Dear Professor: Why Do I Ace Essay Exams but Bomb Multiple Choice Ones?

Herbert T. Krimmel

This is not a scientific study and does not purport to be one.[1] I did not set about to investigate what causes certain good students to experience problems with multiple choice exams. Indeed, I do not really think I would have found the question all that interesting had it not ambushed me.

I like my grades to be tidy, and it bothers me when they are not. I am well aware that a perfect statistical correlation between the different parts of an exam is unachievable and probably undesirable. After all, if two testing vehicles infallibly gave the same results,[2] they would be redundant. And, besides, I think different parts of an exam ought to test different skills,[3] as opposed to just different knowledge. It is, therefore, conceivable, expected even, that a student could perform better on one part of an exam that tests one aspect of lawyering than on another part of the exam that tests other lawyering

Herbert T. Krimmel is Professor of Law, Southwestern Law School. I wish to thank my colleagues Professors Michael Dorff, James Fischer, Warren Grimes, and Danielle Hart who generously took the time to read and comment on the various drafts of this article.

[1] Throughout this article I have generally refrained from citing exact percentages in reporting the results of my study. I have done this intentionally because I do not want to give the misimpression that my study's findings imply a level of precision or exactness that is both unclaimed and unwarranted due to the relatively small number of participating students. I conducted my study over two decades with just shy of 50 students, who qualified after the winnowing process described in this paper. That small of a sample size is just too little to provide any more than an inventory (and perhaps a partial one, at that) of students' multiple choice exam-taking problems. The purpose of my study was to determine *if* I could find a way to identify and diagnose these problems, not to determine their relative frequency. To do that would require a much larger study.

[2] Of course they never do, and a primary reason for using a battery of comprehensive multiple choice questions as part of a final exam is to reduce the element of random luck that results from such things as: a student's extra-legal familiarity with an issue; his absences and illnesses during periods when a given subject matter is taught; and the relative emphasis that he gives a subject in allocating study time. These all can pose a problem for essay exams, which, of necessity, sample a narrower and more selective slice of a course's coverage.

[3] I like to give essays questions with fewer issues and liberal allotments of time but which require the student to go to much greater depths of analysis, and then use the multiple choice questions to provide broad doctrinal coverage of the course

skills. Multiple choice questions are excellent for testing mastery of the rules, but essay questions also assess and reward other important lawyering skills such as rhetorical ability, creativity, composition and organization.

Nevertheless, I kept finding myself disappointed when I assembled the various parts of my exam and discovered the occasional situation where a student wrote an essay scoring in the top quarter but then ranked in the bottom quarter on the multiple choice section of the exam. Initially, I assumed this was a fluke. But I also had the vague fear that somehow this might be my fault, either in grading or exam construction. I would reread the essay thinking maybe the student's answer was not as good as my initial impression of it. But this was never the case. The points I awarded these essays on second reading always were close to my initial evaluation. They were well within an acceptable margin of error; the difference between the two readings never constituted more than one-third of a grade point and usually less.

Worse was the fear that maybe my multiple choice questions were sabotaging good students. But this did not seem to be the case, either. The multiple choice exam evaluated on its own had a sound coefficient of reliability (KR20).[4] And a scatter plot of the multiple choice and essay scores of my exams showed that, in general, they correlated well.[5] The remaining explanation, therefore, was that the disparate essay and multiple choice grades of these students resulted from their differing ability to perform based on the testing vehicle.[6]

I am not sure why, but the converse situation never bothers me as much. When a student excels on the multiple choice part of the exam but "bombs" the essay, I attribute this to a "writing problem." And it usually is. The student's essay shows spots of creativity but is "conclusory," omits rule statements, devises some grand architecture that he cannot complete (for every hand, there is another hand even if it involves questioning the given, ignoring facts or assuming the court will change the rule or make new exceptions to it, for God knows what reason). I am disinclined to give such students much benefit of doubt, perhaps because reading poorly written essays answers is so personally annoying. For these students, it is their good multiple choice scores that look aberrant—something I would like to chalk up to good luck, or, when I am on a less personal rant, something I blame on their undergraduate universities: "They are not doing their job ...These students are bright but they cannot express what they know in writing . . . ."

[4] Described *infra* at footnote 17.

[5] Occasionally I have run a linear regression of the essay and multiple choice scores, but for the most part I prefer the visual aspects of scatterplots. I like being able to see the correlation.

[6] The essay and multiple choice portions of my exam are separately timed. A student, therefore, cannot borrow time from one exam section for the other, by, say, skimping on the multiple choice questions to spend more time on the essay. Consequently, because it simply cannot be done given the exam structure, a different allocation of time cannot explain any discrepancy between a student's relative performances on the two separate portions of the exam.

But, I expect it is not so. Knowing something, and being able to express it, is not the same. And neither is the ability to recognize the applicable legal rule when it presents itself in a structured array the same skill as the ability to summon it forth from a body of facts. The answer to a multiple choice question is quite literally on the page; the answer to an essay question is in the student's mind waiting to be born. It is well known that different testing vehicles can produce different results. Even something as mundane as a spelling test can be constructed in several different ways, and at least for some students, their performance can be strongly affected based on the testing vehicle. Some students apparently will often perform quite differently when asked to spell a word versus identify which words in a list are misspelled, versus find the misspelled words in a document, versus identify which of two spellings of a word is correct.

Twenty years ago[7], I started to make a point of writing "see me"[8] on the exams of students whose essay and multiple choice scores diverged widely, and, by and large, most came to confer. What greatly surprised me was that almost without exception they told a similar story. Rather than blame the exam, quite to the contrary, they told me that they never did well on objective tests. They were frustrated, embarrassed, and fearful they would never be able to pass the bar exam. They wanted to know the "secret" of taking law school multiple choice exams. Of course I told them there was no "secret." The "secret" was diligent preparation, careful reading of the question, etc. I think what I told them was right, but even at the time, my words seemed a bit hollow to me: right words, but not very helpful ones. After all, these were good students. I believed they had in good faith done the work. For the most part they had participated in class. Their essays demonstrated that they could do B+ work or better. Indeed, there was no secret to be imparted, but there was a problem to be overcome. In working with these students certain patterns started to emerge. Certain problems with taking multiple choice exams began to be evident. This article is about what I have learned from working with these students.

I. On Establishing a Method for Evaluating Differential Performance on Essay and Objective Testing Vehicles

My paramount problem was to see if it was possible to develop methods to diagnose what these students were doing wrong. Could a technique be devised so that I and others could understand precisely what these students were doing differently that caused them to be flummoxed by multiple choice exams and to explain why their actions confounded them? Equally important to me was to learn whether these exam-taking problems had any consistently appearing special or distinguishing hallmarks by which they could be identified.

[7] I commenced this study in May, 1991, with my Wills and Trusts exam. In the ensuing 20 years, I have employed the techniques described in this article in analyzing the exams of both my Wills and Trusts classes and my first-year Torts classes.

[8] Like most law schools, the school I teach at uses a blind grading system.

Students perform poorly on multiple choice exams for a multiplicity of reasons: lack of understanding of the doctrines being tested, not being "test wise," poor exam construction, to name just a few. It was necessary to find a way to separate out these different strands. I needed to identify a subgroup of those who performed poorly on multiple choice exams for whom no other explanations of poor performance other than a "test taking problem," could be true. Of necessity this would be a small group but one in which the signal might be clearly distinguishable from the background noise, if it was to be distinguishable at all.

As the first step in this winnowing process, I limited my study to only those students whose performance on the essay and multiple choice portions of the exam was widely divergent. Typically around 4 percent of the students on any given exam qualified under this criterion. All but a handful of the students I have worked with[9] had an essay score that ranked them in the top quarter of that portion of the exam, and a multiple choice score that put them in the bottom quarter of that portion of the exam. I do not doubt that the problems of objective exam taking that I describe in this article affect a much larger group of students than these. But the closer that students' subjective and objective scores correlate to and corroborate with one another: 1) the more likely it is that the nature of the exam vehicle is not a significant factor in their performance; 2) the more likely it is that any exam taking problem that students experience will be masked by other and greater problems they may have, such as a lack of mastery of the rules; and 3) the more difficult it will be to be able to isolate, identify or diagnose their problem. In other words, while I strongly believe that the problems of multiple choice exam taking that I identify in this article in fact affect a rather broad range of students, and that these problems are in no way limited to just the extreme cases of divergence that I chose for my study, for most students any "exam taking problems" they may have will be engulfed by bigger issues. It was necessary to limit my study to those cases where there was an extreme divergence between a student's essay and multiple choice performances in order to have any hope of being able to isolate and identify what these objective test taking problems were.

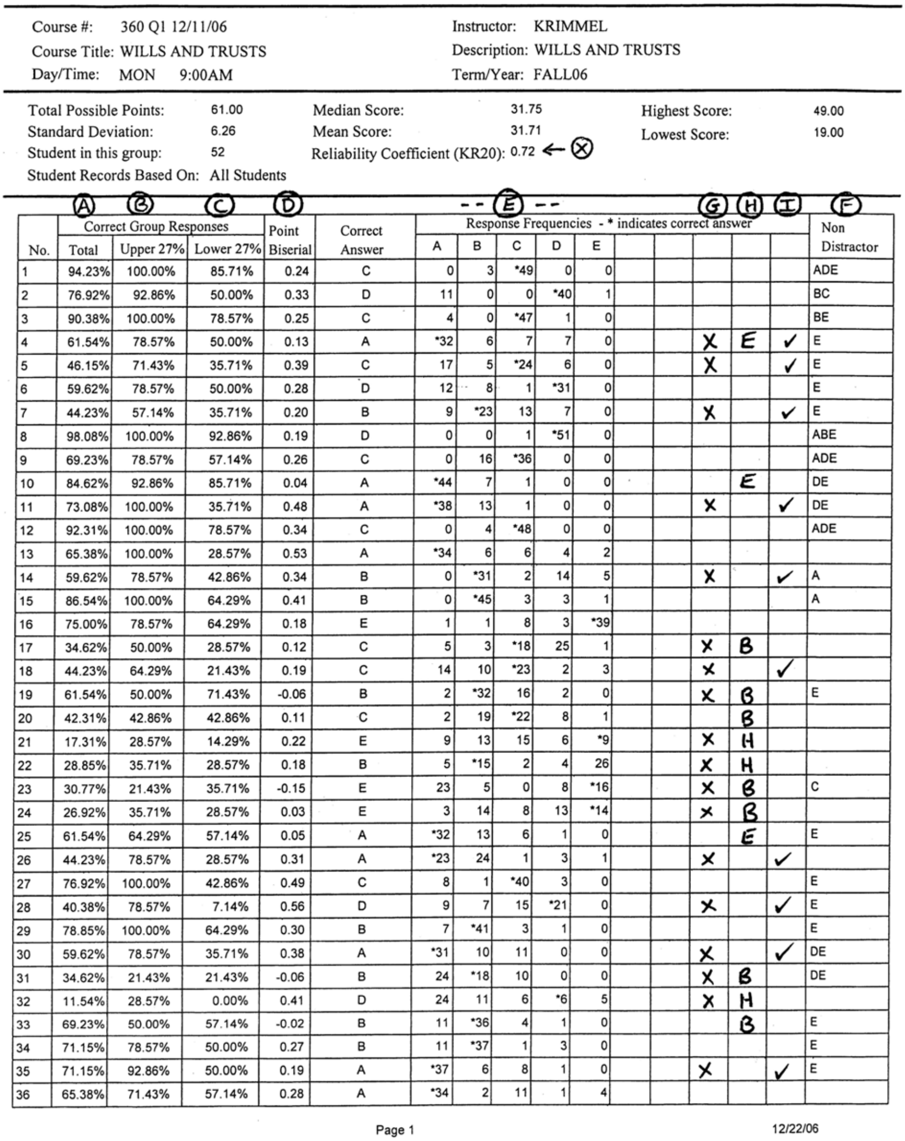
Software provided a critical and necessary element in diagnosing these problems, allowing for analysis of how groups of students perform on each question of a multiple choice exam. The software I have used from the inception of this study is ParSCORE (currently version 6.1) produced by SCANTRON Corporation.[10] The "Standard Item Analysis Report" is this software's basic

[9] One of the difficulties in doing a study such as this, and one reason I have waited 20 years to report my conclusions, is that on any given exam, I can cull only a very small number of students who meet the participation criteria. Typically on a given exam, two or three students qualify but I never had more than four and sometimes none. Fortunately, the students who qualified generally were eager to participate in my study as they already were frustrated by their poor performance on objective exams.

[10] ParSCORE 6.1, ParSYSTEM Software, ParSCORE User's Guide (P/N 908099-001, Rev E) (January 2003); SCANTRON Corporation [hereinafter cited as ParSCORE User's Guide].

tool for accessing the effectiveness, validity and reliability of multiple choice test questions.[11]

Table 1: Standard Item Analysis Report on Exam 1 Version A



Course #: 360 Q1 12/11/06
Course Title: WILLS AND TRUSTS
Day/Time: MON 9:00AM

Instructor: KRIMMEL
Description: WILLS AND TRUSTS
Term/Year: FALL06

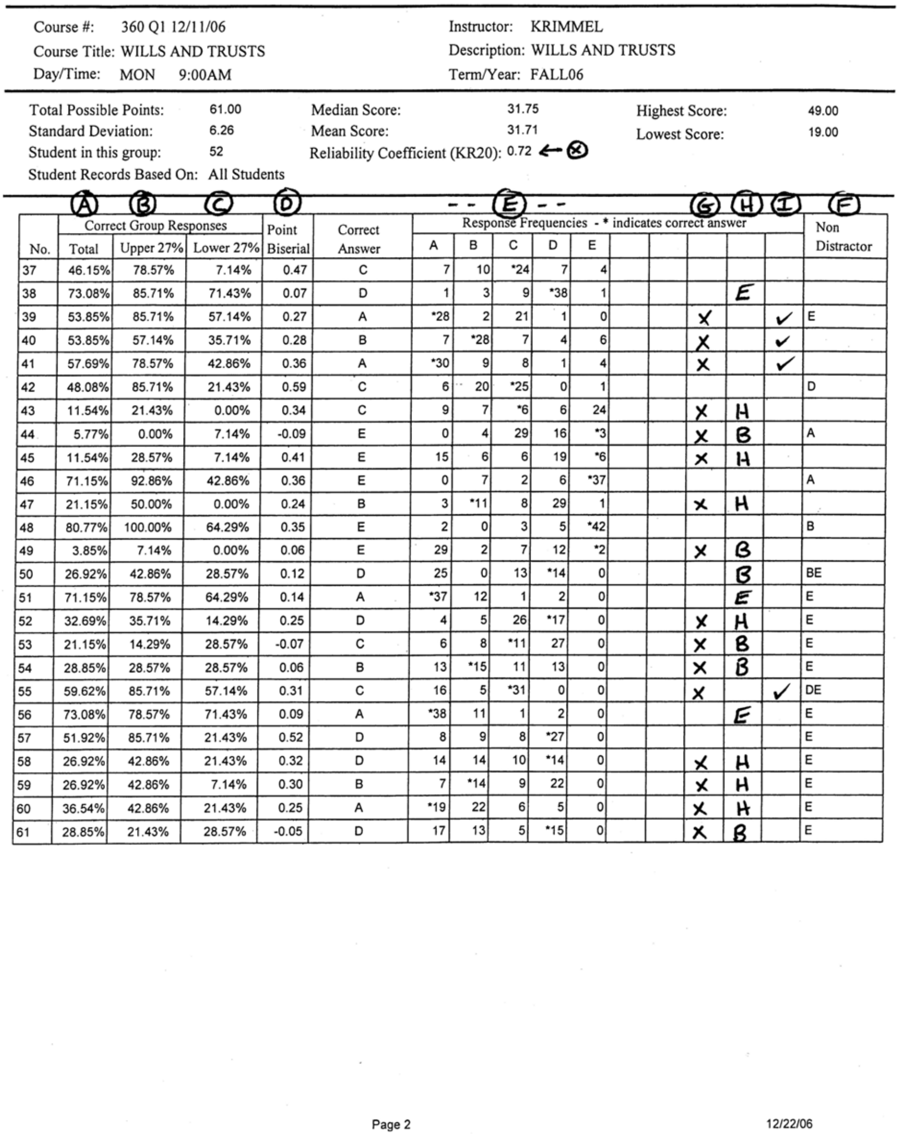
Total Possible Points:	61.00	Median Score:	31.75	Highest Score:	49.00
Standard Deviation:	6.26	Mean Score:	31.71	Lowest Score:	19.00
Student in this group:	52	Reliability Coefficient (KR20):	0.72 ← ⊗		
Student Records Based On:	All Students				

	Ⓐ	Ⓑ	Ⓒ	Ⓓ		-- Ⓔ --							Ⓖ	Ⓗ	Ⓘ	Ⓕ
	Correct Group Responses			Point	Correct	Response Frequencies - * indicates correct answer										Non
No.	Total	Upper 27%	Lower 27%	Biserial	Answer	A	B	C	D	E						Distractor
1	94.23%	100.00%	85.71%	0.24	C	0	3	*49	0	0						ADE
2	76.92%	92.86%	50.00%	0.33	D	11	0	0	*40	1						BC
3	90.38%	100.00%	78.57%	0.25	C	4	0	*47	1	0						BE
4	61.54%	78.57%	50.00%	0.13	A	*32	6	7	7	0			X	E	✓	E
5	46.15%	71.43%	35.71%	0.39	C	17	5	*24	6	0			X		✓	E
6	59.62%	78.57%	50.00%	0.28	D	12	8	1	*31	0						E
7	44.23%	57.14%	35.71%	0.20	B	9	*23	13	7	0			X		✓	E
8	98.08%	100.00%	92.86%	0.19	D	0	0	1	*51	0						ABE
9	69.23%	78.57%	57.14%	0.26	C	0	16	*36	0	0						ADE
10	84.62%	92.86%	85.71%	0.04	A	*44	7	1	0	0				E		DE
11	73.08%	100.00%	35.71%	0.48	A	*38	13	1	0	0			X		✓	DE
12	92.31%	100.00%	78.57%	0.34	C	0	4	*48	0	0						ADE
13	65.38%	100.00%	28.57%	0.53	A	*34	6	6	4	2						
14	59.62%	78.57%	42.86%	0.34	B	0	*31	2	14	5			X		✓	A
15	86.54%	100.00%	64.29%	0.41	B	0	*45	3	3	1						A
16	75.00%	78.57%	64.29%	0.18	E	1	1	8	3	*39						
17	34.62%	50.00%	28.57%	0.12	C	5	3	*18	25	1			X	B		
18	44.23%	64.29%	21.43%	0.19	C	14	10	*23	2	3			X		✓	
19	61.54%	50.00%	71.43%	-0.06	B	2	*32	16	2	0			X	B		E
20	42.31%	42.86%	42.86%	0.11	C	2	19	*22	8	1				B		
21	17.31%	28.57%	14.29%	0.22	E	9	13	15	6	*9			X	H		
22	28.85%	35.71%	28.57%	0.18	B	5	*15	2	4	26			X	H		
23	30.77%	21.43%	35.71%	-0.15	E	23	5	0	8	*16			X	B		C
24	26.92%	35.71%	28.57%	0.03	E	3	14	8	13	*14			X	B		
25	61.54%	64.29%	57.14%	0.05	A	*32	13	6	1	0				E		E
26	44.23%	78.57%	28.57%	0.31	A	*23	24	1	3	1			X		✓	
27	76.92%	100.00%	42.86%	0.49	C	8	1	*40	3	0						E
28	40.38%	78.57%	7.14%	0.56	D	9	7	15	*21	0			X		✓	E
29	78.85%	100.00%	64.29%	0.30	B	7	*41	3	1	0						E
30	59.62%	78.57%	35.71%	0.38	A	*31	10	11	0	0			X		✓	DE
31	34.62%	21.43%	21.43%	-0.06	B	24	*18	10	0	0			X	B		DE
32	11.54%	28.57%	0.00%	0.41	D	24	11	6	*6	5			X	H		
33	69.23%	50.00%	57.14%	-0.02	B	11	*36	4	1	0				B		E
34	71.15%	78.57%	50.00%	0.27	B	11	*37	1	3	0						E
35	71.15%	92.86%	50.00%	0.19	A	*37	6	8	1	0			X		✓	E
36	65.38%	71.43%	57.14%	0.28	A	*34	2	11	1	4						

Page 1 12/22/06

[11] This software also has available an "Enhanced Item Analysis" report providing more statistical information about the exam (i.e. variance, mode, skewness, kurtosis, standard error of measurement) and expands the analysis of the answer choices (the distractors), providing for each a point-biserial coefficient. In other words, for each question it allows you to see how well each wrong answer choice did in discriminating between the good and the poor students. For those who are interested, the software's user's manual provides the technical details. ParSCORE User's Guide, *supra* note 10, at D-11.

Table 1: Standard Item Analysis Report on Exam 1 Version A (Cont'd.)



Course #: 360 Q1 12/11/06
Course Title: WILLS AND TRUSTS
Day/Time: MON 9:00AM

Instructor: KRIMMEL
Description: WILLS AND TRUSTS
Term/Year: FALL06

Total Possible Points: 61.00
Standard Deviation: 6.26
Student in this group: 52
Student Records Based On: All Students

Median Score: 31.75
Mean Score: 31.71
Reliability Coefficient (KR20): 0.72 ← ⊗

Highest Score: 49.00
Lowest Score: 19.00

	A	B	C	D				-- E --					G	H	I	F
	Correct Group Responses			Point	Correct	Response Frequencies - * indicates correct answer										Non
No.	Total	Upper 27%	Lower 27%	Biserial	Answer	A	B	C	D	E						Distractor
37	46.15%	78.57%	7.14%	0.47	C	7	10	*24	7	4						
38	73.08%	85.71%	71.43%	0.07	D	1	3	9	*38	1				E		
39	53.85%	85.71%	57.14%	0.27	A	*28	2	21	1	0			X		✓	E
40	53.85%	57.14%	35.71%	0.28	B	7	*28	7	4	6			X		✓	
41	57.69%	78.57%	42.86%	0.36	A	*30	9	8	1	4			X		✓	
42	48.08%	85.71%	21.43%	0.59	C	6	20	*25	0	1						D
43	11.54%	21.43%	0.00%	0.34	C	9	7	*6	6	24			X	H		
44	5.77%	0.00%	7.14%	-0.09	E	0	4	29	16	*3			X	B		A
45	11.54%	28.57%	7.14%	0.41	E	15	6	6	19	*6			X	H		
46	71.15%	92.86%	42.86%	0.36	E	0	7	2	6	*37						A
47	21.15%	50.00%	0.00%	0.24	B	3	*11	8	29	1			X	H		
48	80.77%	100.00%	64.29%	0.35	E	2	0	3	5	*42						B
49	3.85%	7.14%	0.00%	0.06	E	29	2	7	12	*2			X	B		
50	26.92%	42.86%	28.57%	0.12	D	25	0	13	*14	0				B		BE
51	71.15%	78.57%	64.29%	0.14	A	*37	12	1	2	0				E		E
52	32.69%	35.71%	14.29%	0.25	D	4	5	26	*17	0			X	H		E
53	21.15%	14.29%	28.57%	-0.07	C	6	8	*11	27	0			X	B		E
54	28.85%	28.57%	28.57%	0.06	B	13	*15	11	13	0			X	B		E
55	59.62%	85.71%	57.14%	0.31	C	16	5	*31	0	0			X		✓	DE
56	73.08%	78.57%	71.43%	0.09	A	*38	11	1	2	0				E		E
57	51.92%	85.71%	21.43%	0.52	D	8	9	8	*27	0						E
58	26.92%	42.86%	21.43%	0.32	D	14	14	10	*14	0			X	H		E
59	26.92%	42.86%	7.14%	0.30	B	7	*14	9	22	0			X	H		E
60	36.54%	42.86%	21.43%	0.25	A	*19	22	6	5	0			X	H		E
61	28.85%	21.43%	28.57%	-0.05	D	17	13	5	*15	0			X	B		E

Page 2 12/22/06

Table I represents an analysis of one of my Wills and Trusts exams.[12] At this juncture I want to point out important features of this output to which I will refer in this article. The second column of Table I (marked "A") represents the percentage of students who answered correctly a given question. This number serves as an index of difficulty for a given question. For example, Question 3 was relatively easy: 90.38 percent of the students taking the exam answered it

[12] Fall 2006, Wills and Trusts.

correctly. Question 21, however, was hard, only 17.31 percent of the class got it right.

The next two columns of Table I (labeled "B" and "C") show respectively how the students whose score on just the multiple choice part of the exam placed them in either the bottom or top quarter[13] of the class, performed on this particular question. Referring to Question 28, for example, 40.38 percent of the class as a whole answered it correctly. But the students whose overall score on the multiple choice part of the exam placed them in the top quarter answered Question 28 correctly 78.57 percent of the time. In contrast, students whose overall multiple choice score placed them in the bottom quarter answered the question correctly 7.14 percent of the time (not even as good as to be expected from guessing, assuming five answer choices).

The next column (labeled "D") is a statistical product[14] called a point biserial. It represents how well a given question performs in discriminating between students who are doing well and those who are doing poorly on the multiple choice portion of the exam.[15] To capture this concept, some examples are necessary. Look at Questions 26, 29, 55 and 59. Although they ranged in difficulty from hard (26.92 percent correctly answered) to moderately easy (78.85 percent correctly answered), each has a nearly identical point biserial of .30 to .31. Each of these questions does a nice job of discriminating between those students in the top and bottom quarters of the multiple choice portion of the exam. Now compare Question 8, which is so easy that even 92.86 percent of students in the bottom quarter got it right. Now there are other good reasons for including such questions on an exam,[16] but for evaluation purposes they

[13] Actually top and bottom 27 percent, but for ease of expression throughout this article, I refer to this as the top and bottom quarters.

[14] $$PBCC = \frac{(Mp - Mq)\sqrt{Np \times Nq}}{N \times \sigma}$$

Where:
Mp = Mean total score of the group of students that got the item correct.
Mq = Mean total score of the group of students that got the item incorrect.
N = Total number of students.
Np = Total number of students who got the item correct.
Nq = Total number of students who got the item incorrect.
σ = Standard deviation of scores on the whole test.
ParSCORE User's Guide, *supra* note 10, at 10-12.

[15] Test items with a point biserial of .30 and above are considered to be excellent questions. Those between .20 and .29 are considered to be good questions. Those between .09 and .19 are considered marginal questions. And test items with a point biserial below .09 are poorly performing questions and should be either rewritten or replaced. ParSCORE User's Guide, *supra* note 10, at 10-15.

[16] Psychologically it is useful to start an exam with easy questions. But I also find them a useful test of good faith. I tell my Wills and Trusts students that if they master my intestacy and advancement homework problems, they will be rewarded by finding similar questions on the final exam. I have no qualms about giving a poor grade to a student who fails that part of the exam.

might just as well not have been there: 51 of 52 students answered Question 8 correctly, consequently the question only mattered for one student. Question 61 shows the flip side of this phenomenon: 28.85 percent of the class as a whole answered it correctly. But it appears that nearly everyone, good and poor students, guessed at it. In fact the bottom quarter was marginally better at guessing than the top quarter, which accounts for the negative point biserial. Obviously, I will either want to discard such a question or rewrite it before I use it on a future exam.

Now the reliability coefficient (KR20) that I have labeled "X" on Table I is a measure of internal consistency. It measures how well correct/incorrect answers on each question predict correct/incorrect answers on the other questions of the same exam.[17] The higher the number, the more reliable the exam is. I find that when I first use a new exam, I get KR20 scores in the .4-.5 range, and that after several years of revisions I can move up to the .7-.8 range.[18]

The columns labeled "E" on Table I, represent the frequency that each answer choice was selected by all the students taking the exam. The answer choice with the * next to it is the correct answer. Note that as a class, with some exceptions, the correct answer usually is the single most popular one; this holds true even for questions where a majority of the class did not answer the question correctly. This is done by design.[19] For example, Question 18 was

[17] The Kuder-Richardson Formula 20 is:

$$KR_{20} = \frac{N_i}{N_i - 1} \times \frac{S^2 - (\Sigma PQ)}{S^2}$$

Where:
N_i = Number of items
N = Number of students taking the test
S = Variance
P = $\frac{\text{Number of people responding correctly to a test item}}{N}$
Q = 1.0 - P
ParSCORE User's Guide, *supra* note 10, at 10-12.

[18] According to the software's user's manual we should strive to craft a multiple choice test that will produce a KR20 of .70 or better. ParSCORE User's Guide, *supra* note 10, at 10-12. Nevertheless, it has been my experience that it is much harder to get a high KR20 with my first-year Torts students than with the second- and third-year students in my Wills and Trusts class.

[19] When writing a new multiple choice question there is, of course, no guarantee as to how well it will perform. But, I have two guidelines concerning whether to retain a given question for use on future exams. First, subject to the exception discussed at the end of this footnote, the question must have a point biserial of ≥ .15. Second, a question's correct answer must have been chosen by, if not a majority, at least a plurality, of the exam takers. (See, for examples, Question 14—a majority chose the correct answer B; and Question 28—a plurality chose correct answer D). It is my goal to have a multiple choice exam, where, if the class as a whole were to have to vote on a common answer for each question, and submitted one Scantron that represented the answer on each question that received the most votes, that Scantron would have a score of 85 to 90 percent correct. I do retain some questions that do not meet this majority/plurality criterion, if they have a high point biserial (e.g. Questions 32 and 45). Because these represent hard questions that discriminate well between the top

a moderately difficult question. Only 44.23 percent of the class got it right, but the right answer, in this case "C," was chosen by more students than any of the wrong answers.

The column at the far right of Table I, labeled "F," represents answer choices that were "non-distractors." In other words, these are wrong answers that no student taking the exam chose. Obviously if one mixes four and five choice answers, as I did on this exam, answer "E" sometimes will be listed in column F, simply because no student will choose a non-answer. But, apart from this, for example on both Questions 8 and 9, answer "A" was a legitimate choice, but none of the students taking this exam "fell for it." If I decide to use these questions on a future exam I will try to provide better wrong answers.[20] No point in using bad bait.

The biggest benefit I find from the data in Column E , however, is the information it gives me about what part of the rule the students did not know. I try to structure my multiple choice questions so that each answer choice tests some facet of a rule or exception to a rule. I also try to have a hierarchy of choices so that for a given question there will be one answer, which although wrong, requires a complete understanding of the rule to know this and other choices that are wrong for more obvious reasons. On Question 24, for example, answers "D" and "B" both were wrong but it required an exacting analysis by students to determine this. Answers "A" and "C" on the same question also both were wrong but for more obvious reasons, though I still got some takers. Structuring the exam choices in this way also provides me information about what facet of a rule the students are missing. It gives me a reading on what percentage of the students had a working grasp of a given concept by how many avoided choosing a wrong answer that tested their understanding of those rules.

II. On Diagnosing the Pathologies of Multiple Choice Exam Taking

First, I reread the essay portion of the exam on which the student performed well. This necessary step may provide clues about poor performance on the multiple choice section. I essentially seek to determine where and how this student earns his points. I look to see if her statements of the rules are detailed and precise. For, it sometimes happens that a bright student who

and the rest of the class. I think there should be at least some questions on the exam that only the A students are expected to know the answers to. But these should be the exceptions rather than the rule, and questions that the majority of the class can answer correctly should strongly predominate. Correspondingly, I also retain some easy questions (e.g. Questions 4, 10, 25, 38, 51 and 56) even though they have a point biserial < .15, because, here the lower point biserial is merely an artifact of them being easy questions as opposed to it being an indication of their unreliability.

[20] Sometimes this is difficult. Questions 2 and 6 on this exam are both intestacy questions. Unfortunately for Question 2, there really is only one good wrong answer: "A." Both "B" and "C" mathematically are possible but nothing a student who was paying attention would pick. On Question 6, I had the luxury of two good wrong answers: "A" and "B," and the one student who picked "C" was either out to lunch or marked his Scantron wrong.

writes well can compensate for her less than complete understanding of a rule by focusing her efforts and doing well on the parts of the essay exam that she fully understands. I am especially likely to see this, say, in students who do better in policy-oriented classes than in those that demand strong and exacting emphasis on mastering a complex hierarchy of rules.[21] Consequently, some students perform chronically poorly on multiple choice exams, simply because they lack depth in their understanding of the rules. This becomes apparent when I review the essays.

The next step in the diagnosis is to look at the student's Scantron and the data the computer provides me on how the class as a whole performed on the exam. Using the student's score on just the essay portion of the exam, I compute the number of questions that hypothetically he should have gotten right on the multiple choice part of the exam, and then subtract from this his actual raw score on the multiple choice part of the exam. This difference represents the deficiency for which I need to find an explanation. To illustrate, in column "G" on Table I, I have marked with an "X" the questions a hypothetical student in my study answered incorrectly.[22] For ease of reference, let us call him Wendell. Using the computer data on the relative difficulty and reliability of the various questions on the exam, I now seek to identify which questions Wendell missed, but should have gotten right, for the hypothesis to be true that his multiple choice score is aberrant, and his score on the essay represents an accurate assessment of his mastery of the course's subject matter.

To continue my illustration, let us assume that Wendell had a raw score on just the essay part of the exam which placed him at the 80th percentile. Let us also assume that the other students performing at the 80th percentile on the objective portion of the exam had raw scores of 37 correct answers. Wendell however got an objective score of 27, a score which placed him at the 22nd percentile on the multiple choice portion of the exam. In other words, Wendell missed ten multiple choice questions that he should have gotten correctly if his essay score indeed represents his true abilities. But now, which questions? And, even more importantly, why did he miss them?

First I look at the subject matter of questions missed. I build in a certain degree of redundancy in my exams. Important rules, subjects and concepts often are tested more than once and in slightly different ways. Not only is this useful for identifying guessing, but it also tests whether students can recognize and apply a rule when it presents itself in unconventional ways. But, I am getting ahead of myself here, for the benefit of redundancy is that if the questions I have identified as ones that a student performing at the level of his essay score should have gotten right cluster in a given subject area—will formalities, lapse, etc.—and if this represents all or most of the deficiency in their multiple choice score, then the best explanation of this student's poor

[21] For example, for multiple choice exam purposes, knowing 80 percent of the rules of intestacy is practically useless.

[22] In practice I also circle the answer choice in Column E that the student actually chose. I have not done so here to reduce clutter.

multiple choice score seems to be just that there was a lacuna in his knowledge of certain rules which were tested on the multiple choice portion of the exam, and that he "lucked out" when that subject matter was not tested on the essay portion of the exam. These students do not have a test taking problem. I often can identify their corresponding absences during times when these subjects were covered in class, or, sometimes students will outright admit to me that they were shaky on abatement, spendthrift trusts or wherever their understanding showed gaps.

Now these two explanations—compensating glibness and isolated gaps in preparation—have accounted only for a modest fraction of the students I have worked with. Moreover these students generally do not have a history of doing poorly on all multiple choice exams. Often, however, they will report that sometimes they do poorly on multiple choice, but other times, their performance on essays and multiple choice is consistent. When I ask the students in what other classes they have done poorly on multiple choice exams, that class or subject area often places a powerful and exacting emphasis on mastery of a complex hierarchy of rules (e.g. future interests, evidence) and, therefore, their differential performance can better be explained by the aforementioned phenomena of incomplete mastery and compensating glibness.

The remaining students which make it to this point now have what appears to be a true multiple choice exam taking problem. In other words, the deficiency between their actual raw scores on the multiple choice portion of the exam and the projected scores that they should have received if they performed on the multiple choice portion of the exam at the same percentile as their essay scores, is best explained by a different ability to perform on different testing vehicles.

But how to diagnose the causes of their exam woes? To do so, I need to further refine my inspection of the student's wrong answers and here is where computer analysis of his Scantron is essential. Not every wrong answer offers a diagnostic. A student will miss some hard questions just because of their difficulty. A good student also may miss a question because it is flawed and unreliable. I find the pattern of a student's multiple choice problems emerges most clearly if I filter out such questions.

Let us return to Wendell. He had 34[23] wrong exam answers, not all relevant. Using his essay score, hypothetically, he should perform at about the 80th percentile, or, in other words he should have only gotten about 24 multiple choice questions wrong.[24] While this means he should have answered correctly ten more multiple choice questions than he did, it also means he still should have gotten wrong 24 questions. Therefore, in my analysis, I eliminate Questions 21, 22, 32, 43, 45, 47, 52, 58, 59 and 60, because they are more difficult

[23] 61 questions: 27 correctly, and 34 wrongly, answered.

[24] 61-37.

than I would expect he could answer.[25] I have marked these ten questions with the letter "H," for hard, in Column H of Table I. In short, it is appropriate that Wendell got these questions wrong because only an A student would be expected to successfully negotiate them. I also in my analysis need to eliminate Questions 17, 19, 20, 23, 24, 31, 33, 44, 49, 50, 53, 54, and 61, because they are insufficiently reliable.[26] I have marked these 13 questions with the letter "B," for low biserial in Column H of Table I. Although Question 20 is close to the level of difficulty that a student like Wendell should be able to handle, data show (the point biserial of .11) that students who answered this question correctly did so by guessing.[27] Nevertheless, I do not eliminate from my analysis Questions 4, 10, 25, 38, 51 or 56. Though these questions have a low-point biserial, this merely reflects they are easy not that they are unreliable. I have marked these six questions with the letter "E," for easy, in Column H of Table I.

Now this leaves me with 14: Questions 4, 5, 7, 11, 14, 18, 26, 28, 30, 35, 39, 40, 41, and 55.[28] Wendell missed these all, though they both are reliable and within his expected reach. I have marked these questions with a check mark in Column I of Table I. I want to know why Wendell missed these questions because I have identified them as representing his performance deficiency. I now have him retake those questions in my office and based on this mini exam, I look to see: 1) whether he was consistent: i.e. did he choose the same wrong answer again or did he choose a different wrong answer, or did he now choose correctly? How do these results correlate with erasures? 2) Does he seem to miss only a certain type of question: those with long or complex stems, "K"

[25] Fifty percent or less of the top quarter got these questions right (see, Column B in Table I). I adjust this level based on the particular student's essay performance. The better the essay, the fewer multiple choice questions that I eliminate from consideration as being too hard.

[26] Those questions with a point biserial >.15 (see, Column D in Table I). It should be noted that unreliability is not the only explanation for a question with a low point biserial. A very low point biserial also may be the product of insufficient coverage of subject matter tested by a question. Hence, the students merely are guessing at a right answer because they lack the substantive knowledge necessary to analyze their way to the correct one. Consequently, sometimes a low point biserial can provide exceedingly helpful clues that a different approach to teaching a given doctrine may be needed, or that more emphasis and coverage must be devoted to a given topic. Indeed, I have found that the exact same multiple choice question performs markedly better after I revise how I cover the given topic it tests. Nevertheless, for my study, whether a multiple choice question performs poorly because if was faulty, or because the material was inadequately covered, it has exactly the same effect: The question is unsuitable to determine why the student in my study performs poorly on multiple choice while doing well on the essays.

[27] Even worse are questions with a negative biserial like Question 53 where the bottom quarter was more likely to answer it correctly than was the top quarter, and Question 49 where virtually everyone was misled.

[28] Remember that I sought to explain a discrepancy of ten questions that Wendell should have gotten correct, but did not, to bring his performance up to the 80th percentile, matching his achievement on the essay portion of the exam. Hypothetically, after eliminating questions too hard or insufficiently reliable, I suppose one ought to be left with ten questions. In practice, however, I almost always have more than this hypothetical number and never less.

type questions,[29] paired true-false questions,[30] "best answer," or policy type questions, etc. I now ask him to tell me how he chose his answer. And I ask him a battery of questions on how he takes multiple choice exams, how he manages his time, and so forth.[31] I have, thus far, discovered from this information five different, objective exam taking problems, each with a distinctive footprint.

III. Five Problems Some Good Students Have Taking Multiple Choice Exams[32]

First, let us take note of the rarest issue I have encountered: objective test anxiety. Only two students I have worked with were candidates for experiencing what I felt might be some form of test anxiety. Both had a consistent history of doing very poorly on multiple choice exams relative to essay exams and both had mediocre SAT and LSAT scores but excellent GPAs. And most interesting of all, when I sat them down to take my diagnostic mini exam, without warning or preparation, they choose the correct answer >80 percent of the time! These two, quite uncharacteristically compared with other students that I studied, utterly were unable to articulate why and how they chose their answers on their Scantrons. They also were at a loss to explain how they now in my office could choose correct answers. In fact, they both looked at their Scantrons as something foreign. I referred both to an

[29] "K" type questions (also known as multiple-multiple choice questions) have complex answer choices such as: "both (a) and (b) are correct," or instruct the student to mark all the correct answers of which several may exist for any given question.

[30] Paired true-false questions present the students with two propositions, usually statements of black letter law, and then asks the student to mark answer choice (a) if both propositions are correct, answer choice (b) if proposition I is correct and proposition II is false, answer choice (c) if proposition I is false and proposition II is correct, and answer choice (d) if both propositions are false. Incidentally, these questions, based on their point biserials, often perform very reliably, but, in practice, I find that students loathe them.

[31] I attempt to keep my interview with the student free ranging and fluid. But the typical information I want to elicit includes: Has their poor multiple choice performance been consistent? If not, is there any pattern that would explain why they do poorly on some multiple choice exams but not on others? How did they perform on standardized tests, such as the LSAT and the SAT? Do they mark and return to some questions for further review? Do they change answers? How often do they run out of time on an exam? How worried are they about doing so? How do they budget their time on an exam? Have they been told any "rules" for taking multiple choice exams? How do they narrow down the choices to select their answer? Do they treat questions that they perceive to be hard differently than those they think to be easy? How do they go about selecting a guess answer? As my study progressed, and I started to form some conclusions about why some good students had problems with multiple choice exams, my interview included questions that appeared to be diagnostic of problems I had already encountered and had started to be able to identify. I always, however, strived to keep my interview as broad and open as possible so I wouldn't foreclose the identification of more or previously undiscovered problems. I did not want to fall into the trap of finding only what I was looking for.

[32] During the course of this study, I looked for, but did not find, significant differences based on gender. Because of the small size of the study (see, footnote 1, *supra*) I think it is unlikely that, either I would have found any, or that any findings would be statistically significant.

educational psychologist who specializes in exam anxiety. Only one reported back, with substantial improvement in her objective test results after receiving professional help.

The next problem, although comparatively rare, is one of the most interesting. As odd as it may sound, this student falls in love with a certain answer and *wants* it to be the right one. (I already was sensitive to this problem because I had it myself on multiple choice exams in my first year of law school.) Of the seven or eight students I identified with this problem, most wrote exam essays that were among the best, most had participated extensively in class, enjoyed arguing and defending their positions with passion and all had better than average LSAT scores.[33] When I gave these students the mini exam in my office, they consistently choose the same wrong answer as they had on their exam Scantrons. When asked to explain their answer choices, they commonly said something like: "It seems/feels like the right answer," or "It reminds me of the case where…." When I ask them why they had not chosen the correct answer "B," instead of responding with something about the pertinent doctrines tested, i.e., rather than giving a legal reason for their choice, they instead would say something like: "I didn't like it," or "It didn't seem right to me." These students, instead of analyzing their way to the answer, apparently attempt to either intuit their answer or to select an answer consistent with some internal *weltanschauung*. It is especially fun to work with these students because, although bright, they are clueless about what they are doing. When I demonstrate to them the difference between their way and figuring an answer by analysis, their epiphany is palpable. They are amazed (as I was when my Contracts professor[34] pointed it out to me) that they have been wishing away inconvenient facts, forgiving missing elements and creating new law ad hoc from policy to come to the answer they want. Generally, I have found these students can correct their problem quickly and those who have reported back to me recount a substantial improvement on their objective test scores.

The next problem plagues the largest number of my exam-troubled students. This group unduly frets about time, although they almost never report they cannot finish the multiple choice part of the exam, and most report finishing early so they can go back to reexamine some answers. These students typically chose the "least worse" wrong answer[35] on their Scantrons. When they take the mini exam in my office, they will either choose the same wrong answer again or pick the correct answer but verbalize that they got it down to two choices and had trouble deciding between them. In other words, they tackled the multiple choice section exactly as they should. They eliminated wrong answers until they get the field down to the two best candidates, but rather than continuing this process, they, in essence, now guess, dooming themselves to a 50 percent score. These students' exams often show they have marked the question for

[33] This may not be inconsistent. The LSAT purportedly tests aptitude and not knowledge.

[34] Michael Levine, then (1971) at the University of Southern California.

[35] "Best wrong answer"?

further review "as time permits." But they fail to realize that two half analyses of a question do not equal a full one. When they return to a question for further review, the forest looks the same as it did before but instead of pushing farther into the clearing, they merely retrace their old steps and it all looks no different than before. *In other words, they give themselves the misimpression they have done more analysis without actually doing it.* The cure for this problem? Demonstrate to the student that it is a better strategy, once they commit[36] to answering a question, to spend all the time necessary to analyze it, even at the risk of running out of time. On a 50 question test, if two half analyses result in only narrowing the choices to two, and then I, in essence, select between them by guessing, my score will be 25. But if by doing a full analysis on 40 questions, I can select the correct answer 80 percent of the time, and then I fill in "B" in the remaining ten questions that I lack time for, my score (assuming five choices of answers) will be 34 (40 x .8 + 10 x .2). I present this to students as a worst case scenario, but actually typical students will not run out of time. This exam tactic gets them to reallocate their time so they do a thorough analysis once and do not split their time inefficiently. But in order to do this, they often must give themselves permission to take the risk of running out of time.

The next problem is perhaps related to the one I have just described. These students "give up" on certain types of questions. The indicia here is the questions the student missed all tend to be of a certain type: questions with a long or complex stem, k type questions,[37] paired true-false questions,[38] multiple questions based on the same stem (especially those where factual variations are introduced as the series of questions develops).[39] What typifies these questions is that they look hard and the student tends to give up on them at the beginning. Some aspect of fatigue also seems to be present, because there often will be a marked difference between how the student fares in the first versus the last half of the exam, even when it is structured with questions of varying difficulty throughout. The cure for this problem is to build the student's confidence by showing him that even hard questions can be made simple by breaking them into pieces: by treating the answer choices as true-false questions and by methodically going through the elements of a rule and its exceptions and mapping them onto the question. In other words, asking themselves what rule or exception each answer choice is designed to test.

The last problem, of those I have identified so far, seems to result from students buying into dubious "rules" of test taking. When I analyze their

[36] This is not to say that it is impermissible to altogether skip a question that appears confounding or to defer attempting any answer to it until the end of the exam. These are acceptable strategies, if sparingly employed. The problem is with making halfhearted efforts on every question that seems to be in any way challenging.

[37] *See supra* note 29.

[38] *See supra* note 30.

[39] For example, the first question is based on the stem as initially presented. The next questions also are based on the same stem but introduce variations such as: "same facts as Question 34, but X dies before Y," or "same facts as Question 34, but T's net estate is worth only $100,000.

exams, I find they missed all questions where “none of the above” or “all of the above” was the correct answer. These students blithely explain they rejected these answers because “none of the above” or “all of the above” can never be correct. In a related phenomenon, students tell me they rejected “B” because they had chosen it in their prior two answers and “there just cannot be three ‘B’s in a row.” These students hunt for patterns in their answers and trust these and their preconceived rubrics of test-taking more than their own ability to analyze a question. What a curious mixture students show of lack of confidence and wishful thinking. A cure seems to be to address both prongs of the problem. First, I point out the obvious: not only can “none of the above” be correct, but it was for this question. Not only can there be three question in a row for which “B” is correct, this was so on this exam. And because of their formula thinking, they missed getting these questions correct.[40] Second, these students also must be reassured that own abilities are more than adequate to the task. If they did superior work on their exam essays, they are fully competent in analyzing multiple choice questions.

These are the primary problems I have seen to date that account for a poor performance on multiple choice exams by students whose essays, viewed in isolation, are B+ or better. The categories described account for all but a handful of the students with which I have worked. There were four students in my study whose multiple choice exam problems I never could figure out or characterize. For this reason alone, I am sure my list of five problems is not exhaustive, but I also feel confident that the five I have identified probably represent the most common problems some good students, and probably many others,[41] confront in taking multiple choice exams.

[40] Besides, if it were indeed true that three “B”s in a row is high unlikely, why is it the last “B,” and not one of the other two, that is suspect? For some reason these students never seem to consider this dilemma.

[41] *See* text accompanying notes 9–11, *supra*.