All of the Above: Computerized Exam Scoring of Multiple Choice Items Helps To: (A) Show How Exam Items Worked Technically, (B) Maximize Exam Fairness, (C) Justly Assign Letter Grades, and (D) Provide Feedback on Student Learning

Lynn M. Daggett

Introduction

Many years ago in college and graduate school, I studied psychometrics, the branch of psychology that deals with standardized testing. In courses in psychological and educational testing, I ran a weekly lab in which students took standardized tests. I scored these tests, and also constructed and scored weekly multiple choice quizzes. I punched holes in manila folders, placed the hole-punched folder "scoring keys" over my students' answer sheets, and manually scored their exams. To calculate item analysis (difficulty and discrimination) statistics, I sorted answer sheets into endless piles and tediously hand calculated the pass rate for each item and each pile.

Today, we law faculty are lucky to have access to computerized scoring for multiple choice exams. This technology not only eliminates tedious hand scoring, but also provides a wealth of statistical information about how students performed on the exam and each item. This data can be helpful to law faculty. First, the data identify potential "bad" items—ones for which the pattern of student performance suggests a possible flaw. Teachers can review these potentially flawed items to determine if in fact a flaw exists; if it does, the teacher can then decide what to do about the item (perhaps, for example, to double

Lynn M. Daggett has a Ph.D. in education and is a professor at Gonzaga University School of Law.

This paper began as a handout for talks I gave at my own and several other law schools on this topic. I thank the audiences at those schools for their interest in the issue and for their specific comments and questions, which helped to shape this article. I also offer my profound (and unfortunately, posthumous) gratitude to my mentor in my "first life" in education; noted psychometrician Dr. Julian C. Stanley piqued my interest in and then generously and expertly guided my learning in psychometrics.

Journal of Legal Education, Volume 57, Number 3 (September 2007)

key it) to maximize the fairness of the exam for the students. Second, the data offer specifics about how each item worked—particularly, the extent to which each item sorted out different levels of student learning of the material. Third, the data provide guidance about when scores are statistically different enough to warrant different letter grades. Finally, the data suggest specific areas where the class, overall as well as individual students, did (or did not) master the material.

The data cannot substitute for the teacher's judgment, be it about whether an item is bad or an exam is fair, whether the students' performances on an exam truly reflect their varying levels of mastery of the material, when two scores on an exam are truly different enough to warrant different letter grades, and what both individual students and the class as a whole have (not) learned. The data provide some useful information to the teacher for making those judgments.[1]

The second part of this article briefly reviews a few core psychometric concepts: validity and reliability (two properties of all good exams), formative and summative evaluation (two different purposes of exams), and the specific norm-referenced and criterion-referenced categories of exams. The third section describes the typical sorts of data available from computerized scoring of multiple choice exams and the uses of this data for law faculty. The Appendices provide some concrete examples for the interested reader.

This article does not review the pros and cons of multiple choice formats for law school exam items. There is a wealth of existing literature on this issue.[2] In the spirit of disclosure, I believe multiple choice questions have

1. As one commentator notes, student evaluation is a part of our job that is institutionally marginalized and for which we are provided with little training or other support. See Steven Friedland, A Critical Inquiry into the Traditional Uses of Law School Evaluation, 23 Pace L. Rev. 147, 174-79 (2002) (Student evaluation is not given significant law school training or other resources, incentives, or oversight and in fact is largely performed outside of the semester and away from school.). That does not make the judgments and choices involved in student evaluation any less important.

2. For an examination of the strengths and weaknesses of multiple choice exams, see 2 Michael S. Josephson, Learning & Evaluation in Law School 318-33 (Washington, D.C., 1984). For a debate on the relative merits of multiple choice and essay exams, see Norman Redlich and Steven Friedland, Challenging Tradition: Using Objective Questions in Law School Examinations, 41 DePaul L. Rev. 143 (1991); Michael S. Jacobs, Law School Examinations and Churchillian Democracy: A Reply to Professors Redlich and Friedland, 41 DePaul L. Rev. 159 (1991); Norman Redlich and Steve Friedland, A Reply to Professor Jacobs: Right Answer, Wrong Question, 41 DePaul L. Rev. 183 (1991).

   For commentary espousing the use of multiple choice questions, see Howard J. Gensler, Valid Objective Test Construction, 60 St. John's L. Rev. 288 (1986) (multiple choice format offers opportunity to test material broadly, is good at separating out all levels of student performance, and is easy to score); Linda R. Crane, Grading Law School Examinations: Making a Case for Objective Exams to Cure What Ails "Objectified" Exams, 34 New. Eng. L. Rev. 785 (2000).

   For commentary espousing the use of all-essay exams, see, e.g., Kenney F. Hegland, On Essay Exams, 56 J. Legal Educ. 140 (2006). For musings on multiple choice questions, manual versus computer scoring, and grade curves, see Marcella David, A Funny Thing

both real strengths and significant limitations as a means of evaluating law students. Consequently, in large classes I regularly use a multiple choice exam component (typically one fourth to one third of a three hour exam and less than half of the grade for a course), but I am uncomfortable assigning a course grade based solely or even primarily on performance on multiple choice items.[3] I offer some of my reasons for taking this approach throughout the article.

A Primer on Some Core Psychometric Concepts

Exam Validity[4]

The most important psychometric property of any exam is that it be "valid," which means that the exam measures whatever it is supposed to measure.[5] An exam that is not valid is not worth much and is not a good basis for assigning

---

Happened on the Way to the Multiple-Choice Exam: Or, The Schoolroom Lessons from Bush v. Gore, 51 J. Legal Educ. 1 (2001).

One concern about essay exams is the risk of subjective and/or arbitrary grading. See, e.g., Gregory S. Munro, Outcomes Assessment for Law Schools 108 (Spokane, Wash., 2000) (on California bar exam, the same person grading the same question twice resulted in the same (pass or fail) result only 75 percent of the time, and when two different examiners reviewed the same answer, they agreed on whether it passed or failed only 67 percent of the time).

Possible gender and/or racial bias, discussed *infra* note 7, is identified as a concern with multiple choice item format.

3. See Friedland, A Critical Inquiry, *supra* note 1, at 195-96 (proposing that ideal law school evaluation involves "multiple evaluations with varying item types" so that student evaluation will not depend on the strengths and weaknesses with regard to various types of exam questions).

4. There are some excellent treatises on psychometrics for the interested reader. See, e.g., Peter W. Airasian, Assessment in the Classroom: A Concise Approach (2d ed., Boston, 2000); Anne Anastasi, Psychological Testing (6th ed., New York, 1988); Lee J. Cronbach, Essentials of Psychological Testing (5th ed., New York, 1990); Robert L. Ebel and David A. Frisbie, Essentials of Educational Measurement (5th ed., Englewood Cliffs, N.J., 1991); Norman Edward Gronlund and Robert L. Linn, Measurement and Evaluation in Teaching (6th ed., New York, 1990); Kenneth D. Hopkins and Julian C. Stanley, Educational and Psychological Measurement and Evaluation (6th ed., Englewood Cliffs, N.J., 1981) (in the interests of full disclosure, I worked on this book as a graduate student); Robert M. Kaplan and Dennis P. Saccuzzo, Psychological Testing: Principles, Applications, and Issues (3d ed., Florence, Ky., 1993); John Salvia and James E. Ysseldyke, Assessment (6th ed., Boston, 1995); William Wiersma and Stephen G. Jurs, Educational Measurement and Testing (2d ed., Boston, 1990).

The most extensive review of psychometric issues specifically for law professors is Josephson, Learning & Evaluation in Law School, *supra* note 2.

For more extended discussions of exam validity, see Anastasi, Psychological Testing, *supra*, at 139-201; Hopkins and Stanley, Educational and Psychological Measurement, *supra*, at 76-112; Josephson, Learning & Evaluation in Law School, *supra* note 2, at 6-15.

5. For other more detailed discussions of exam validity, see Friedland, A Critical Inquiry, *supra* note 1, at 157-60.

letter grades, determining which students should be academically dismissed, and the like.

Types of Exam Validity

Exam validity depends on the purpose(s) of the exam. An IQ test is designed to measure the "construct" of general intelligence; if it truly does measure intelligence it is said to have *construct* validity. The LSAT is designed to measure raw legal reasoning ability to predict success in law school; if it does indeed measure legal reasoning ability it would have construct validity. Moreover, if LSAT scores correlate with an external criterion, law school grades, the LSAT would have *predictive* validity. Classroom exams, in contrast, typically are designed to measure student learning/achievement of the content and skills covered in the class; the extent to which a classroom exam truly measures this classroom content and skills is its *content* validity.

Law School Exams and Validity

Law school exam validity is primarily about content validity—does a torts exam, for example, measure the torts concepts (such as the concept of tortious intent) and skills (such as being able to spot viable legal claims in a fact pattern and predict their likely outcome in court) students are expected to have learned in the course. There is also a dimension of construct validity to law school exams, as they are designed in part to measure the students' ability to "think like a lawyer." In some cases they have a predictive component, as when exams in first-year courses may be designed in part to identify which students appear to be (or not be) on track to successfully continue in law school, be able to pass a bar exam, and eventually competently practice as attorneys.

Multiple Choice Exam Components and Exam Validity

Multiple choice components help me to be comfortable that my course grades are valid measures of individual students' mastery of the course concepts and skills. All-essay exams can measure student performance at a deep level, but because of time limits such exams tend to measure a small subset of course concepts and skills and tend to lack breadth.[6] Student performance on such exams thus reflects in part whether the few tested concepts and skills were ones individual students happened to learn well, which lessens validity. On the other hand, for me course grades based solely on multiple choice would not test certain of the skills I try to teach (for example, coming up with each party's arguments about how tort doctrines apply to a complex fact pattern) and thus would not be a highly valid basis for assigning course grades. A course evaluation package including both essay and multiple

6. For example, in the second (two credit) semester of Torts, I cover damages, products liability, defamation, misrepresentation, statutes of limitation, an introduction to consumer law, invasion of privacy, interference with contractual and business relations, and tort reform. An all essay exam could not hope to cover all of this ground.

choice components allows me to measure a wider subset of concepts and skills, some deeply, and thus enhances the validity of my course grades.

Research identifying concerns about multiple choice formats, particularly that which focuses on group differences in performance, suggests a significant potential limitation on the validity of multiple choice items and is another reason I am not comfortable assigning grades based solely on performance on multiple choice exams. Briefly, research suggests that on the LSAT and some multiple choice exams, women and students of color score somewhat lower than do white male students; there is also research suggesting that women score somewhat higher than men on essay exams.[7] I have compared the performance of my male and female students on my essay and multiple choice questions and have not identified any statistically significant gender differences.[8] Although I have not identified bias in my exams, avoiding any bias is essential and another reason for me not to base grades solely or primarily on multiple choice questions.

Exam Reliability[9]

Whatever an exam measures, *reliability* is about whether the exam does so consistently. Exams that are not reliable are ones on which students' scores vary considerably, and for the wrong reasons. Exams that lack reliability produce scores that are due significantly to chance and other irrelevant factors, rather than scores that are due to different levels of student performance, and are thus not a good basis for assigning letter grades, or any other purpose.

7. See, e.g., William C. Kidder, Portia Denied: Unmasking Gender Bias on the LSAT and Its Relationship to Racial Diversity in Legal Education, 12 Yale J. L. & Fem. 1, 6 (2000) (as a group, women's LSAT scores are lower than men's scores). See also Roy Freedle, How and Why Standardized Tests Systematically Underestimate African-Americans' True Verbal Ability and What to Do About It: Towards the Promotion of Two New Theories with Practical Applications, 80 St. John's L. Rev. 183 (2006); see generally Derrick A. Bell, Jr., Law School Exams and Minority-Group Students, 7 Black L. J. 304 (1981). For a comprehensive look at gender issues in evaluation, see Warren W. Willingham and Nancy S. Cole, Gender and Fair Assessment (Mahwah, N.J., 1997).

   There are common sense precautions to minimize the pitfalls identified by these commentators in classroom multiple choice tests, such as clear instructions with regard to guessing (e.g., if the test is scored by the number of correct answers with no penalty for guessing, make it clear to students that they should choose an answer for all items), creating items which are not more familiar or friendly to one gender or racial group, and writing an exam which is not overly time-pressured.

8. Unfortunately, I have never had a large enough group of students of color in my classes to perform an analysis of their performance.

9. For more extended discussions of exam reliability, see Anastasi, Psychological Testing, *supra* note 4, at 109-38; Hopkins and Stanley, Educational and Psychological Measurement, *supra* note 4, at 113-40; Josephson, Learning & Evaluation in Law School, *supra* note 2, at 15-26; Friedland, A Critical Inquiry, *supra* note 1, at 160-63.

Types of Exam Reliability

There are many ways to measure exam reliability—for example, comparing the same students' scores on different forms of the same exam (*alternate-form reliability*), comparing student scores on odd- versus even-numbered exam items on the same exam (*split-half reliability*), and correlating students' scores on each item with their performance on the exam overall (*Spearman's alpha coefficient,* or *Kuder-Richardson 20 or 21* are common formulae for measuring this kind of reliability).

The Relationship between Exam Reliability and Validity

Exam reliability is necessary, but not sufficient, for exam validity. In other words, an exam cannot be valid if it is not reliable. Measurement can be reliable without being valid. If, for example, I measured each of my torts student's height at several points during a semester and assigned letter grades based on average measured height (for example, students 6 feet 2 and taller receive an "A"), that measurement would be highly reliable, since adult student height is quite consistent. However, this evaluation technique would have no validity, since height has no relationship to mastery of torts concepts and skills.[10]

Reliability and Exam Length

As a general matter, long exams tend to have higher reliability than short ones.[11] This makes sense; a ten-item multiple choice torts exam is testing only a few things out of many that were taught. Scores will depend significantly on whether those few tested concepts were ones individual students knew well. In contrast, a forty-item torts exam is likely to cover many of the concepts covered in class, and thus scores will be more likely to reflect their level of mastery of the concepts. This does not mean that short quizzes should be avoided. While individual quizzes may not have high reliability, combining scores on several quizzes with final exam scores is likely to result in a semester-long evaluation "package" that is a reliable basis for assigning course grades.

Multiple Choice Exam Components and Exam Reliability

For me, multiple choice exam components are an important way to assure that when I assign course grades, my basis for doing so is reliable.[12] More specifically, multiple choice exam components allow me to measure student mastery more than once during a course. They reduce the potential problem that course grades based on a single exam may reflect only whether

10. Such an evaluation would also result in gender biased grades, since male students on average are significantly taller than female students and thus male students would on average earn higher grades.

11. See Anastasi, Psychological Testing, *supra* note 4, at 121; Hopkins and Stanley, Educational and Psychological Measurement, *supra* note 4, at 126, 131.

12. See Crane, Grading Law School Examinations, *supra* note 2, at 793 (for exams of same length, multiple choice exam is more reliable than essay exam).

each student had a "good" or "bad" final exam day. As discussed above, they also allow me to measure a broad variety of the course's concepts and skills, which also enhances the reliability of my grades.

Formative and Summative Evaluation[13]

Formative Evaluation

Basing course grades on more than one evaluation creates an evaluation system with a significant "formative" (or diagnostic) component. *Formative* evaluation offers the student feedback about performance before a final judgment is passed on the student, thus offering students an opportunity to identify areas where they need further work. For example, a student's performance on a torts quiz might indicate the student is confused about a specific concept or needs more work on a skill.

Summative Evaluation

In contrast to formative evaluation, which looks forward to ways in which students might continue to learn, *summative* evaluation looks back and judges how much students have learned.[14] A final exam (which in law schools is often the sole basis for course grades)[15] is an example of summative evaluation. Law

13. For more extended discussions of formative and summative evaluation, see Gronlund and Linn, Measurement and Evaluation in Teaching, *supra* note 4, at 12-13, 111-13, 459-61.

14. Another context in which law faculty may be sensitive to evaluation as formative versus summative is evaluating colleagues for retention, tenure, and promotion. In my experience, junior faculty understandably very much want formative evaluation from more senior colleagues so that they can build on strengths and work on weaknesses in the pretenure period. Faculty retention/promotion/tenure committees' primary responsibility, however, may be summative evaluation; to pass judgment on the candidate's demonstrated abilities in teaching, scholarship, and other job responsibilities.

    In my experience serving on such committees, it is difficult for a faculty committee to have sole responsibility for both kinds of evaluation, and the candidate and the school are better served with the two kinds of evaluation being performed by separate bodies. For example, faculty teaching and scholarship mentors may offer formative evaluation to a junior colleague, while a faculty tenure committee performs a summative evaluation recommendation of the colleague's demonstrated fitness for retention, tenure, or promotion. Even if the committee performs the only evaluation, however, the retention recommendations, for example, offer the candidate feedback about strengths and weaknesses before the tenure decision is made.

15. See Steve Sheppard, An Informal History of How Law Schools Evaluate Students, With a Predictable Emphasis on Law School Final Exams, 65 UMKC L.Rev. 657, 665-88 (1997). Sheppard notes that use of multiple choice questions in law school began following the creation of the LSAT in 1947. *Id.* at 683-86. Clinical Legal Education Association, Best Practices for Legal Education at 214 (August 2005 draft), available at <http://professionalism.law.sc.edu/downloads/bestpractices/20050831-Text.pdf> (last visited Jan. 15, 2008) ("For the most part, law schools and law professors...administer a single exam at the end of each semester-long or even year-long course. Typically, these exams are written essay exams requiring students to apply legal principles to hypothetical fact patterns. For many reasons, this practice is inadequate.... Law schools and law teachers should develop and use more

students can (and hopefully are encouraged to) review final exams and course grades to get feedback about apparent strengths and weaknesses, and thus final exams can have a formative component; however, they are primarily summative. Final exam scores and grades are first and foremost judgments about a student's mastery of the course material.

Multiple Choice Exam Components as a Vehicle for Formative Evaluation

Some legal commentators agree with educators that it is important to provide students with formative evaluation to maximize learning.[16] Formative evaluation can be mid-course quizzes or can be somewhat less formal and ungraded, such as providing students with practice exams and problems and offering feedback to students who participate in class.

For me, the multiple choice exam format, with the help of computer scoring, gives students timely formative evaluation. Given my normal teaching load and other responsibilities, it is not feasible to give and grade in a timely fashion an essay mid-term to a large class. A multiple choice quiz offers prompt, formal feedback during the course so that students can adjust as needed before the course is over.[17]

Norm-Referenced and Criterion-Referenced Exams[18]

Criterion-Referenced Evaluation

Criterion-referenced tests measure student performance against an external objective standard. For example, giving A's to students who demonstrate mastery of 80 percent or more of course learning objectives as measured on an exam would be a criterion-reference based grading system. In this approach, all students can earn A's (or C's or F's) depending on the degree

---

comprehensive methods of measuring law student performance than the typical end-of-the-term examination. Students should be given detailed critiques of their performance.").

16. One commentator notes that the single end-of-course examination can enable the teacher to become disengaged from her students' learning. Philip C. Kissam, Law School Examinations, 42 Vand. L. Rev. 433, 471-74 (1989). Basing course grades on a single exam may also increase student perceptions that the exam, and/or its grading, is unfair, Sheppard, An Informal History, *supra* note 15, at 693-94, and likely also increases the stress level of the examinees, thus potentially reducing the validity of the exam as it is measuring stress level rather than solely mastery. Another notes that law schools "have undervalued assessment as a teaching tool and overvalued evaluation as an accurate, objective measuring device." Friedland, A Critical Inquiry, *supra* note 1, at 152.

17. For commentary on the importance of prompt, during-the-course, student feedback to learning, see Munro, Outcomes Assessment, *supra* note 2, at 151; Terri LeClercq, Principle 4: Good Practice Gives Prompt Feedback, 49 J. Legal Educ. 418 (1999).

18. For more extended discussions of norm-referenced and criterion-referenced evaluation, see Anastasi, Psychological Testing, *supra* note 4, at 71-106; Hopkins and Stanley, Educational and Psychological Measurement, *supra* note 4, at 102-07; Josephson, Learning & Evaluation in Law School, *supra* note 2, at 4-5.

to which they demonstrate mastery of the criteria; it does not matter how their classmates perform.

Norm-Referenced Evaluation

In contrast to a criterion-referenced exam, a *norm-referenced* test is designed to separate out levels of learning within a group; the students earning A's under this approach do so because they have the top scores in the group. For example, the norm-referenced LSAT measures legal reasoning ability of individuals in comparison to the national pool of test-takers. For example, if a 153 on the LSAT is the 50th percentile or median, individuals earning this score are squarely in the middle of the large group of test-takers, and have average legal reasoning ability for this group. The 153 LSAT score says little or nothing about the student's absolute level of legal reasoning ability; it is about comparing this individual's performances to the group's.

Course Grades and Norm-Referenced Versus Criterion-Referenced Evaluation

An "A" in a course in which grading is completely criterion-referenced means the student demonstrated excellent mastery of the course concepts and skills, without regard to how other students did. An "A" in a course in which grading is completely norm-referenced means the student's performance ranked at the top of the class, without regard to whether this student or the class as a whole performed at an excellent, mediocre, or poor level.

Law School Grades, Grade Curves, and Norm-Referenced Versus Criterion-Referenced Evaluation

Law school grades have both norm-referenced and criterion-referenced components. At most law schools, grade curves require law teachers to assign grades with a certain average, and/or a certain distribution.[19] At my law school, for example, in first-year courses other than Legal Research and Writing, grades must average between 2.6 and 2.9 (where 2.7 is a B-).[20]

Norm-referenced Aspects

The grade curve forces a significant level of "norm-referenced-ness" into grades; even if every student in my torts class demonstrated excellent mastery of course concepts and skills, they cannot all receive A's. On the other hand, the central limit theorem[21] posits that in a large group, performance approximates a

19. See Nancy H. Kaufman, A Survey of Law School Grading Practices, 44 J. Legal Educ. 415, 417-18 (1994) (two thirds of law schools in a national survey use a grade curve).

20. Gonzaga University School of Law Academic Rule 2(a), Letter Grades and Numerical Values, available at <http://www.law.gonzaga.edu/Files/Students/2006-2007StudentHandbook.pdf> (last visited Nov. 7, 2007).

21. According to the central limit theorem, law school exam scores will approximate a bell curve the larger the group being measured. See, e.g., Hopkins and Stanley, Educational and Psychological Measurement, *supra* note 4, at 57.

bell curve, in which case typical law school grade curve requirements will likely not prevent course grades from reflecting each student's level of mastery as well as each one's standing in comparison to the group. In fact in fifteen years of full-time law school teaching I have had only one large class in which the grade curve requirements prevented me from assigning course grades that reflected levels of mastery as well as class rank.[22]

In many law schools, grade curves do not apply to certain small elective courses. At my law school the grade curve does not apply to elective courses with enrollments of less than twenty-five.[23] This appropriately reflects that in small classes class performance may not resemble a typical bell curve. In those classes, a grade curve could force grades that do not reflect absolute levels of performance. First, students who will achieve at high levels in the course, because they are top students or are interested in the subject, may self-select such a course, resulting in an overall high level of class mastery. Second, my experience in small courses is that students tend not to fall through the cracks and perform at a low level.[24]

Finally, in some law schools certain courses are exempt from the grade curve, or the teacher can present a criterion-referenced evaluation plan for a course and seek approval to grade on this basis rather than the grade curve, or in the event of a really good or bad set of exams can request a waiver from grade curve requirements. For example, at my law school, faculty may request administrative waiver. Moreover, students in our in-house clinics are graded based on their performance on a set of specific criteria and the grade curve does not apply.

Criterion-referenced Aspects

Most law teachers assign grades at least in part because of their judgment about the student's performance level. Law teachers have significant discretion in assigning letter grades even with a grade curve. If I do not see a lot of variation in performances in a class, I can submit a compressed curve with few extreme (very high or very low) grades. I can, and do, consider the average level of class performance in deciding where within the required 2.6 to 2.9 range to put the average grade. I also consider the absolute levels of performance demonstrated and when performances seem truly different in assigning letter

22. The class in question was a heavily enrolled elective class in labor law with multiple evaluations based on small group simulations, for which many top students self-selected and invested enormous time and energy. I was unsuccessful in getting the then-academic dean to waive the grade curve, and more than a dozen years later still feel bad that students' grades underrepresented their absolute level of performance (for example, some students who performed at an excellent level received grades of B and B+).

23. Rule 2(a), *supra* note 20.

24. I think this is so because in small classes each student participates more, learns more because the stress level is somewhat lower than in large classes, and as the teacher I am better able to identify students who are not doing well early and intervene to help them perform better.

grades (for example which point totals will be A's) and in making letter grade cutoffs (for example which scores will be B-'s and which C+'s).[25]

Finally, I perform somewhat of a gatekeeper function in assigning low letter grades, particularly in first year classes. For example, I will not fail a student solely because her exam performance is much lower than her classmates'. For me an F means the student has not demonstrated minimally adequate mastery of course concepts and skills and should retake the course if it is required and she remains in law school. Despite a nickname reportedly given to me by certain students,[26] a grade of D+ or D from me means the student has demonstrated barely adequate learning of course concepts and skills and more generally should not continue in law school unless grades in other courses reflect considerably more mastery. Especially for first-year students, a very low course grade also represents my judgment that this student may not be one who can be successful in law school, and ultimately on the bar exam, and practicing law.

Multiple Choice Questions and Norm- and Criterion-Referenced Evaluation

When using multiple choice (or other) exam item formats, law teachers have to decide what the purpose(s) of the items is, including whether the item is designed to separate out levels of learning within a class (norm-referenced evaluation), or to measure whether students have mastered specific concepts or skills (criterion-referenced evaluation). The two types of evaluation involve somewhat different kinds of items. For example, if the primary goal of the exam is norm-referenced, items that most of the class answers correctly (or incorrectly) do little to separate out levels of performance and thus to form a basis for assigning different letter grades.[27] Conversely, if the goal of the exam is to determine whether students have achieved a certain level of mastery of specific concepts and skills, items that are properly designed and that most of the class answers correctly are a positive event, suggesting widespread mastery within the class.

As the next part discusses in greater detail, for criterion-referenced evaluation, computerized scoring provides useful information about the overall levels of class performance. For norm-referenced evaluation, the computerized scoring information specifically measures whether individual exam items are successfully separating out levels of performance on the one hand, or appear to have significant technical difficulties on the other.

25. To help with this, I often record a holistic impression of an exam (e.g. "not bad" or "deserves a C+") after grading it using a point-based rubric.

26. "D+ Daggett."

27. For example, if the average student answers 85 percent of items correctly, there is not much space between the average score and the maximum score, and letter grade cutoffs will result from very small score differences.

The Computerized Scoring Information–What It Is and How It Can Be Used

Appendix I consists of excerpts from a computerized scoring report for the multiple choice component of a recent actual quiz ("Torts Quiz") I gave to two sections of torts students after we completed the first course unit on intentional torts.[28]

Basic Information

Computerized multiple choice scoring services provide basic information about overall exam performance. The reports include the number of exams graded (142 students for Torts Quiz) and the number of items graded (10 items for Torts Quiz), which is helpful in making sure all students took the exam and all exams were scored. Also typically reported are the low and high scores in the class (2 and 10 respectively for Torts Quiz) and the range of scores (the term for the difference between highest and lowest scores; the range is not reported for Torts Quiz but is 8 (10-2)).

Measures of Central Tendency–Eeny Meany, Median, Mode

The reports also typically include various measures of central tendency.[29] Three statistics indicate the center of the students' scores. The mean score (6.32 for Torts Quiz) is the arithmetic average.[30] The median (6.0 for Torts Quiz), also the 50th percentile, is the "middle" score after scores are ranked from high to low.[31] The mode is the most commonly earned score (not reported for Torts Quiz, but is 7.0).[32]

The mean is the best measure of central tendency unless the distribution is skewed (meaning scores tended to be clustered disproportionately at the high or low end of the range);[33] in that case, the median is the best measure. In a large class, because the central limit theorem predicts a normal distribution (bell curve) of performance, the mean is usually the best measure of the "center" of the students' exam performance. In the Torts Quiz, the mean and

28. The full quiz, answers, scoring matrix, and explanatory handout are available from the author.

29. For more extended discussions of measures of central tendency, see Anastasi, Psychological Testing, *supra* note 4, at 74-75; Hopkins and Stanley, Educational and Psychological Measurement, *supra* note 4, at 21-32.

30. The mean is computed by adding all scores and dividing by the number of students (X = sum of all scores/N).

31. If there are an even number of scores the median is the midpoint between the two middle scores. For example if there were six students, the median would be halfway between the third-highest and fourth-highest scores. If there were five students, the median would be the score of the third-highest exam.

32. See the frequency distribution, which shows 7 is the most common score.

33. For graphical examples of skewed distributions and related discussion, see Anastasi, Psychological Testing, *supra* note 4, at 207-09.

median are fairly close (6.32 and 6.00 respectively) suggesting little skewing. As a contrasting example, student end-of-semester teacher evaluations are often skewed toward the higher scores and often contain a few very low scores from unhappy students, which drag down the mean. The median is the more accurate measure of central tendency in such cases.[34]

Confidence intervals for the mean may also be reported, although they are not reported by the software used at my university and I do not find them particularly useful. These calculate the range of the true mean for the exam, given its reliability. Often, for example, a 95 percent confidence interval may be reported, which shows the range of scores within which we can be 95 percent sure is the true mean for the class. Typically these are small ranges and indicate we can be confident the calculated mean is close to the true one, which in turn suggests the reliability of the exam is fairly high. Law teachers may not be comfortable assigning different letter grades to scores that are within the confidence interval of the mean.[35]

Measures of Variability—How Closely Do Scores Cluster Around the Mean?

Measures of variability are about how closely (or not) scores cluster around the mean. The range (8 for Torts Quiz) (spread from highest to lowest score) is a very crude measure of variability.[36] Certain benchmarks may also be reported, such as the 25th percentile (the score which marks the bottom 25 percent of the class or bottom quartile) (not reported for Torts Quiz but is 5) and the 75th percentile (the score which marks the top 25 percent of the class or top quartile) (not reported for Torts Quiz but is 8) and the Inter Quartile range,

34. As an example, a class of 50 rates their teacher on a 7 point scale, with 7 the highest. The ratings are as follows:

Rating	N
1	6
2	0
3	0
4	11
5	19
6	14
7	0

The median is the rating between student # 25 and 26, if all 50 ratings are ranked from high to low. In this example the median is 5. The mode is also 5. However, the mean of this same set of scores is 4.38. The few students who gave the teacher very low ratings pull the mean down considerably from what appears to be the middle of the ratings. While "1" ratings from 12 percent of the class should not be ignored, 5 is the more accurate measure of how the class as a whole ranked the teacher.

35. For example, if the 95 percent confidence interval for the 6.32 mean on the Torts Quiz were 5.62 to 7.02, it may not be appropriate to assign different letter grades to scores of 6 and 7 respectively.

36. If the computerized scoring system does not compute this, it is easily done on a spreadsheet. Excel and other spreadsheets have "min" and "max" functions, or the data can simply be sorted by score to identify the high and low scores.

or difference between the top and bottom quartiles (not reported for Torts Quiz but is 3).

Standard Deviations

The standard deviation (1.76 for Torts Quiz) is the most important measure of variability. Variance is the standard deviation squared (not reported for Torts Quiz but is 3.10). Standard deviations assume a normal distribution of scores, or a bell curve.[37]

Standard deviations show the range of class performance and help both the teacher and students to understand how much differently than average a student scored. About two-thirds of the class will be within one standard deviation from the mean. Students who score +/- 1 standard deviation from the mean are within the mainstream, or middle two-thirds of the class. Thus on the Torts Quiz students can be told that if they scored within a certain point range (5 to 8) they were in the middle two-thirds of the class for the multiple choice component. If they scored lower than that point range (2, 3, or 4), they scored significantly below average (bottom sixth of the class), and if they scored higher than that point range (9 or 10), they scored significantly higher than average (top sixth of the class). Ninety-five percent will be within two standard deviations (6.32 +/- 3.52, or from 3 to 9), and 99.7 percent will be within three standard deviations (6.32 +/- 5.28, or scores of 0 or 1; the highest possible score of 10 is less than three standard deviations above the mean).

Standard Deviations as a Measure of Variation in Performance

Standard deviations show how much performance (or ratings in the case of student evaluations) varies. On an exam, the standard deviation shows how much student knowledge/mastery apparently varies. On the Torts Quiz a standard deviation of 3 would indicate wide variation in student performance and mastery. If the Torts Quiz standard deviation was 0.5, it would indicate student performance was closely bunched together (and consequently small score differences would translate into different letter grades).[38]

37. For more extended discussions of standard deviations, see Anastasi, Psychological Testing, *supra* note 4, at 75-77; Hopkins and Stanley, Educational and Psychological Measurement, *supra* note 4, at 34-40. If computerized scoring is not available, Excel, Quattropro, and other spreadsheets have a built-in function for computing standard deviations. For more extended discussions of variance, see Anastasi, Psychological Testing, *supra* note 4, at 75-77.

38. As another example, on teacher evaluations, the standard deviation shows how much student opinion about a teacher varies. For example, two teachers may get evaluation ratings with identical medians of 5.0. However, Teacher A has a standard deviation of 1.3, and Teacher B a standard deviation of 0.4. There is a close consensus within the class on the "5" rating for Teacher B. Students gave widely differing ratings to Teacher A; she has much more mixed reviews than does Teacher B.

Z-scores[39]

Standard deviations may be used to calculate z-scores, the standardized scores representing the number of standard deviations a student's score is from the mean.[40] Computing z-score or other standardized scores on multiple assignments such as quizzes and a final exam (each with its own mean and standard deviation) allows a teacher to calculate course grades with each assignment weighted as the instructor has set out at the beginning of the semester. For example, on the Torts Quiz a score of 8 equals a z-score of +0.95 ((8-6.32/1.76).

Using Standard Deviations and Z-scores to Assign Letter Grades

Standard deviations can help the teacher to determine when a score represents a truly different performance than another, for example, when assigning letter grades. On a reasonably reliable exam, a difference of a full standard deviation, or even a half standard deviation, between two scores is a real one; conversely, a difference of less than one quarter standard deviation between scores may be due to chance rather than real differences in performance.

Hypothetical Example

On a final exam with an average score of 69 points and a standard deviation of 9, a score of 64 is almost certainly not statistically significantly different from a score of 66 (difference of only 0.22 standard deviations). The 64 score is 0.55 standard deviations below the mean, which is the point at which the teacher may find the performance to be significantly below average and deserving of a lower letter grade than average (for example, in a class with a B- curve, the score of 64 may be a C+, or could reasonably be a low B- in a curve with a wide point range for B- grades).[41]

Appendix II provides some additional information and examples of using standard deviations and z-scores to combine grades on multiple exams and to assign course grades.

The Torts Quiz

The Quiz included both multiple choice and short essay components. With a spreadsheet I computed a standard deviation of 6.26 for the exam overall.

39. For discussions of z-scores and other standard scores, see Anastasi, Psychological Testing, *supra* note 4, at 84-86, 90-91; Hopkins and Stanley, Educational and Psychological Measurement, *supra* note 4, at 52-53; Paul T. Wangerin, Calculating Rank-in-Class Numbers: The Impact of Grading Differences Among Law School Teachers, 51 J. Legal Educ. 98, 101-03 (2001).

40. Thus in a normal distribution the mean of a set of z-scores will be zero, and the standard deviation will be one. Anastasi, Psychological Testing, *supra* note 4, at 84-86.

41. In classes where I wished a 2.7 or B- average, I have sometimes assigned B- grades to scores within 1/3 of a standard deviation of the mean (66 to 72 in this hypothetical), and sometimes to scores within ½ of a standard deviation of the mean (64.5 to 73.5 in this hypothetical). In all cases, I look at natural breaks in the point scores and my holistic impression of exams on the cusp to tinker with the letter grade cutoffs.

Appendix III explains the Torts Quiz scores and shows z-score calculations. I chose not to assign letter grades for this quiz, as it was early in the first year and I wanted to help my 1L students keep the results in perspective.[42] If I had assigned letter grades, given the mandatory B- curve at my law school, I would have used standard deviations and z-scores and the process outlined in Appendix II to help set letter grade cutoffs.

Test Reliability

Computerized scoring typically includes reliability coefficients such as KR (Kuder Richardson) 20 or 21 (0.47 and 0.28 respectively on Torts Quiz) and/or Cronbach's Alpha coefficient.[43] As discussed earlier, reliability is about how consistently a test measures whatever it is measuring. These coefficients are two different formulae for computing a specific kind of reliability–internal consistency, or the extent to which the test is measuring a single thing (e.g. Torts mastery). These formulae compute the correlation, for all items, between getting the item correct and doing well on the exam overall. If the test measures one thing, the correlation should be high (+1.0 is the maximum); if it is measuring many things the correlation should be low (0 is the minimum). Although on the low side of the possible range, internal consistency reliability coefficients around 0.30 are about what I would expect for a rather short (twenty items or less) multiple choice quiz or component of a law school final exam where many concepts are being assessed.[44] If on a short multiple choice quiz or exam component these figures were closer to 0 (0.15 or below) I would be concerned about the test's reliability. Longer exams, as well as the overall internal consistency reliability coefficient for all the exams in a course, would be expected to have much higher reliability.[45]

42. As Appendix III notes, I did not and would not assign letter grades based on the Torts Quiz as it was too small a sample of student learning, and I was concerned about students, one month into law school, keeping the proper perspective on their first formal feedback.

43. Kuder-Richardson reliability fomulae essentially average all possible split halves of an unspeeded exam. Hopkins and Stanley, Educational and Psychological Measurement, *supra* note 4, at 130-32. They are heavily dependent on test length. Kuder-Richardson 21 is somewhat easier to compute than Kuder-Richardson 20, but is less accurate as it assumes items of equal difficulty. The formulae are not useful for speeded exams. The Kuder-Richardson reliability formulae are more specific applications of and attempts to measure the same kind of reliability as the Cronbach alpha coefficient and normally produce similar results. *Id.* at 130-33; Cronbach, Essentials of Psychological Testing, *supra* note 4, at 202-05.

44. Contrast Crane, Grading Law School Examinations, *supra* note 2, at 795 (Spearman-Brown coefficients of at least .5 are desirable for tests measuring a single domain, but "a test with good internal reliability yields items that correlate anywhere from .25 to .5."); *id.* at 796 (law school exams measure several domains, for example on a Property exam adverse possession and future interests are separate domains).

45. *Id.* at 796 (suggesting at least forty questions are needed per item to achieve desired internal consistency and other reliability levels). I would also expect higher reliability coefficients on a lengthy exam for other kinds of reliability such as split-half reliability. *Id.* at 795-96 (for these kinds of reliability, a coefficient of 0.8 or higher is desirable).

Frequency Distributions

Computerized scoring reports typically include a graphic frequency distribution,[46] which shows the number of students who earned each possible raw score on the exam. For example, the frequency distribution for the multiple choice component of the Torts Quiz is:

Score	# of students	Graphic Representation
10	3	***
9	11	**********
8	25	**********
7	31	**********
6	29	**********
5	17	**********
4	17	**********
3	7	*******
2	2	**
1	0	
0	0	

The Torts Quiz shows the bell curve distribution expected in a large law school class. As the scores taper off at the maximum (10) or minimum (0) scores, there appears to be enough "ceiling" and "floor," or room at the top and bottom, to separate out the students within the highest and lowest performing subgroups. This distribution has a single high point at 7 and is thus "unimodal," as expected with a bell curve. Sometimes the distribution is bimodal or trimodal,[47] where instead of the single central bump of the bell curve, there are several bumps, suggesting there are two or more subgroups of students with similar levels of mastery.[48]

46. For more extended discussions of frequency distributions, see Anastasi, Psychological Testing, *supra* note 4, at 73-77, 207-09; Hopkins and Stanley, Educational and Psychological Measurement, *supra* note 4, at 22-23.

47. For discussions of bimodal and trimodal curves, see Ebel and Frisbie, Essentials of Educational Measurement, *supra* note 4, at 58-59.

48. I have had both bimodal and trimodal distributions, and when this happened it matched my perception that there were several "clusters" of students in my large class with similar levels of mastery.

Item Analysis

Computerized exam item analysis is part of the typical computerized scoring report and provides a wealth of information about the effectiveness of the items and the performance of the students.[49]

Identifying Possible Scoring Errors

Computerized scoring typically generates codes or marks a student's response as "other" to a question (as in the Torts Quiz item analysis) when the computer finds any unanswered questions or when multiple answers have been recorded, for example when a student has not fully erased a first response. The simple solution is to hand check the individual answer sheets to be sure the computer scored the item correctly and to note this on the scantron sheet in case the student is confused about scoring.[50]

Item Analysis

Item analysis statistics offer information about how well each item worked in specific technical ways, as well as specifics about what areas students have (not) mastered.

Item Difficulty

Item difficulty is the percentage of students who chose the correct answer, expressed as a fraction of 1.00 or as an actual percentage.[51] For example, for Torts Quiz Question 5, "E" has been keyed as correct and was chosen by 59 of 142 students, or 41.5 percent, so its item difficulty is 0.415 or 41.5 percent. For Torts Quiz Question 10, "D" has been keyed as correct and was chosen by 127 of 142 students, or 89.4 percent, so its item difficulty is 0.894 or 89.4 percent.

In a criterion-referenced test, where the goal is for all students to demonstrate mastery of the criteria, a very high pass rate is desirable. As discussed earlier, most law school exams are in significant part norm-referenced tests designed to

49. For more extended discussions of item analysis, see Anastasi, Psychological Testing, *supra* note 4, at 202-36; Hopkins and Stanley, Educational and Psychological Measurement, *supra* note 4, at 269-88.

    This article does not address the design of effective multiple choice questions. Effective design involves both representative and broad coverage of the content to be tested, and careful thought about the level of learning to be tested. As to the former, I normally make a list of the specific concepts and skills covered and try to write a set of items that covers most of this list. As to the latter, I generally try to write items that test at the analysis or higher levels of Bloom's Taxonomy of levels of learning, see Benjamin Bloom, Taxonomy of Educational Objectives (New York, 1956), rather than, for example, items which merely ask students to recall concepts or apply them in a simplified manner, and I also try to write items of varying difficulty to aid in separating out levels of student learning.

50. I do this and also make a photocopy of all scantron sheets before they are scanned. In the (rare but real) event that a student claims a scoring error on the scantron sheet, I can review the photocopy to be sure the scantron has not been altered. Making a copy also protects me and the students in the event the scantron sheets are lost or damaged during scanning.

51. So, 0.63 means 63 percent of students answered the item correctly.

separate out different levels of student mastery (and thus to assign letter grades), and so a very high (or very low pass) rate is undesirable. If virtually all students perform the same on an item (passing or failing it), that item cannot separate out levels of mastery.

To enhance exam fairness, I find it helpful to reconsider any item on my exams with a pass rate under 50 percent, such as Torts Quiz Question 5. First, I reread the item to make sure I have not made an error in wording or miskeyed the item. Second, I look at what the low pass rate on the item suggests in terms of a concept likely not yet mastered by the class and/or not taught effectively. I list these apparently-not-yet-mastered concepts and share them with my students as part of general feedback about the exam, or for retesting mastery of those concepts later in the course, and/or for possible additional instruction or practice opportunities later in the course.

It is helpful to look at the percent of students who chose each of the incorrect options (called "distractors") for each question. I reconsider any distractor chosen by at least 25 percent of students, first to verify I have not made a wording error or miskeyed the question, and then to reflect on what caused students to choose this distractor. As with items not answered correctly by most students, a list of "attractive" (frequently chosen) distractors can be shared with students as part of general feedback, or for retesting mastery of those concepts later in the course, or for possible additional instruction or practice opportunities later in the course. For example, on Torts Quiz Question 5, not many students chose options A (0 of 142) or C (11 of 142); more (28 of 142) chose option D, while quite a few students (44 of 142) chose option B, making it an attractive distractor. I would take a close look at attractive distractor B to make sure there is no error and then reflect on what it tells me about the students' (mis)learning.[52]

One way to check for the test being too long for students is to see if the item difficulty and discrimination decline with the last few items. Such a pattern may indicate that students did not have enough time to perform well on the last few items of the exam.[53]

Item discrimination, or (Point) Biserial Correlation

Item discrimination, or (point) biserial correlation, looks at how students who did well or poorly overall performed on an item. In other words, these are measures of how performance on a single item correlated with overall performance.[54] As discussed earlier, the most important kind of validity for

52. Although not the attractive distractor for this question, on my exams the option "two or more of the above-listed options are likely correct," is sometimes a popular option both for students whose mastery is incomplete and also for those who lack confidence in their answers.

53. Unless the teacher intended the exam to measure performance under speeded conditions, this pattern would raise a concern about exam validity.

54. Note that the default computation of the discrimination index and point biserial coefficient relates performance on one multiple-choice item with performance on the multiple choice

a law school exam is content validity (e.g., is a torts exam measuring torts mastery). Item discrimination or (point) biserial correlations are a rough index of item validity for a norm-referenced test–how well is it separating out levels of overall mastery. Computerized scoring systems generate item discrimination or biserial correlation statistics; they can also be computed by hand.[55]

Item discrimination/(point) biserial correlation coefficients range from

-1.00 (which indicates all the low overall scorers got the item right and all the high overall scorers got it wrong) to

0 (the pass rates on the item for the low overall scorers and high overall scorers were identical) to

+1.00 (which indicates all the high overall scorers got the item right and all the low overall scorers got it wrong).

For a norm-referenced test, a fairly high positive item discrimination/(point) biserial correlation index is desirable.[56] A guideline for point biserial correlation[57]/item discrimination coefficients on a norm-referenced test is:[58]

---

overall, not with overall exam performance including both multiple choice and essay components.

55. To compute a discrimination index for an item: 1. Identify the overall scorers in the top 27 percent on the test and those in the bottom 27 percent (rounded off as appropriate) (some formulae use 25 percent, 30 percent, or 33 percent rather than 27 percent). 2. Calculate the percent pass rate for each item for these top and bottom scorer groups. 3. The pass rate for top group minus pass rate for bottom group = item discrimination (a number somewhere between -1.0 and +1.0). See Hopkins and Stanley, Educational and Psychological Measurement, *supra* note 4, at 271.

56. For extended discussion of item discrimination indices, see Anastasi, Psychological Testing, *supra* note 4, at 210-19; Hopkins and Stanley, Educational and Psychological Measurement, *supra* note 4, at 270-83.

57. The point biserial correlation assumes a true dichotomy (for exams, that there is a truly right and a truly wrong answer to items) but does not assume a normal distribution. See Paul Kline, A Handbook of Test Construction: Introduction to Psychometric Design 138-39 (New York, 1986). The point biserial coefficient is thus the better formula for multiple choice items, unless they are scored with some options being given partial credit.

Biserial correlation coefficients are consistently at least one-fourth larger than point biserial correlation coefficients. See Ebel and Frisbee, Essentials of Educational Measurement, *supra* note 4, at 232.

As compared with the item discrimination index, the biserial correlation coefficients depend less on item difficulty; thus, one can obtain high biserial coefficients for hard or easy items for which discrimination indexes are low. *Id.*

58. Hopkins and Stanley provide the following guidelines regarding item discrimination indexes:

Index of discrimination	
.40 and up	very good item
.30 -.39	good item

Point biserial correlation	Item discrimination	
+0.25 and up	+0.40 and up	Item is separating out levels of performance to a significant degree
+0.15 and below	+0.20 and below	Not a good item in terms of separating out levels of performance
Less than 0	Less than 0	Something is wrong. Item may be miskeyed or unfairly ambiguous

Mathematically, items with pass rates between 25 and 75 percent have the highest potential for high discrimination among scorers.[59]

Looking at Torts Quiz item 5, the discrimination index for correct option E is +0.67, indicating that this item is doing a very powerful job separating out levels of performance:

Item Analysis

	Upper Quartile	Lower Quartile	Total Count	Total %	Discrimination Index	Diffculty Factor
A	0	0	0	0	0	0.415
B	7	15	44	31	-0.22	
C	1	5	11	8	-0.11	
D	1	13	28	20	-0.33	
E	27	3	59	42	+0.67	
Other	0	0	0	0	0	

In contrast, Torts Quiz item 10 shows a discrimination index of only +0.19 for correct option D, indicating that this item is not doing much to separate out levels of performance:

.20 -.29 reasonably good item
.10 -.19 marginal item, subject to improvement
Below .10 poor item, to be rejected or revised

Hopkins and Stanley, Educational and Psychological Measurement, *supra* note 4, at 276.

See Kline, A Handbook of Test Construction, *supra* note 57, at 143 (suggesting point biserial coefficients of 0.20 and higher are desirable); Salvia and Ysseldyke, Assessment, *supra* note 4, at 164 (suggesting point biserial correlations of .25 to .30 and higher are desirable).

59. See Hopkins and Stanley, Educational and Psychological Measurement, *supra* note 4, at 272-73.

Item Analysis

	Upper Quartile	Lower Quartile	Total Count	Total %	Discrimination Index	Diffculty Factor
A	0	0	0	0	0	0.894
B	0	2	1	1	0	
C	0	1	0	0	0	
D*	35	28	127	89	+0.19	
E	1	7	14	10	-0.17	
Other	0	0	0	0	0	

When one looks at the item difficulty for Torts Quiz item 10 (0.89), the reason is clear. As almost everyone in the class got this item correct, it had little potential to separate out levels of performance. Note that Torts Quiz item may be a wonderful, extremely valid item content wise, as it appears to indicate that most of the class has mastered the concept(s) it tested. It just wasn't great at separating out levels of performance.

Looking at the discrimination/biserial correlation coefficients for the distractors is also extremely helpful. The distractors with high negative discrimination/biserial correlation coefficients were ones that were extremely attractive to low overall-scoring students but not to high overall-scoring students.[60] For example, on Torts Quiz item 5, the attractive distractor D has a high negative discrimination index of -0.33, indicating it was chosen much more often by overall low scoring students. This suggests that even though it was a difficult item that less than half the class answered correctly, it is not a "bad" item; the attractive distractor was apparently chosen because it was attractive to students with lower levels of overall mastery.[61]

Any distractor with a significant positive correlation (+0.15 or higher) was attractive to high overall scoring students, but not to low overall scoring students. These distractors should be reviewed carefully to see if they indicate a wording error or miskeying.[62]

Addressing Item Errors

If an item has been clerically miskeyed, (e.g., "A" was keyed as the correct answer to an item when "B" is the correct answer) it is simple enough to redo the key and rescore the exams. Somewhat more complicated is the situation where the item turns out to have been poorly worded, resulting in the initially

60. They were thus successful in separating out levels of mastery. These successful distractors also indicate common specific errors/areas of confusion for low overall-scoring students.

61. Distractors B and C for Question 5 were moderately attractive, tempting more low scorers than high scorers. Distractor A tempted no one.

62. There were no such distractors on the Torts Quiz.

keyed correct answer not being the clearly best option. Some teachers throw out those items from student scores. I identify the subset of options for the "bad" item that are the best answers, and by double (or triple) keying, mark each of them as correct. I choose this approach because even on a badly worded item, students spent time trying to choose the best answer, some with more success than others since some options are still clearly wrong (or at least more wrong than other available options for the item).[63]

Sharing Information with Students

In my syllabus, I explain how grades will be calculated.[64] As shown in Appendix I, printouts typically include a page of data for each student, including some overall class performance data—number of students, high and low scores for the class (mean and median), and data on this student's performance (in the example in Appendix I, 7 of 10 correct answers). It also shows the student what answer was recorded for the student on each question (e.g., in the example in Appendix I the student correctly chose option C on Question 1) and the correct answer for the questions this student got wrong (e.g., this student chose option D on Question 5 when the correct option was E).

I give my students their copy of this page, hopefully enhancing the formative evaluation function of my exams. I think doing so helps avoid student perceptions that the exam or my grading are unfair. I also choose to share most of the test statistics with students in an explanatory handout. Appendix III is the explanatory handout for the Torts Quiz. I share overall statistical performance information with students. For each item, I share and explain the item difficulty and discrimination/biserial correlation coefficient of the correct option so that students can both see how their responses compare to classmates and, presuming the items worked well technically, have some confidence that the test was effective. When there are common errors, I try to describe them to students to give those who need it some feedback on fertile areas for review.

Conclusion

Computerized multiple choice exam scoring does not eliminate the need for the difficult and time consuming work of writing good exam items, nor the complex judgments law teachers must exercise when translating point totals to letter grades. Rather, the information offers useful guidance to law teachers and law students: in identifying and dealing appropriately with any technical errors and thus enhancing the fairness of the exam, in getting and giving feedback on overall level of student learning as well as specific areas of mastery and lack of mastery for the class overall and for individual students, and in helping to understand when levels of performance are really different

63. I think my approach thus appropriately rewards those students who made one of the better choices on this item.

64. I specifically describe how I will combine scores on exams to arrive at course assignments. See Appendix II for alternate approaches to combining scores.

and consequently to enhance the fairness of assigned letter grades. It is well worth the law teacher's time to obtain and decipher the mysterious computer printouts that accompany our scored multiple choice exam items.

Appendix I
Torts Quiz (10/4/2005)
Excerpts from printout for Exam Number 59259

59259	Answer Key
	1234567890
	3233532114
Score	
7	· · · · 4 · 45··

Key Symbols		Respondent Symbols	
*	= Any response correct	1-9, 0, ?	= Wrong Answer
( )	= Question ignored	.	= Correct Answer
1-9, 0 , ?	= Correct response	*	= Multiple Responses
		-	= Not answered
		( )	= Not scored

Number of respondents	142	Average score	6.32
Total in Upper Quartile	36	Median Score	6.00
Total in Lower Quartile	36	Highest Score	10
Number of Test Items	10	Lowest Score	2
Kuder Richardson 20	0.47	Standard Deviation	1.76
Kuder Richardson 21	0.28		

Appendix II

A. Using z-scores to assign letter grades to point scores on an exam

First, decide what the grades signify: The student's rank within class (norm-referenced evaluation)? The student's level of accomplishment without regard to other students' performance (criterion-referenced evaluation; also called outcome assessment)? Some combination of the above (my usual approach)? Then:

1. Decide roughly where the middle or average grade will be, based on teacher's belief about level of class's performance as a whole (within any restrictions imposed by school grade curve policy). To do this, it can help to reread several exams which scored in the middle—do they read like a B-? a C+? a B? an F? It also helps to form holistic impressions as you score the exams (e.g., if you want the average to be B-/C+ (2.5), make the mean the cutoff between C+ grades and B- grades; if you want the average to be B- (2.7) make the mean the middle of B- grades; if you if want the average to be 2.85 make the mean the cutoff between B- and B grades).

2. Compute the average, standard deviation, and z-scores = number of standard deviations score is from mean = ((score-mean)/SD) for all students.

3. Set initial letter grade cutoffs. Mathematically, a good rule of thumb is that a 1/3 or 1/2SD between scores is a real and significant difference in performance. After establishing mean, set tentative letter grade cutoffs at 1/3 or 1/2SD increments around mean.

4. Finalize letter grade cutoffs by fudging for cusp grades and for natural breaks in scores.

5. Calculate letter grade average, SD, and distribution.

B. Combining grades on quizzes, etc. to compute course grades:

I am not advocating any specific method for computing course grades; in fact, I have used several different methods in recent classes. What is important is to understand the options and their consequences and to make a deliberate and informed choice. Specifically choosing whether to combine raw scores on exams, or standardized z-scores, can significantly affect the resulting grades. Examples of this are worked out below. As a matter of fairness, the method for calculating grades should be shared with students at the beginning of the course.

EXAMPLE 1:

Course has a midterm and final, which students are told each "count 50 percent."

On the midterm:	Average is 54 out of 100; standard deviation is 6
On the final:	Average is 88 out of 100; standard deviation is 12

a. To weight each test at 50 percent, calculate z-scores for each student for midterm and final

Student A got 57 on the midterm and 91 on the final.

Midterm Z	= ((57-54)/6)	= +0.5
Final Z	= ((91-88)/12)	= +0.25
Z Average	=((0.5 + 0.25)/2)	= +0.38

Student B got 54 on the midterm and 88 on the final.

Midterm Z	= ((54-54)/6)	= +0.0
Final Z	= ((88-88)/12)	= +0.0
Z Average	= ((0.0 + 0.0)/2)	= +0.0

b. If the raw points are added up instead:

Student A
57+91 = 148
% of grade from midterm points: 57/148 (39%)
% of grade from final points: 91/148 (61%)

Student B
54+88 = 142
% of grade from midterm points: 54/142 (38%)
% of grade from final points: 88/142 (62%)

Note that the final, which had the higher average, ends up being weighted much more heavily than 50 percent in course grade calculations. There is nothing invalid about calculating course grades this way; students had a chance to earn an equal number of points (100) on the midterm as on the final. However, because the class did much better on the final, scores on it end up being weighted more heavily. Also note that adding a constant or multiplying scores by a constant on one test to get equal averages usually will not result in them being equally weighted since the standard deviations are different.

EXAMPLE 2:

Course has a midterm and final, which students are told each "count 50 percent." Same average score of 65 on midterm and final.

On the midterm: Average 65 of 100; standard deviation is 5

On the final: Average 65 of 100; standard deviation is 20

Student A gets 75 on midterm and 65 on final.
Total points = 130.

Student B gets 65 on midterm and 75 on final.
Total points = 130.
Student A's z-scores are +2.0 on midterm, 0.0 on final = +1.0 z-score average.
Student B's z-scores are 0.0 on midterm, +0.5 on final = +0.25 z-score average.

In this example, even though the means on the two tests were identical, computing grades by adding points results in the scores on the test with the bigger standard deviation being weighted more heavily. In this example, each student did exactly average on one test and well on another. Student A was much further above average on the midterm than B was on the final. However, if grades are determined by adding points, B's performance on the final with its bigger standard deviation gives B the same point total as A.

Converting scores on the midterm and final to letter grades and averaging those will preserve weighting. However, letter grades are less precise than z-scores (e.g., a B- on the midterm above on my typical B- curve would include a range of point scores).

Note though that z-score conversion assumes the teacher finds the average class performance on each exam to be equal. If a teacher thinks the average student did C work on the midterm and B+ work on the final, z-score averaging will not take this into effect. In this situation, assigning letter grades to scores on each test and then averaging may well be more appropriate than using z-scores. (Alternatively, a constant could be added to z-scores on exams where the class did particularly well.)

Appendix III
TORTS
LYNN DAGGETT
FALL 2005
QUIZ RESULTS

	Multiple Choice (worth 3 points each)	Essay I (12 possible points)	Essay II (12 possible points)	Essay total (24 possible points)	Total Exam Points (54 possible points)
Low score	2	2.5	0	3.5	15
High score	10	9	7.5	15.5	43.5
Average score	6.32	5.63	4.06	9.69	28.66
Standard Deviation	1.76	1.21	1.50	2.19	6.26

As I told you earlier, I have high expectations of you, and I write difficult exams where I expect about ½ of the possible points to be the average score. The class exceeded my high expectations on the multiple choice. On the essays, the class almost met my high expectations. Congratulations on the solid start.

Multiple choice

Question	% who got it right	Item discrimination index*
1	70	+.28
2	67	+.64
3	88	+.33
4	75	+.47
5	42	+.67
6	79	+.39
7	32	+.78
8	18	+.25
9	72	+.50
10	89	+.19

* An item discrimination index is a statistic to see how well the item is working by comparing the performance on the item of the overall high-scoring and overall low-scoring students. An index of +0.40 or higher suggests the item is working very well. An index of 0 or less suggests the item needs retooling. It is very difficult to get a high discrimination index for an item which more than 75 percent, or less than 25 percent, of the class answers correctly.

I expected the average score to be about half right (5 of 10). In fact, the average was 6.3 of 10 (the median was 6 and the mode was 7). Good job. The scores ranged from 2 to 10, with a standard deviation of 1.85. That means 2/3 of you were in the 5 to 8 correct range which is the middle pack of the class. Being in the middle pack of a class of bright, hard-working, highly motivated students is no mean feat, so pat yourself on the back. If you got 9 or 10, you did significantly better than average (and an 8 is right at the top of that big middle pack); keep it up! If you got a 2, 3, or 4, that may indicate you need to make adjustments in your approach to studying and/or test-taking (see how you do on the short essays).

Here is the multiple choice score distribution:

[omitted]

Essays

On the essays, most folks showed they have already learned how to think like a lawyer. Most did a good job zeroing in on what to talk about on the first (consent) essay. Many of you had more difficulty sticking to the issue (state of mind) on the second (IIED) essay, telling me about the whole prima facie case rather than the one state of mind element. Most of you were pretty well organized. Most folks seem to know the rules, except that many of you seem to need to do some rethinking of intent. Intent for IIED must be with regard to the P having severe emotional distress; general intent is not enough. Moreover, intent is subjective; if D should have known something that is not determinative. Also some of you need to rethink what is reckless vs. what is intentional. Finally, some of you need to be more thorough in laying out the rules. Remember your job on the exam is to convince me you know the rules, and that means defining them in lay language and including the details. Some folks find it helpful to pretend they are writing to someone who doesn't know the law (e.g., your mom or dad?), and explain things at that level.

As is to be expected only 6 weeks into law school, pretty much everyone still has work to do to get to lawyer-level analysis of how the rules apply to the facts. Some of you are IRC'ing (or even IC'ing): doing the work in your head and writing down your conclusion, but not your thought process. It is the latter which is actually more important.

For example, some essays read like this:

Consent can be implied from one's behavior. Clearly Student's behavior in opening the cage strongly implies consent.

This is missing analysis, with the exception of one fact (opening the cage) thrown in to justify the conclusion, and without explaining the details of all the relevant consent rules. An argument like this to a court would not be very convincing.

Other essays had more analysis, but only the arguments for one side, e.g.

Director had intent because she threw a dead animal in Miscreant's lap at a staff meeting and used her superior power to berate him.

This is better, but still not there–both sides must be looked at and argued. Again, thinking about arguments to a court, when you don't mention the other side's points, the door is wide open for the other side to come in and blast your credibility by listing all the things you conveniently forgot to mention. From another perspective, whether P or D is your client, you need to be able to lay out both the strengths and weaknesses of their case in order to serve them well.

I am handing back the computer printout for your multiple choice, your essays, and the scoring sheet for them. You can keep all of this. If you need to see it, we can also provide access to your computerized answer sheet and the exam questions (many of you did not write your exam numbers on your exam– otherwise we would give that back too; the questions are posted on the public folder and you can dig through the pile to try to find your exam if you'd like). The multiple choice printout will look like this

		Key
		1234567890 (question number)
		1152335343 (right answer A=1, B=2, etc.)
exam #	score	
12834	6	231... 2... (your answers. A dot means you got it right, a number means the wrong answer you chose. Here, e.g., on Q1 the correct answer was A and the student chose B. On Q4 the student got it right).

I don't write a lot of comments on your actual essay answers; instead I use a pretty detailed scoring sheet which hopefully shows you exactly what I was looking for and what I saw and didn't see in your answer. In some cases I also wrote a few general comments at the end.

If you like, I will go over your exam with you. You may want to talk to me about your exam for one or more of several reasons: (1) you think I made a calculation error (I checked the math but it is possible!) (2) you want some general advice about testtaking, (3) you have specific questions about certain items, (4) you think I made a scoring error. I will regrade exams on request, as fairly as I can, but please note I will rescore the entire exam not just the part you think was in error.

This quiz counts 25 percent of your semester grade (or about 15 percent of your yearlong grade). I use z-scores to calculate grades in order to appropriately weight exams with different average scores and standard deviations. Your z-score is the number of standard deviations you are from the mean. (On this quiz = ((your score-28.66)/6.26). For example, a point score of 24 is 0.74 standard deviations below the mean, so its z-score is -0.74. I am not assigning letter grades to this exam, but since this class is subject to the grade curve (average grade I turn in must be between 2.6 and 2.9) it is safe to assume that the average grade of 28.66 is about a B-. I also did not see any exams that deserved a D or an F, so you may also assume the grade curve would run from A to C-. If you scored between 22.5 and 34.5, you are within the 2/3 of the class that was within 1 standard deviation of the mean. If you scored 22 or less, you did significantly below average. THIS DOES NOT MEAN YOU ARE HEADING FOR DISASTER (none of the exams were really awful) but you may want to think about making some study and/or exam-taking adjustments. If you scored 35 or better, congratulations on doing significantly above average.

Below is a list of all the scores and associated z-scores, so you can see exactly where you are. Your total score (circled at the bottom of your essay scoring sheet) is 3 points for each multiple choice question, plus your essay points. You can look at the list below and find the row which corresponds to your score. You can see your z score at the row's end.

A word about grades

Because (1) we have a mandatory grade curve, (2) most of you had quite high grades in college, and (3) the level of competition in law school is therefore much higher than in college, most students' GPAs unfortunately drop between college and law school. I know this is very frustrating, especially since most of you are working harder than you ever have in school. It is easy to fall into the trap of grades becoming some sort of referendum on your worth as a person, which of course they are not, especially when your life is filled with little but legal studies. Look around at how bright, hard working, and motivated your classmates are—being in the same ballpark with them is something to feel good about! Remember too that the main point is to learn the material, and that grades are secondary and at best an imperfect measure of your achievements. Finally, know that while law school grades do make a difference in terms of getting some jobs straight out of law school, the research is clear that grades are not relevant to lawyer success ten years out of school.

Points	N	Z-score
43.5	1	2.37
42	1	2.13
41.5	1	2.05
38.5	3	1.57
38	4	1.49

[remainder of table omitted]