St. John's Law Review

Volume 60
Number 2 *Volume 60, Winter 1986, Number 2*

Article 2

Valid Objective Test Construction

Howard J. Gensler

VALID OBJECTIVE TEST CONSTRUCTION

HOWARD J. GENSLER*

The use of the objective multiple choice examination is on the rise in legal education. Both the Multistate Bar Examination (MBE)[1] and the Multistate Professional Responsibility Examination (MPRE) implement the multiple choice question format. Forty-seven jurisdictions employ the MBE,[2] which has been highly successful as an examination tool.[3] For example, results on the MBE correlate well with results on the essay portion of the New Mexico Bar Examination.[4] Thirty jurisdictions require the MPRE.[5] Furthermore, law schools are increasingly using objective tests in their examinations.[6] Unfortunately, law professors do not have the resources and expertise that are available to the National Conference of Bar Examiners and the Educational Testing Service in constructing objective examinations.[7] The objective law school test is

---

* Dean, Northrop University School of Law; J.D., University of California, Berkeley; M.P.P., University of California, Berkeley; B.A., University of California, Irvine.

[1] Eckler, *The Multistate Bar Examination: Its Origins and Objectives*, 50 B. EXAMINER 15, 18-19 (1981). The MBE Committee, in arriving at its decision to use multiple choice questions, relied on the widespread acceptance of the objective examination in the law school entrance test. *Id.* at 16.

[2] Quade, *Multistate Bar Exam Challenged in Kentucky*, 8 A.B.A. B. LEADER 29, 29 (1983); *see* Bernstein, *Preparation of MBE Questions*, 51 B. EXAMINER 4, 4 (1982). *But cf.* Oliver, *Testing the Bar Exam*, 5 CAL. LAW., June 1985, at 52, 88 (42 states use MBE); Germany, *Report of the Multistate Bar Examination Committee*, 50 B. EXAMINER 27, 28 (48 jurisdictions use MBE).

[3] *See* Klein, *Summary of Research on the Multistate Bar Examination*, 52 B. EXAMINER 10, 15 (1983). *But cf.* Oliver, *supra* note 2, at 88 (many law school deans would eliminate multiple choice section of bar exam).

[4] *See* Brown & Levay, Melendez v. Burciaga: *Revealing the State of the Art in Bar Examinations*, 51 B. EXAMINER 4, 7 (1982); *see also* Covington, *The Preparation and Operation of the Multistate Bar Examination*, 50 B. EXAMINER 21, 23 (1981) (states using MBE have found strong correlation between essay and MBE scores).

[5] Covington, *Multistate Professional Responsibility Examination Statistics*, 54 B. EXAMINER 31, 31 (1985).

[6] *See* Martineau, *Review Essay: Legal Education and Training Artists of the Law*, 57 N.Y.U. L. REV. 346, 349 (1982); *cf.* Eckler, *supra* note 1, at 17.

[7] *Cf.* Hudson & Hudson, *Suggestions on the Construction of Multiple Choice Tests*, 49 AM. J. PHYSICS 838, 838 (1981) (physics teachers do not have resources available to Educational Testing Service in constructing multiple choice tests). The National Conference of

valuable for both evaluating the law student[8] and exposing him to multiple choice questions prior to the bar examination. This Article will provide background and recommendations for understanding and constructing multiple choice tests.

I. MEASUREMENT

The objective test has several advantages as a measurement of learning. Because the objective test question item is short, many items can be employed to cover a great quantity of material in significant detail. A single essay question cannot examine as many diverse areas in as great detail as, for example, forty objective test questions.[9] The mastery of legal concepts and the ability to apply the reasoning of a substantive area can, therefore, best be tested with a well-constructed objective test.

The objective test, of course, cannot test writing proficiency or creativity as well as the essay test.[10] Verbal skills, however, can be demonstrated more appropriately in a writing course.[11] An instructor of an introductory law course should be concerned with ascertaining whether or not the basic substantive material has been mastered. Since writing skills cannot properly be tested in a pressured, timed situation, the objective test cannot be considered inferior because it does not test writing ability or creativity. Finally, objective tests have provided a reasonable basis for identifying the entire spectrum of test takers from low to high achievers, at all levels of education. It can accomplish this result through quantitative variety to test the entire scope of the course, and qualitative variety to identify poor, average, and good students.

II. SCOPE

The greatest strength of an objective examination is its ability to test a great number of narrow areas. Accordingly, scope becomes one of the most important considerations in developing an objec-

---

Bar Examiners, in conjunction with the Educational Testing Service, constructs the MBE and the MPRE.

[8] ASSOCIATION OF AMERICAN LAW SCHOOLS BAR EXAMINATION STUDY PROJECT FINAL REPORT (1976) at 119; *see* Winterbottom, *Use of Essay and Objective Techniques in Bar Examinations*, 38 B. EXAMINER 5, 8 (1969).

[9] *See* Nickles, *Examining and Grading in American Law Schools*, 30 ARK. L. REV. 411, 447-48 & n.121 (1977); Winterbottom, *supra* note 8, at 7-8.

[10] Winterbottom, *supra* note 8, at 7.

[11] *See* Nickles, *supra* note 9, at 444-47.

tive test. The objective test should be planned, with every major area of the course represented in a proportional amount.[12] Within each area, the items should be reasonably distributed among substantive subdivisions. Finally, the professor should ensure that within each area there are both moderate and highly difficult items. If a professor were to put only highly difficult items in one subject area, he would be unable to determine whether the student was simply an average student or whether he had omitted the subject area entirely. For instance, suppose a property professor gave a hundred-item exam and allocated ten items to estates. If all ten items were extremely difficult questions on the Rule in Shelly's Case, the Rule against Perpetuities and the Statute of Uses, even a good student might miss them all despite having a solid foundation in property law generally. As an instrument of measurement the test would have failed because it would not have reflected any aspect of the student's knowledge of property law.

III. DESIGN

It is important to understand the mechanical design and implications of the objective test. The basic objective test presents a problem and then has a number of solutions. The first question is: how many solutions should there be from which to choose? The answer is that it does not matter.[13] Uniformity is not important. A student will not be incapacitated because some of the questions have only three answers while most of the questions have four. More answers reduce the odds of guessing correctly for a student who does not know the answer.[14] If a professor, however, rotely invents two obviously incorrect answers (known as distractors), the odds of choosing correctly will not be reduced. The important aspect, therefore, is that each of the distractors be viable yet defective or inferior to the correct choice.[15]

---

[12] *See* Trieber, *The Use of Multiple-Choice for Testing*, 34 TRAINING & DEV. J., Oct. 1980, at 24, 26.

[13] *But see* Kolstad, Wagner, Kolstad & Miller, *The Failure of Distractors on Complex Multiple-Choice Items to Prevent Guessing*, 8 EDUC. RESEARCH Q. No. 2, at 44, 45 (1983) (three multiple choice items considered optimum for reliability); Straton & Catts, *A Comparison Of Two, Three And Four-Choice Item Tests Given A Final Total Number Of Choices*, 40 EDUC. & PSYCH. MEASUREMENT 357, 364 (1980) (same).

[14] *See* Duncan, *An Appropriate Number of Multiple-Choice Item Alternatives: A Difference of Opinion*, 15 MEASUREMENT & EVALUATION IN GUIDANCE 283, 290-92 (1983).

[15] *Cf.* Bernstein, *supra* note 2, at 8 ("each distractor must contain some plausible element so that it will in fact distract. . . .").

Assuming that the distractors are equally viable, so that the probability of selecting any of the choices is equal, the odds of selecting the correct choice (depending on the number of choices) would be as follows:

Table 1

Test Choice Probabilities

Items	*Odds*	*Marginal Gain*
2	.50	—
3	.33	.17
4	.25	.08
5	.20	.05
6	.17	.03
7	.14	.03
8	.13	.01

Table 1 demonstrates that very little is gained by adding a sixth choice to a test. The probability of guessing correctly is reduced by only three percent. Moreover, it becomes increasingly difficult to write viable distractors. Accordingly, the difficulty of writing a sixth choice will generally outweigh the usefulness that the choice will actually add to the test. Restriction of an item to four or five choices,[16] will maintain a reasonably low probability of guessing correctly without requiring a professor to develop untenable alternatives. Of course, an item is not fatally flawed if only three alternatives are written.

One way to correct for guessing is to deduct points for incorrect answers.[17] This is an important technique to eliminate bias in favor of poor students. If a student is permitted to guess without penalty, a poor student will accrue more undeserved points than a good student because a poor student will have more opportunities to guess. The following table demonstrates the effect of guessing on various students' scores on a hundred-item, four-choice test.

---

[16] *See, e.g.*, Trieber, *supra* note 12, at 28 (best format for multiple choice question is to include one correct, one almost correct, one incorrect and one inapplicable answer).

[17] *Cf.* Hudson & Hudson, *supra* note 7, at 840 (in examinations prepared by ETS, one-quarter of one point deducted for each incorrect answer primarily to give credit to those who narrow down possible answers).

Table 2

The Effect of Guessing on Grades

Known Answers	*Remaining Questions*	*Correct Guesses*	*Final Score*
50	50	13	63
60	40	10	70
70	30	8	78
80	20	5	85
90	10	3	93
100	0	0	100

A student who should have failed (50%) undeservedly improved the final grade earned to a solid D (63%) through guessing. A student who earned a D (60%) improved to a C (70%) through guessing. The "A" students (90% and 100%) were unaffected by guessing (93% and 100%). Accordingly, incorrect answers should be deducted to eliminate the effect of guessing. If a test has four items, one third of a point should be deducted for an incorrect guess. Out of four questions, one right answer and three wrong answers are expected. A student would lose one third of a point three times and gain one point with a net effect of zero. Of course, if a test has items with different numbers of choices, then the penalty must vary with the item, thus:

Table 3

Weight of Incorrect Guesses

Choices	*Penalty*
3	.50
4	.33
5	.25
6	.20

The penalty, of course, equals 1/(x-1) where x is the number of choices.

Penalties do not eliminate guessing, nor should they. If a student knows something about the question and can eliminate some of the distractors, then the student should and will guess. If a student can eliminate two of four choices on each of ten questions, the

student should statistically guess correctly five times and guess wrong five times. The student will earn five points and lose 1.67 points, thereby improving the final score 3.33 points. There is nothing wrong with improving the score in this way. Employing probabilities in this way is a valuable measurement of the student's actual ability. Since the student knew enough of the material to eliminate some wrong answers, he should benefit proportionately for knowing something. The student does not benefit excessively, however, because wrong answers are deducted. Therefore the student benefits only to the extent of his learning, which is the point of the test. That is why it is important not to manufacture by rote unconvincing distractors. If only three good choices can be developed, then only three should be put on the test. A wrong answer will then result in a greater penalty (.50 off instead of .33). If a useless distractor is added, the penalty will be unnecessarily diluted thereby rewarding poor students.

IV. FORMAT

There are several formats in which to structure an objective test item. The basic format is to pose a problem in a short paragraph of approximately 150 words and then to present four solutions from which to choose. The advantage of employing this format is that the item is sufficiently short so that a great many can be used on the test. Consequently, a great number of subject matters can be specifically tested.

A variation on this theme is the long fact problem with a series of questions either on the one set of facts, or dependent on the one set of facts with certain factual modifications or supplements for each question. The advantage to this format is that certain sophisticated problems can be developed that cannot be presented in a short paragraph. The time it takes to digest the long fact pattern is then amortized over a series of questions. The problem with this format, however, is that often the long fact pattern must be reviewed for each or most of the questions. Furthermore, longer fact patterns are more confusing, particularly in a pressured, timed situation. While sometimes long fact patterns cannot be avoided, as a general rule, a series of short items is preferable to a long one.

The next objective question format is the tri-level structure: the first level is the fact problem; the second level is a series of legal conclusions; and the third level is a series of choices that consists of statements as to which combination of the legal conclusions

is true. This format is fundamentally flawed and should not be employed.[18] It is possible for students who know very little to answer these questions correctly while students who know a fair amount to fail. For instance, suppose that after a legal fact problem five legal conclusions follow: I, II, III, IV, and V. The four alternatives (A, B, C, and D) from which to choose are as follows:

A. I and II are true.
B. I and IV are true.
C. II and IV are true.
D. II, IV, and V are true.

Now assume that I, II, and V are true. Suppose further that the student knows that I and V are true and that III is false, but has no opinion on II and IV. The student who knows sixty percent of the substantive content of the question is unable to eliminate any of the four choices. Another student who only knows that IV is false is able to answer the question correctly. Consequently, this structure of examination should not be used.

V. SCORING

A great advantage of the objective test is that it is mechanically easy to score.[19] The problem, however, with scoring an objective exam is that a wrong answer from a poor student looks the same as a random response to an ambiguous or flawed question. Nevertheless there is a relatively simple process that can be employed to identify ambiguous or suspect questions. First, score all the exams; next, rank order the exams from highest to lowest and separate them into four quarters; finally, tally the number of correct responses to each question by quarters. A good test item will have more correct responses in the higher quarters than in the lower quarters. A good item does not mean that only a few students answered it correctly. A good item is an item that was answered correctly more often by the better students (higher quarter) than the poorer students (lower quarter). An ambiguous or flawed question will have a random response pattern (since it cannot be correctly answered because of the flaw), which will result in a lack

[18] *See* Kolstad, Briggs, Bryant & Kolstad, *Complex Multiple-Choice Items Fail To Measure Achievement*, 17 J. OF RESEARCH & DEV. IN EDUC., Fall 1983, at 7, 8-10.

[19] Trieber, *supra* note 12, at 25; Winterbottom, *supra* note 8, at 8; *see* Nickles, *supra* note 9, at 451.

of an appreciable difference between the quarters. In other words, poor students will do as well as good students because the answer does not depend on understanding the material.[20]

For instance, suppose a class of one hundred students takes an objective test. The exams are divided into ranked quarters and the following correct responses are given:

Table 4

Exemplary Test Response Distributions

Quarter:	*1st*	*2nd*	*3rd*	*4th*
Question 1	22	18	15	10
Question 2	7	2	0	1
Question 3	24	21	22	21
Question 4	15	7	6	7
Question 6	7	6	6	5

The first question provides the expected result of a moderate question. The good students answered correctly. Less students answered correctly as the overall test score declines. This item is a good one in that it helps separate the students into their respective classifications.

The second question gives the results of a difficult question. Only a few of the top students correctly answered the question. This was probably a situation in which the class picked what looked like the obvious answer, but some exception to the rule or technicality applied, making another response the correct answer. Only a few of the best students were aware of this.

The third question was an easy one. Only a few students missed it, therefore it did not help to separate the students into their respective classifications very well. However, every question cannot separate the class into the proper spectrum and this one may have tested a basic area that needed to be represented on the exam. Although a better question may need to be developed in time, this question need not be stricken from the current exam.

The results of the fourth question indicate that there may be something wrong with the question. Although the best students answered the question correctly more often than the rest of the class,

---

[20] *Cf.* Klein, *supra* note 3, at 15 (well structured items are reliable and unaffected by random factors such as guessing).

the separation ends there. The next three quarters provided random results. This indicates that it is a very difficult question that only about ten good students answered correctly and at which everyone else guessed, or that the question was flawed and the better students were able either to deal with the flaw or to second guess the professor's idiosyncrasies. This question should be carefully reviewed by several other professors and students.

The fifth question is classically poor. A random response was given, indicating that the question is either too difficult to use or seriously ambiguous. The question should be removed from consideration.

These are a few of the basic response patterns that can be expected. Any response pattern that is not weighted more heavily with the higher scores should be suspect. If the answers follow the pattern in question five, the question should be stricken from the exam.

Once the flawed questions are identified and stricken, the exams must be rescored excluding the stricken questions. Then the incorrect responses must be deducted (according to their weight) from each score to arrive at the final score. For example, suppose a student scored eighty-two points on a hundred point test, but four of the items were stricken for ambiguity. Of the four items, this student had two right answers. The student's score would be reduced to eighty. Of the eighteen remaining questions, suppose twelve were answered incorrectly and six were left blank. Of the twelve wrong answers, ten had four choices and two had three choices. The student's score should be reduced by 4.33 points (10 x .33 plus 2 x .50). The student's final score is 75.67 out of a possible 96 (100 - 4 ambiguous questions) or 79%.

Grades must then be assigned to the final scores.[21] There are two basic methods of grading: a straight objective scale or a relative scale. A straight objective scale is the usual 90 - 100% = A, 80 - 90% = B, 70 - 80% = C and so on. A relative scale matches grades to the distribution of class exams. For example, suppose forty-two students earned the following points on a seventy point test:

---

[21] *See, e.g., id* at 11 (MBE scores are converted to scale scores).

Table 5

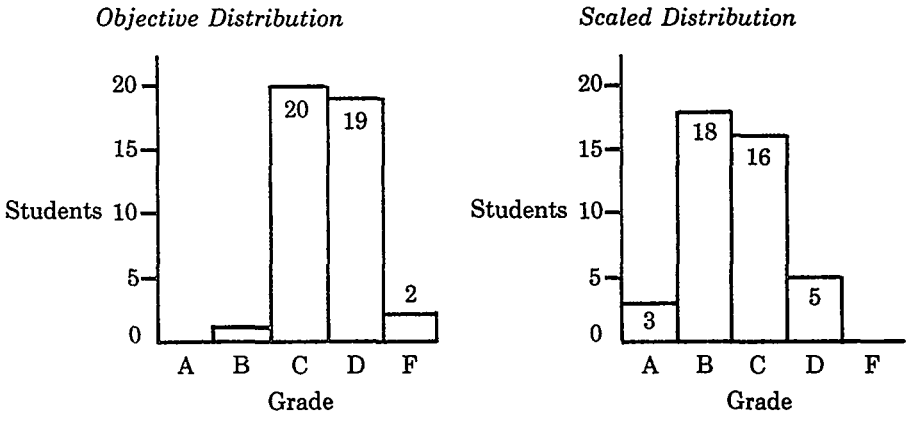
Exemplary Test Scores Distribution

Score	*Number of Students*	*Percentage Score*	*Objective Grade*	*Scaled Grade*	*Relative Scale*
56	1	80	B−	A	A 53-56
54	2	77	C+	A−	B 49-52
52	4	74	C	B+	C 45-48
50	7	71	C−	B	D 41-44
49	7	70	C−	B−	
48	6	69	D+	C+	
47	5	66	D	C	
45	5	63	D−	C−	
43	3	60	D−	D	
41	2	57	F+	D−	

Employment of a strictly numerical percentage scale would result in no A's and generally low grades. This may be appropriate if the professor believed the test was both fair and not unduly difficult. However, since the purpose of an examination is to spread the students over a spectrum, the better policy is to rank the performances to a grading scale tailored to the exam and the performance. In the above table, the distribution of grades has a sixteen point range: three students are in the top quarter, eighteen in the second, sixteen in the third and five in the bottom. Minimum standards should always apply. It is not inconceivable that an entire class could do poorly. However, assuming the basic competency of both the students and the professor, the relative scale set forth in the last column provides a more descriptive analysis of the class performance. Performance and evaluation are more closely aligned with the scaled grades than with the objective grades.

Graph 1

Exemplary Objective/Relative Grade Distribution Comparison



VI. Conclusion

The objective test is a valuable instrument for measuring students' mastery of substantive knowledge. More material can be covered with greater precision than by other test formats. Objective tests can also be utilized to distribute students over the entire spectrum of academic competency, because such tests not only reflect knowledge of correct answers, but also proportionately measure partial comprehension of a test item by discounting for incorrect responses. Discounting removes from objective tests the bias that would otherwise benefit poor students. Tri-level objective test formating prevents proper assessment of student comprehension and should be avoided. In assigning final grades to an objective test final point score, a relative scale, matched to the actual score distribution, will be more descriptive and useful in meaningfully assessing student performance.