. The background material is contained in an introductory section, often called the facts. The call of the question is the stem. The responses available for selecting are the answer choices. Incorrect responses are known as distractors. Steven Friedland, *Test Builder*, L. TCHR., Fall 1999, at 6–7.

117. Even on true-or-false questions, which present only two alternatives, the risk of a student's guessing ten out of ten right is less than one in one-thousand. This is because the chance of guessing one question correctly is 1/2. The chance of guessing two questions right is 1/2 x 1/2, or 1/4. The chance of guessing questions right is 1/2 multiplied by itself ten times, or 1/1024. Multiple-choice questions are less susceptible to correct guessing, because they offer more alternatives. On a four-alternative multiple-choice question, the odds of being correct on all ten questions by random guessing is less than one in one million. In general, the chance of guessing r questions correctly in a test consisting of a total of n questions, where the probability of a successful guess is p, is

additional questions will often require more time for students to answer than will fewer, but more complex, questions.

Reducing inadvertent clues can be accomplished by using true-or-false questions, because they contain no connections between the answer options and the rest of the question. However, true-or-false questions are more vulnerable to successful, random guessing than multiple-choice questions, because they contain only two possible answers. Therefore, it requires twice as many true-or-false questions than four-option multiple-choice questions to achieve an equal degree of immunity from random guessing.[118] In addition, twice as many true-or-false questions will consume more test-taking time than the number of multiple-choice questions that guarantee equivalent immunity from guessing.[119] This is because both multiple-choice questions and true-or-false questions contain base fact patterns that require reading.

Because using well-drafted multiple-choice questions instead of true-or-false questions will conserve examination time and hence maximize coverage, improving the quality of questions is worthwhile, even at a cost of increased drafting time. Poor drafting can happen in several ways.

First, some poorly drafted multiple-choice questions contain only one alternative with the correct legal standard. Students can answer these merely by recognizing the language of legal rules, even if they cannot apply them. Such questions do not test the ability to apply law to fact or to recognize underlying legal concepts. Good sets of questions will contain multiple choices that provide correct statements of the legal rules, so that students cannot rely simply on memorization. This forces students to apply the law to arrive at a correct answer.

Second, poorly drafted materials may provide clues to their own answers.[120] For example, students sometimes guess by selecting the longest answer. On poorly drafted materials, this is often the correct option because it contains many qualifications needed to make it correct.[121] Answers containing such words as "always" or "never" are

---

$\{n! \div [r!*(n-r)!]\} * p^r * (1 - p)^{n-r}$. HARRY G. COSTIS, STATISTICS FOR BUSINESS 256 (1972). Applying this formula, the risk of someone guessing nine of ten right is less than 1/100.

118. *See supra* note 117; JACOBS & CHASE, *supra* note 5, at 84.

119. JACOBS & CHASE, *supra* note 5, at 86–87. This may explain why the Multi-State Bar Examination uses multiple-choice questions. Certainly, teacher-drafted questions that echo the Multi-State Exam will give the questions more credibility and acceptability among students.

120. *See id.* at 60–62.

121. *Id.* at 60–61.

often incorrect because of the absence of qualifications.[122] On very poorly drafted materials, grammatical inconsistencies between the stem of the question and incorrect options will rule out some answers.[123]

Third, poorly drafted materials may have an insufficient number of alternative answers. Expanding the number of alternative answers will be fairly easy where a situation could lead to multiple, plausible legal rules. For example, if mental state is an issue in a legal rule, the options could reflect requirements for strict liability regardless of mental state, negligence, recklessness, and intent. The following provides an example:

1. Wendy was strolling in the park. Margaret owned some land next to the park. Wendy strolled onto the land owned by Margaret. In a suit by Margaret against Wendy, which is the most likely outcome?
   a. Wendy will be liable, because she walked on Margaret's land.
   b. Wendy will be liable if Margaret can show Wendy could not reasonably have believed that the land was park land.
   c. Wendy will be liable if Margaret can show that Wendy recklessly disregarded the risk that the land was not park land.
   d. Wendy will not be liable, because she did not intend to trespass.

Unfortunately, legal issues often have only two plausible rules, so that a question testing a single issue will allow only two alternatives. For example, because the intent requirement for battery is well known, questions turning on whether one could commit battery through negligence would be unlikely to produce wrong answers. In such a case, generating multiple-choice questions requires combining two rules in a single question or adding answers that have no connection with the legal rule at issue. An example of the former approach is question 7,[124] which asks students to assess the truth of two separate statements, leading to four different options.[125]

A risk of including multiple rules or requiring explanations is that the additional material provides clues to the correct answer. Consider the following question for an example of this defect:

2. David was enjoying himself, throwing darts at the local pub. He was mildly intoxicated, and in that state, he believed he had the ability to face away from the board and then quickly whirl around and throw the dart accurately at the board. Although the path from David to the board was

122. *Id.* at 62.
123. *Id.* at 61–62.
124. *See infra* p. 498.
125. This is functionally identical to what Jacobs and Chase describe as a "multiple true-false" question. JACOBS & CHASE, *supra* note 5, at 92–93.

clear, David knew that many people were standing close to either side of the board watching the dart throwing. When David tried to demonstrate his accuracy by turning and throwing suddenly, he threw wildly. The dart went through Paul's sleeve and, still piercing the sleeve, stuck in the wall. Paul immediately removed the dart. In an action by Paul against David, which is the most likely outcome?

a. David is liable for battery, because clothing is identified with the person.
b. David is liable for false imprisonment, because even brief confinement is sufficient for liability.
c. David is liable for battery, because he intentionally threw the dart.
d. David is not liable for battery, because he did not believe the contact was substantially certain.

The correct answer is (d) because substantial certainty is judged from the actor's point of view.[126] One distractor, choice (a), identifies an element for battery that is necessary, but not sufficient, to a successful claim. Choice (c) identifies another element that is necessary but not sufficient. This question is vulnerable to guessing because, if a battery occurred, either (a) or (c) might be correct answers since both contain necessary elements. Hence, a student could conclude the correct answer had to be a choice that did not impose liability.

That leaves (b) and (d). In a class discussion of the question after it was administered as a practice exam, answer (b) did not attract many students, perhaps because it was not a plausible case of false imprisonment. So, the answer has to be (d). Class discussion showed that this structure of the question, which eliminated (a) and (c), did not seem to tip off the students, but professors may choose to avoid questions that have as options two factually correct statements that address elements necessary but insufficient to a claim. In this case, the question might be improved by drafting new alternative (a): "David is not liable for battery, because the dart touched only Paul's clothing."

C. Evaluation of Questions

Most people who draft multiple-choice questions agree that drafting them is very difficult. Ambiguity on essay questions can be removed by including an answer that identifies the student's perspective on an ambiguity that the questions create. Ambiguity on multiple-choice questions cannot be addressed through the forced choices available.

One mechanism for finding ambiguities involves the examination results themselves. If those students who do well on the examination as a whole do poorly on a particular question, the question should be given

126. RESTATEMENT (SECOND) OF TORTS § 8A (1965).

special scrutiny and perhaps discarded.[127] The persuasiveness of the inference will depend on how strong the relation is and on the number of people taking the examination. Interestingly, giving the same examination to two different classes, which cover the same material, occasionally produces substantial differences in correlation between questions, even though the classes do not differ substantially in their overall performance. If the exam question itself seems satisfactory, then the instruction in the course should be examined because it is otherwise unlikely for students generally doing well to do unsatisfactorily on a particular question.

Those who administer multiple-choice questions often try several versions of a question before using it to evaluate students. Part of the purpose for repetition of questions on the LSAT and Multi-State Bar Examinations is to identify reliable questions and assure that scores mean the same thing from year to year.

Other ways of identifying ambiguities go beyond the pure multiple-choice question format. Class discussion of multiple-choice questions can provide a basis for identifying ambiguities and improving questions for a subsequent administration to a different class. Requiring students to explain their choice in a brief essay in the examination can also help evaluate a question and provide an opportunity to recognize ambiguity. The objective answers provide focus, but the requirement of an explanation greatly reduces guessing and provides a modest test of writing skills. Grading becomes relatively more time-consuming, yet less time-consuming than in grading purely open-ended questions.

Another alternative approach gives students the option of objecting to a question that the student believes has either no correct answer or more than one correct answer from among the choices supplied. To reduce a burden on the instructor, the instructor can allow a student to challenge only a limited number of questions, and to provide an incentive, the instructor can provide a bonus for successful challenges. This is less of a burden on a teacher than grading explanations for all questions. Unfortunately, there may be a very small number of questions as to which an alternative interpretation of an ambiguity in the question or fact pattern will produce only one correct answer, but a different one from that which the professor believes correct. In such a case, the student has no occasion to challenge a question, but if an explanation

127. *See, e.g.*, JACOBS & CHASE, *supra* note 5, at 178–90.

was required for all questions, the ambiguity would surface.

Even some questions that pass the statistical tests can be improved. If a possible argument or interpretation of the question is something that only a few students notice, their scores will be hurt, even though the question as a whole is a reliable tool. Thus, opportunities or obligations to explain answers can improve tests. As noted above, an opportunity to challenge questions will usually be sufficient. However, some questions that pass statistical muster and reasonably appear to both the taker and the test-writer as unambiguous are actually ambiguous. There are benefits to requiring explanations to all answers, especially for the first administration of a question. Paul Wangerin requires showing a bad statistical effect, as well as a persuasive explanation that the professor's answer is wrong.[128]

D. Questions Evaluating Existing Knowledge

Multiple-choice questions that assess existing knowledge of legal rules can be drafted with a variety of calls. Simple yes-or-no questions can ask students whether the elements of a certain claim or defense have been met. Such a question requires students to evaluate facts according to standards that they are supposed to know, but which are not provided in the question.

A more common possibility with multiple-choice questions is to provide alternatives that include legal rules as justifications. This is the typical sort of question on the Multi-State Bar Exam.[129] Such questions can test students' ability to identify the correct reasons for a given result. Distractors that add explanations to a decision on liability can also combat students' reliance on an intuitive assessment of whether a party should be held liable. The disadvantage is that adding explanations to the answers can provide the students with clues that help them answer that question or another one.[130]

A simple, but more complex, question supplies the students with facts and asks them to evaluate claims to see which is strongest or weakest. Conversely, a question can ask students to assess which additional fact from a menu of several would improve a claim most.

---

128. Wangerin, *supra* note 1, at 69 & n.23. The author does not require this because of the unreliability of a bad statistical effect and because an ambiguity in an answer may be perceptible to only a few people, especially if those making the mistake are high scorers. *See supra* notes 45–48 and accompanying text.

129. Example questions 1 and 2 in this Article are of that type. *See supra* pp. 489–90.

130. *See supra* p. 489 (showing an example question).

E. Question Evaluating Skills

1. Introduction

Skills evaluation and multiple-choice questions are often thought to be inconsistent.[131] Certainly, multiple-choice questions cannot easily test writing skills. Moreover, instructors consistently overestimate the extent to which their multiple-choice questions test skills.[132]

On the other hand, law professors do use multiple-choice questions to assess skills,[133] and in some respect they offer more sophisticated tools for analyzing skills than essay questions.[134] Multiple-choice questions can ask only about the facts, pinpointing students' difficulties in reading facts, or include a rule of law, pinpointing students' difficulties in applying rules. Moreover, multiple-choice questions provide fast feedback and statistical verification of the reliability of questions. The following portion of this Article explains some of the uses of multiple-choice questions to evaluate skills, starting with an analysis of the limitations of the essay and traditional multiple-choice exams.

2. The Limits of Complex Questions in Identifying Student Mistakes

Essay questions can test a student's ability to identify relevant facts, apply the law to them, and organize and write an answer. Unfortunately, the very complexity of essay questions limits their usefulness in identifying where students make mistakes. The thought process in writing an exam answer is a chain with many links[135] and when the chain breaks, it is often impossible to tell which link failed. A failure to address an issue can result from careless reading, which caused the student to miss a relevant fact; from failing to know the applicable legal rule, which caused the student to miss the significance of a fact that the student did read; or from inability to interpret the language of a memorized rule, which caused the student to miss the applicability of the rule.

---

131. JACOBS & CHASE, *supra* note 5, at 51.
132. *Id.* at 52.
133. Wangerin, *supra* note 1, at 64 & n.18.
134. *See* Greg Sergienko, *Skills Evaluation with Multiple-Choice Questions*, L. TCHR., Fall 2000, at 3.
135. The thought process in answering an exam question includes several chain-linked steps: (1) read the question, (2) recall the rule, (3) apply the rule, and (4) write an answer to the exam question.

The traditional multiple-choice exam tests knowledge of the law by asking questions about the legal rule and forcing the student to select among alternative statements of the law or by providing a fact pattern and alternative answers applying the law to fact. This seems to be the exclusive type of multiple-choice question in law schools and on the bar. Multiple-choice questions requiring reading, recall, and application are quite similar to essay questions in their complexity, but also share the defects of essay questions. Because the student can go wrong in reading, recalling, and applying a rule, a wrong answer to such a question does not reveal where the student went wrong. The lack of an explanation in an answer to a multiple-choice question exacerbates this difficulty.

3. Skills-Oriented Multiple-Choice Questions

The limitations of essay exams and traditional multiple-choice questions have led to the development of skills-oriented multiple-choice questions. These questions examine separately the ability to read facts and cases and the ability to apply an unfamiliar rule of law.

Questions testing for the ability to read facts provide reading material and ask factual questions about it. This is quite similar to many questions on the SAT and LSAT. Reading passages in law school are generally factual situations that might give rise to legal issues, but some of the questions asked pertain solely to the facts.

Questions testing for law application provide a rule of law and ask the student to apply it. The rule can be in canonical form, such as a definition, or can be from a case. The test then asks the student to apply the rule.

Such exams have important features for diagnosing students' difficulties and assessing students' skills. Using multiple-choice exams for skills testing makes possible frequent evaluations during the semester. This allows the instructor to evaluate the success of instruction and provides students with the chance to evaluate their own progress. In addition, the speed of feedback vastly increases its effectiveness. With multiple-choice tests, results are often available on the afternoon of a morning exam, or even in the second half of a one-hour class. The desire for better grades provides students with a constant incentive to develop skills, because they know they will be tested often.

Breaking down legal analysis into pieces also makes the test more effective as a teaching tool. The information about where students are going wrong is often a surprise to them. Two aspects of this are especially noteworthy. First, students tend to overestimate how

carefully they read questions.[136] Quantifying exactly how often students miss relevant facts appears far more effective in encouraging careful reading than just telling them to read carefully. Second, students who make errors are often genuinely unsure about whether those errors came from their misremembering the words of the rule or from their inability to apply those words. Breaking down the process into rule knowledge and rule application provides them with this information.

4. Examples of This Approach

The examples that follow test skills. The material on reading opinions and rules is sometimes described as questions based on a "context-dependent item set."[137] In these examples, the articles, rules, and opinions are the item set, and they are followed by questions. Use of multiple questions based on each set offsets the longer amount of time necessary to read the material. Such questions are excellent ways of measuring higher-level cognitive skills, but the material must be novel to avoid merely testing recall.[138]

Experience suggests that multiple-choice exams are harder to draft when they test higher-level skills,[139] which deters drafting such questions. Using one of the systematic descriptions of intellectual skill levels, such as Bloom's taxonomy or Gagné's categorization of intellectual skills, helps ensure that questions are drafted to test an appropriate range of skills.[140] Allowing ample time to write such questions, preferably before or while the material is taught, will lead to better questions.[141]

---

136. This problem is confirmed by a study, comparing the reading strategies of lawyers and law students, which found that lawyers spend more time ascertaining the facts than law students. Peter Dewitz, *Legal Education: A Problem of Learning from Text*, 23 N.Y.U. REV. L. & SOC. CHANGE 225, 230 (1997).

137. JACOBS & CHASE, *supra* note 5, at 68.

138. *Id.* at 68.

139. *Id.* at 52.

140. *See* BENJAMIN S. BLOOM, ET AL., TAXONOMY OF EDUCATIONAL OBJECTIVES: THE CLASSIFICATION OF EDUCATIONAL GOALS (1956); JACOBS & CHASE, *supra* note 5, at 17–20 (applying Bloom's taxonomy to exams generally); ROBERT M. GAGNÉ, THE CONDITIONS OF LEARNING AND THEORY OF INSTRUCTION (1985); Paul S. Ferber, *Bloom's Taxonomy: Teachers' Framework*, L. TCHR., Spring 1997, at 4–5 (discussing the use of Bloom's taxonomy on law school exams).

141. JACOBS & CHASE, *supra* note 5, at 52.

a. Reading Opinions

By asking students to read and apply a case, we can test their ability to read a case, extract from it a rule, determine critical facts to the application of the rule, and recognize parallel situations in which the rule could reasonably be applied. Questions 3 through 6, which provide a set of examples, refer to the following case. In this version, but not (of course) the one given students, the correct answers have been italicized.

Lamson v. American Ax & Tool Co., 177 Mass. 144, 58 N.E. 585 (1900).
Holmes, C. J.

This is an action for personal injuries caused by the fall of a hatchet from a rack in front of which it was the plaintiff's business to work at painting hatchets, and upon which the hatchets were to be placed to dry when painted. The plaintiff had been in the defendant's employment for many years.

About a year before the accident new racks had been substituted for those previously in use, and it may be assumed that they were less safe, and were not proper, but were dangerous, on account of the liability of the hatchets to fall from the pegs upon the plaintiff when the racks were jarred by the motion of machinery near by. The plaintiff complained to the superintendent that the hatchets were more likely to drop off than when the old racks were in use, and that now they might fall upon him, which they could not have done from the old racks. He was answered, in substance, that he would have to use the racks or leave. The accident which he feared happened, and he brought this suit.

The plaintiff, on his own evidence, appreciated the danger more than any one else. He perfectly understood what was likely to happen. That likelihood did not depend upon the doing of some negligent act by people in another branch of employment, but solely on the permanent conditions of the racks and their surroundings and the plaintiff's continuing to work where he did. He complained, and was notified that he could go if he would not face the chance. He stayed and took the risk. He did so none the less that the fear of losing his place was one of his motives.

Exceptions overruled.

3. Who won in this case?
   a. The defendant won in the trial court and on appeal.
   b. The plaintiff won in the trial court and on appeal.
   c. The defendant won in the trial court, and the plaintiff won on appeal.
   d. The plaintiff won in the trial court, and the defendant won on appeal.

4. Which of the following facts is most significant to the court's result?
   a. The plaintiff signed a contract assuming the risk that axes would fall off.
   b. The plaintiff knew that the axes were likely to fall off.
   c. The accident happened as a result of negligence by people in another branch of the employment.
   d. The new racks made the drying process less safe.

5. In what situation would the *Lamson* precedent be most relevant?
   a. *The plaintiff smelled alcohol on the breath of the defendant, but allowed the defendant to drive him anyway, and was injured in an accident caused by the intoxication of the driver.*
   b. The plaintiff was an employee and was injured as a result of a defect in the defendant's flooring, which was unknown to the employee and caused an ax to fall on the employee from a higher floor.
   c. The plaintiff was an employee in the painting department, and the defendant was an employee in the ax-head installing department, who injured the plaintiff by throwing an ax head at his supervisor and hitting the plaintiff by mistake.
   d. The plaintiff was a minor employed by the defendant, and was suing for loss of brain function suffered as a result of continued exposure to paint fumes in the defendant's factory.
6. Which change in the law would be most likely to change the outcome of this case?
   a. *Different rules on assumption of risk in negligence cases.*
   b. Different rules on battery and other intentional torts.
   c. Different rules on extreme and outrageous conduct (intentional infliction of emotional distress).
   d. Different rules on the damages recoverable for personal injury.

Question 3, which asks who won the case, addresses procedural knowledge and case reading. The "exceptions overruled" at the end demonstrates that the same person who lost in the trial court lost on appeal. From the opinion, it is clear that the plaintiff loses, although the opinion nowhere states that in so many words.

Question 4, which asks which facts are most significant, assesses reading skills and ability to apply a rule to facts. Its incorrect choices combine possibilities for missed facts and erroneous understanding of the court's opinion. The first option, the signed contract assuming the risk, is legally plausible but factually erroneous. Someone with a correct understanding of the law but poor reading skills would be tempted to pick that. The second option is correct. The third option, that the accident was caused by people in another department, is factually incorrect and would support an outcome different from the court's. The fourth option, that the new racks are less safe, is assumed by the court to be factually correct, but is not important for the court's assumption of risk argument.

Question 5 asks the reader to select the most analogous case. This tests the ability to transfer knowledge from one case to another. To do this, readers must understand the salient features of the plaintiff's decision to encounter a known risk and give that more weight than scenarios that have facts that are superficially similar to those of the

Lamson case. Readers with poor knowledge of legal principles are likely to focus on superficial similarities.[142]

Question 6 asks the reader to classify this case by selecting the area where changes would most likely alter the result in this case. The question is quite easy, because there is no hint of an intentional tort, emotional distress, or a need for damage rules in the case. A student who cannot correctly classify the case is likely to have weak knowledge of general tort law or extreme problems in identifying the facts.

b. Reading Rules

Questions on reading rules provide students with the rule and then ask them to apply it. A failure to succeed indicates difficulty in rule-reading skills. Often, students confuse this deficiency with a failure to memorize the rule.

> 7. Assume that the intent to accomplish a result exists if the actor desires the result or believes that the result is substantially certain to occur as a result of his acts. In which of the following situations does the intent to make contact exist?
>
> I. Calvin has thrown 100 slushballs at Susie, but because of his rotten aim, he has missed each time. He realized the odds are strongly against him, but he desperately wanted to hit Susie, so he threw another slushball, which hit Susie.
>
> II. Calvin was outside a stadium testing his new slingshot for propelling slushballs, and he aimed his slushballs to fall inside the stadium. Signs outside the stadium advertised the sell-out crowd and Calvin heard cheering, but Calvin did not read the signs and failed to realize that there was anyone inside the stadium. Calvin hit someone inside the stadium, as was substantially certain to happen.
>
> a. Neither I nor II.
> *b. I only.*
> c. II only.
> d. I *and* II.

The rule is based on the standard definition of intent in torts.[143] The correct answer is (b). Intent exists in I because Calvin hopes to hit Susie, even though he expects not to hit Susie. Intent does not exist in II, because Calvin was not substantially certain he would hit someone, although it was substantially certain in reality. Here, providing the student the rule to work through means that the student learns that it is his or her ability to apply the rule that needs improvement, not rule memorization.

It is also possible to use a single question to assess for a variety of

142. *See* SMITH & RAGAN, *supra* note 8, at 136 (discussing differences between novice and expert problem solvers).

143. RESTATEMENT (SECOND) OF TORTS § 8A (1965).

errors. The following example, substantively very much the same as the one immediately preceding, illustrates this approach. This time, the information in italics indicates the sort of error made.

> 8. Intent to accomplish a result exists if the actor desires the result or believes that the result is substantially certain to occur as a result of his acts. Intent to make a harmful or offensive contact is an element of battery.
>
> Calvin is testing his new slingshot for propelling slushballs. He is on the outside of a stadium and aims his slushballs to fall inside the stadium. Although there are signs outside the stadium advertising the sell-out crowd, Calvin neglects to read the signs, and does not think about the implications of the noise emanating from the stadium and so does not know that there are people inside the stadium. Calvin hits someone inside the stadium, which any impartial observer would have realized was substantially certain. Which of the following alternatives is correct? *In this question, the alternatives are evaluated in italics.*
>
> a. Calvin is liable for battery because he knew that it was substantially certain that he would hit people in the stadium. *An incorrect choice, because it ignores the statement in the facts that Calvin "does not know that there are people inside the stadium." The error type is one of not reading facts.*
>
> b. Calvin is liable for battery because it was substantially certain that he would hit people in the stadium. *An incorrect choice, because it disregards the requirement that the actor know that the contact was substantially certain. The error type is one of not reading rules.*
>
> c. Calvin is liable for battery because he ought to have known that it was substantially certain that he would hit people in the stadium. *An incorrect choice, because the rule requires purpose or knowledge. The error type is one of not reading rules.*
>
> d. None of the other alternatives is correct. *This is the correct choice, and by being in a none-of-the-above format, it makes the correct answer harder to guess, by requiring the specific elimination of all alternatives.*

Where none-of-the-above choices are present, research indicates that they tend to make exam items more difficult.[144] The none-of-the-above option cuts down on successful guessing by requiring the students specifically to reject the other three alternatives, instead of allowing them to compare a true alternative with false alternatives.

A limitation of this sort of question is that it does not account for the

144. JACOBS & CHASE, *supra* note 5, at 63–64. This might also be inferred from the restrictions that a none-of-the-above alternative puts on the ability to infer a correct answer from partial information. In most multiple-choice selections, one can work by the process of elimination, so that identifying the truth or falsity of all but one of the alternatives guarantees a correct answer. With a none-of-the-above question, the ability to identify the truth of the last alternative requires correctly identifying the truth or falsity of all of the preceding alternatives.

possibility that the student will commit both factual and legal errors. Such a student will receive feedback on only one type of error.

c. Reading Facts

Questions based on a statement of facts that are not contained in a legal opinion may also be considered. These are very much like any reading comprehension question, although tending to involve material with a legal context. Using newspaper or magazine articles as a starting point for these questions will limit the influence of the instructor's idiosyncrasies in writing and perhaps provide more readable material.[145]

An example follows, with the correct answers italicized.

Officials Condemn Bandit Cabs

Los Angeles County Supervisor Yvonne Brathwaite Burke called for more enforcement against unlicensed or "bandit" taxis. Los Angeles city officials said they intend to increase enforcement after a Blue Line train and an unlicensed cab collided, killing six.

The cab driver, Romaldo Gonzalez, had been convicted twice for drunk driving. He was driving an unlicensed cab on a suspended driver's license when he and his girlfriend picked up three brothers and a young woman at a party in Compton. Everyone in the cab died when it was crushed by the train.

Main City's officials said they had legitimately sent a cab to pick up people at the party in Compton, only to find that Gonzalez had beaten the driver there. However, Compton police spokesman criticized Main City Taxi. Main City Taxi is not licensed in Compton, the spokesman said, so it should not have taken the call.

Gates Proposed at Blue Line Intersections

The Blue Line rail accident that killed six last weekend might have been prevented if four barrier gates, instead of the standard two, had barricaded the intersection at which a train collided with a taxicab, the MTA's safety chief said.

The empty train, which was going out of service and heading back to its service yard, was traveling 55 mph when the cab went around a barrier gate. At almost all its intersections, the MTA uses two gates, which is standard throughout the United States. The gates close to stop oncoming vehicle traffic. Often, though, as happened in Saturday's collision, motorists enter the open lanes on the wrong side of the street and try to beat the train through the intersection.

"The staff feels that in this particular accident a four-quadrant grade-crossing protection device may have deterred the driver of the taxicab from entering the intersection, thus avoiding the collision," safety director Paul Lennon told the MTA board.

---

145. There are several measures of reading difficulty that can help in constructing sets of questions with equivalent difficulty. SMITH & RAGAN, *supra* note 8, at 321.

Although Lennon said that he believes the driver 100% responsible, MTA directors voted to re-assess the agency's safety measures and press for installation of quadrant gates at problem intersections, although overcoming regulatory hurdles could take months or years.

Since it opened in 1990, fifty-three people have lost their lives along the Blue Line right of way, by far the worst record of any of California's five light-rail systems and said to be one of the worst in the nation. All those who died in the accidents were either on foot or in motor vehicles hit by the trains.

While investigations have cleared the MTA in the accidents, Los Angeles County Supervisor Yvonne Brathwaite Burke, who heads the thirteen-member MTA board, said steps have to be taken to protect people from themselves. "We are going to have to do something that takes into account human error," Burke said.

Burke said that in addition to installing quadrant gates, the MTA should explore grade separations along the Blue Line tracks. Grade separation would either raise or lower tracks through heavily trafficked intersections, a costly countermeasure that transit agency planners rejected in building the system.

The Blue Line runs mostly through densely populated neighborhoods at street level, traveling over crosswalks and through intersections heavily used by cars, trucks and buses, at speeds that are among the fastest in the nation for a light-rail system. This has led critics and system administrators to conclude that accidents are inevitable.

Installation of four-quadrant gates could take months or years, because the safety measures would have to be approved by the Public Utilities Commission and then receive funding. Some MTA board members called for immediate action, such as publicizing the recent increase from $104 to $271 in the fines for ignoring train warning signals and getting legislative permission to increase fines to as much as $1,000.

Burke rejected this idea and observed that the Blue Line goes through many low-income communities, where a $271 fine already assessed would be a substantial loss to a household.

9. An element of negligence per se is a violation of a statute, regulation, or ordinance. According to the facts in these articles, which of the following violations did the driver commit at the time of the accident?
   I. Driving while drunk.
   II. Driving around the crossing gates.
   III. Driving without a valid driver's license.
   IV. Driving without a valid taxicab license.
   a. I and II
   b. II and III
   c. *II, III, and IV*
   d. I, III, and IV
   e. I, II, III, and IV

10. Vicarious liability can make an employer liable for the conduct of the employee. According to the facts in these articles, against which defendants might the element of employment exist in a vicarious liability claim?
   a. Main City Taxi, for the conduct of the cab driver in the accident.
   b. The MTA, for the conduct of the conductor or engineer of the train in the accident.
   c. Against both Main City Taxi, for the conduct of the cab driver in the accident, and against the MTA, for the conduct of the conductor or

engineer of the train in the accident.
d. Against neither Main City Taxi nor the MTA.

11. Which of the following would *least* support requiring the MTA to install four-quadrant crossing gates?
    a. Society believes that it should protect people against the consequences of their own negligent law breaking, even if that costs other people money.
    b. Installing four quadrant gates is much less expensive than grade separation.
    c. Innocent passengers can be harmed by drivers going around gates that block only the right half of the road.
    d. Installing the gates would require a long approval process.

12. If all of the following facts were true, which one would *least* support raising the fines for ignoring train warning signals?
    a. Many people who drive across the railroad tracks of the Blue Line are from neighborhoods other than the one through which the tracks run.
    b. People can be fined for a traffic offense even if they would not be killed in an accident.
    c. The neighborhoods through which the Blue Line passes are low-income communities.
    d. Raising fines does not require expensive construction.

Using these questions revealed that students often infer too much from a factual situation, and fail to separate out what they know from what they infer. For example, the students often believed that the article stated that the driver was drunk at the time of the accident, although the article stated only that the driver had been convicted of drunken driving on two prior occasions.

In addition, students often make mistakes about the call of the question. Question ten asks about whether the element of employment exists. However, a substantial minority of students addressed the issue of negligence, and chose an option that excluded liability for the MTA.

Statistical analysis of the questions shows that they have a high correlation with traditional multiple-choice questions and skills-based multiple-choice questions. However, students who did poorly on reading facts often had difficulty understanding the relevance of these questions to legal education. This suggests that these students' improvement on more conventional legal tests may encounter a major unrealized obstacle (unrealized, that is, to the students involved), poor critical reading skills.

5. Preliminary Results from Testing Skills with Multiple-Choice Exams

The following section of this Article reports the preliminary results of an examination study conducted in a Remedies course offered in Spring 2000. Six quizzes and a final exam were administered as part of the class. The exams included:

(1) A multiple-choice skills quiz,
(2) A multiple-choice quiz mixing skills questions with questions testing ability to identify and apply legal knowledge,
(3) A conventional essay quiz,
(4) A case-based essay performance quiz,
(5) A rule-based essay performance quiz,
(6) A take-home short-answer quiz on verbal knowledge, and
(7) A case- and rule-based essay performance exam final.

First, use of skills-based multiple-choice questions has already provided some benefits. Because multiple-choice questions provided fast information to me about students' strengths and weakness in a variety of skills, student weaknesses were identified and teaching was modified accordingly.

In particular, students did quite poorly on skills tests. This led to the use of more skill-based testing on subsequent quizzes, administered in essay formats. In addition, teaching methods were modified in first-year courses to include substantial formative evaluation of reading skills pertaining to rules and cases. It is too early to tell whether this will help, but students have noticed the change in emphasis and seem receptive to it.

As between different sorts of skills test, students did best on reading cases and most poorly on questions dealing with rule application and recognition of facts. A tentative explanation for these results is that the context provided by a case allows even students with mediocre reading skills to grasp more of the implications of a case. Those disparate results led to the development of substantial instruction on reading rules. It is too soon to determine how successful this instruction will be.

The ability of one test to predict the results of another is called concurrent validity.[146] The Remedies course uses a variety of test styles, and therefore provides a modest database for tentative conclusions about the efficacy of multiple-choice skills testing.

The best predictor of all scores was the final. This was to be expected because it was much longer and more comprehensive than the quizzes.[147] Clustered relatively closely together were two essay performance exams

146. HARRIET TALMADGE, STATISTICS AS A TOOL FOR EDUCATIONAL PRACTITIONERS 115 (1976).

147. The more questions on an exam, the more reliable it is, so long as nearly everyone finishes. JACOBS & CHASE, *supra* note 5, at 38. The quizzes were from twenty-five to forty minutes long; the final was three hours long.

and the skills-based multiple-choice exam. The second best overall, after the final, was a case-based performance exam. Because cases provided an explanation and analysis of the rules that were given on the final, rule-based testing may have had a slightly smaller effect on performance. Third best overall, and about one percentage point behind the case-based performance exam, was a skills-based multiple-choice test similar to the one that supplies the examples for this Article. This is especially striking because it was the first quiz, so to the extent that one would expect performance to improve over time, tests administered in the middle would better represent average performance in the course. A rule-based essay performance exam trailed about two percentage points behind that.

Three tests performed relatively poorly, the take-home quiz on factual knowledge and concepts, the mixed performance and knowledge multiple-choice quiz and a conventional essay quiz. The poorest predictor of all was the conventional essay quiz.

The take-home quiz was quite different from all the other quizzes. It was the last of the quizzes, and because the students' final grade would consider only their best five quizzes, students who had done well before the take-home quiz would have little incentive to do well on it. This may have contributed to the quiz's poor predictive value. On the other hand, the quiz was long, which increases reliability, and it did provide a comprehensive review of terms and concepts for the final exam. That may explain its higher correlation with the final than with any of the other quizzes.

The reasons for the poorer performance of the other quizzes in predicting students' overall success in the course are less clear. The quiz consisting of mixed performance and knowledge questions may have been a poor predictor because of the knowledge component. It was also shorter than the first quiz, and its internal measures of reliability were lower than for the purer performance multiple-choice-question quiz.

The failure of the essay exam is quite striking. Its inability to predict performance on a variety of performance exams suggests that skills-based multiple-choice exams are better predictors of performance on written-out skills exams than standard essay exams. Moreover, because multiple-choice skills can predict performance approximately as well as written skills exams, the role of essay exams can be reduced without compromising the quality of evaluation.

This would suggest that the act of writing out an answer is not an important factor in differentiating among students' performance. If writing out an answer were significant, one would expect the essay exams to be much better predictors than the multiple-choice exams. As a result, well-drafted multiple-choice questions seem to have significant

potential for assessing students' skills in a way that is fast and relatively easy, once the questions have been drafted.

6. Concluding Thoughts on Testing Skills with Multiple-Choice Exams

By traditional measures of the efficacy of multiple-choice exams, these skills-oriented questions are very successful. The only substantial disadvantage is that drafting these questions is even more time-consuming than drafting ordinary multiple-choice questions. However, the benefits and long-term results appear to be worth it.

VI. CONCLUSION

Conventional law school testing procedures fail to assess learning and do not provide information to students in time for the students or instructors to improve. Better testing is possible at little or no overall increase in the faculty's workload, although the workload will shift from evaluating student answers to creating and administering examinations. Such testing can promote student learning by telling students what they have not yet learned and telling instructors what instruction is effective.