DOCUMENT RESUME

ED 403 863 HE 029 945

AUTHOR	Wangerin, Paul T.
TITLE	Lies; Damned Lies; Statistics; and Law School Grades. Grade Conferences from Hell: Measurement Error in Law School Grading.
PUB DATE	Jul 94
NOTE	106p.; Paper presented at Annual Conference on The Science and Art of Law Teaching (Spokane, WA, July 15-16, 1994).
PUB TYPE	Speeches/Conference Papers (150) -- Viewpoints (Opinion/Position Papers, Essays, etc.) (120)
EDRS PRICE	MF01/PC05 Plus Postage.
DESCRIPTORS	Civil Rights; *Court Litigation; *Educational Malpractice; *Error of Measurement; *Evidence (Legal); Grade Prediction; Grades (Scholastic); *Grading; Higher Education; *Law Schools; Legal Problems; Psychometrics; Scoring; Student Rights; Teacher Made Tests; Testing Problems; Test Reliability; Test Validity

ABSTRACT

This paper addresses problems confronting law school teachers in grading law school exams and assigning letter grades. Using prototypical dialogue and scenarios, the paper examines mathematical and statistical issues that contribute to grading errors. Discussed in relation to real world data and the bar exam are: differential weighting, combining scores, test reliability, consistency in measurement, and standard error issues. The paper also reviews two sets of court cases. In so-called "academic challenge" cases, case law is clear--the burden of proof is on test-takers who must show that tests violate accepted norms. In "high-stakes testing" on the other hand, the burden of proof is upon test-scorers, who must prove that tests comply with accepted academic norms. Since such cases often involve claims of civil rights, court rulings are more ambiguous. This raises the question of whether most law school grades are high-stakes tests or simply academic challenge situations. Appended to the paper are sample test questions which instructors can used to evaluate their own grading biases. Also appended is a chapter, "Constructing and Using Essay and Product Development Tests" from the book, "Measuring and Evaluating School Learning", by Lou M. Carey. (Contains 50 references.) (CH)

INSTITUTE FOR
LAW SCHOOL TEACHING

GONZAGA
UNIVERSITY

SCHOOL OF LAW ● P.O. Box 3528, Spokane, Washington 99220-3528 ● 509/328-4220 (Ext. 3740)

THE SCIENCE AND ART OF LAW TEACHING
July 15-16, 1994

Lies; Damned Lies; Statistics; and Law School Grades

by

Paul T. Wangerin

Paul Wangerin is Associate Professor Law at The John Marshall Law School. He received an A.B. from the University of Missouri in 1969 and a J.D. with high honors from The John Marshall Law School. From 1978 to 1979 he clerked for the Illinois Supreme Court, and from 1979 to 1982 he practiced law with Winston & Strawn in Chicago. Professor Wangerin teaches courses in Remedies and Contracts and principally writes about Remedies and legal education issues.

Grade Conferences From Hell: Measurement Error in Law School Grading

Paul T. Wangerin
John Marshall Law School
Chicago, Illinois

Abstract

An enormous amount of anecdotal evidence, and at least some empirical evidence, suggests that many law school teachers make a number of serious "measurement errors" in connection with the grading of law school exams and the assignment of letter grades. Though some of these errors involve difficult "discretionary" issues, some involve essentially mathematical or statistical issues. The present paper discusses the second of those two kinds of error, the kind involving statistical or mathematical issues. The present paper also discusses two separate but parallel sets of judicial decisions. Both of these sets deal with judicial responses to grading disputes. One of those sets, a set dealing with what is sometimes called "high-stakes" testing, places the burden on test-scorers to prove that the tests involved comply with accepted academic norms. Conversely, the other of those sets, a set dealing with the notion of "academic challenges" to classroom grades, places the burden on test-takers to prove that the tests involved violate accepted academic norms. Law school grading, the analysis concludes, is a kind of "hybrid" situation, a situation existing somewhere between academic challenges and high stakes testing. Thus, courts must deal with the kind of law school measurement errors described herein in some sort of hybrid fashion.

Grade Conferences From Hell: Measurement Error in Law School Grading

I. Introduction

Most law school teachers have vague feelings of uneasiness regarding grading practices in the law schools. Most such teachers understand, for example, the awesome consequences of minor letter grade differences between the grades that they give to different students. The GPA's of students at the very top of a graduating class, after all, often differ by only tenths, or even hundredths, of points. Such minuscule differences, however, often lead to life-changing consequences for the students involved. Likewise, minuscule grade differences can lead to life-changing consequences for students at the bottom of a class. The difference between a 1.99 GPA and a 2.00 GPA, and thus the difference between staying in school and flunking out, is minuscule.

Many law teachers also have vague feelings of uneasiness about the very process of grading exams, and of assigning letter grades. Many law teachers realize, for example, that the grades they assign to essay type questions are inherently subjective. Further, many law school teachers probably acknowledge that the standard single exam system of law school grading, a system wherein the grade for a course reflects nothing other than performance on a single final, is, at best, educationally problematic.[1] Finally, many law teachers probably acknowledge that the different letter grades that they give to students whose test scores differ only slightly perhaps do not reflect much in the way of real differences in performance.

All of these things, and other related ones, cause law school teachers to dread student grade conferences perhaps more than any other thing. "You received the most points that I could possibly give on that question," law teachers must constantly say to desperate students during such conferences. "Your paper was significantly worse than the next higher one up." While teachers say these things, however, they frequently are thinking just the opposite. "Of course I could have given an extra point for that answer," these teachers are thinking as they mouth opposite words. "In fact, there probably was no real difference between those two exams despite the fact that I gave them different grades."

Fortunately for most teachers, students aggrieved by grading decisions generally know very little about two things. First, these students generally know very little about the legal rights that they might have vis a vis their teachers and their schools. Second, these students generally know very little about the principles of educational testing and measurement.

The following analysis addresses both of those shortcomings in most students' knowledge. First, building on widely accepted ideas from the world of educational measurement, the analysis suggests that the grades that many law school teachers give in probably are affected by a number of very serious "measurement" errors.[2] Those errors, the analysis suggests, involve (1) "weighting" issues, (2) "reliability" issues, and (3) "standard error of measurement" issues. Second, the analysis describes the difference in burdens of proof between "academic challenge" lawsuits involving traditional classroom grades and lawsuits involving the results on "high stakes" tests. Students have virtually no chance of success if they file academic challenge suits. Conversely, a substantial possibility of success exists for students in high stakes testing suits. Law school grading practices, the analysis then suggests, are a hybrid between traditional grading and high stakes testing. Thus, courts of law should be allowed to intervene in connection with the kind of measurement errors described herein.

Three introductory points must now quickly be made. First, in connection with the preparation of the present analysis, this writer _informally_ collected 13 sets of grades from 11 teachers at a large urban law school. While collecting these sets of grades, this writer made no attempt to make certain that these sets of grades, these teachers, and this school, are representative generally of law school grades, teachers and schools. Thus, it is possible that the real world grading problems described herein are entirely isolated. Having said that, however, it must also be noted that no reason exists to think that the data collected is _not_ representative of grades, teachers and schools generally. Thus, it is surely possible that the real world grading problems described herein are widespread. Second, although the following discussion regularly refers to ideas from the world of statistical analysis, readers who know nothing about that world have nothing to fear herein. Unlike some discussions of statistics, discussions which seem to revel in abstruseness, the present analysis strives for simplicity above all else.[3]

The third point yet to be made relates closely to the second. Although most law school teachers seem to have little knowledge of the kinds of measurement and statistical issues discussed herein, and thus make the kind of grading mistakes described, the ideas described herein are not totally foreign to the field of law. _The Bar Examiner_, for example, frequently publishes articles addressing the kinds of measurement and statistical ideas discussed herein.[4] Further, and perhaps more importantly, as the last part of this analysis reveals, courts that have dealt with "high stakes" tests regularly address these issues.

II. Measurement Error in Law School Grading

Countless law students have argued in countless grade conferences that their teachers have "erred" by not giving them additional points on particular essay questions. And countless law students have argued in such conferences that their teachers have "erred" because the teachers gave too many C's or D's or the like. The present analysis completely ignores these kinds of alleged grading errors. It does so, in turn, for a straight forward reason. Grading errors of the kind just described are errors of judgment or discretion. Thus, obviously, errors of this kind cannot easily be quantified or turned into hard and fast numbers. Further, careful analysis might well reveal that errors of this kind are not errors at all. Rather, these errors might turn out to be nothing more than differences of opinion.

The present analysis concentrates on an entirely different kinds of measurement error. These errors involve purely mathematical or statistical issues. Thus, discretion plays no role whatsoever here. Thus, if as demonstrated herein a teacher makes statistical errors when combining scores from different parts of exams, those teachers cannot simply point to discretionary decisions. If the numbers are wrong, then the numbers are wrong. Period. Likewise, if as demonstrated herein a teacher fails to employ a test that is sufficiently reliable, or fails to take into account an appropriate amount of measurement error, that teacher again cannot simply hide behind the notion of discretion. Again, if the numbers are wrong, then the numbers are wrong.

Note carefully now an important point. Although the present analysis addresses nothing other than statistical or mathematical types of grading errors, the errors discussed herein are by no means simple or obvious ones. The present paper, for example, spends no time on situations in which teachers simply miscalculate addition sums. And the present paper spends no time on situations in which teachers overlook the second of two blue books, or inadvertently fail to read an answer to a particular question. Rather, the present analysis concentrates on kinds of statistical or mathematical errors that measurement experts fully understand but that law school teachers for the most part know nothing about.

Three different kinds of measurement error are described below. The first of those, herein called "weighting" problems, occurs when teachers total up the scores from different parts of exams. The second of those, dealing with the "reliability" of tests, involves questions about the likelihood that a particular test will produce consistent results. The third of these errors deals with the notion of the "standard error of measurement." Standard error of measurement problems occur when teachers assign different letter grades to students whose point totals differ by relatively small amounts.

A. Differential "Weighting"

It is probably safe to say that the most dramatically obvious "measurement error" that law school teachers routinely make involves the "weighting" of different parts of exams.[5] Teachers make this mistake, in turn, because they assume that they can simply add up the scores from different parts of exams. Teachers who give a test that contains four "equally weighted" essay questions, for example, usually assume that they can just add up the scores from those different questions and assign grades in light of the totals. Likewise, teachers who give exams that combine essays with objective-type questions usually also assume that they can just add up point scores. As the following will demonstrate, however, such simple addition can lead to very, very serious mistakes.

1. Teacher 1

Assume that the following figure contains scores and grades that Teacher 1 gave on a test that contained four "equally weighted" essay questions. (For convenience only, students here are identified by names rather than numbers.)

Figure 1

NAME	#1	#2	#3	#4	TTL	GRD
JOHNS	4	2	2	5	13	F
DIAZ	2	4	5	4	15	D
CHENG	4	7	2	4	17	C
PHILL	4	5	3	5	17	C
PLURS	3	4	9	4	20	C +
HOPPE	4	9	5	5	23	B
SMITH	5	6	6	6	23	B
ADAMS	4	7	7	5	23	B
JONES	4	10	9	5	28	A
WENDI	3	10	10	5	28	A
MEAN:	3.7	6.4	5.8	4.8	20.7	

Note carefully some of the details regarding these grades. First, Teacher 1, like most teachers, has calculated the "mean" (or average) scores so that she can tell roughly where students stand in relation to each other. Further, Teacher 1, like most teachers, has ranked her students by overall point totals, and assigned letter grades in light of those point totals.

Imagine now that students Smith and Adams appear for a joint grade conference with Teacher 1. Teacher 1 tells these two students that they both received the same number of points (23), a number that put them somewhat above the middle of the class. These scores, the teacher then tells these students, earned them grades that were just above average, namely, B's. But, the students press on, producing the first of a group of "grade conferences from hell."

Smith: How did I do on the individual questions?

Teacher 1: Well, Smith, on questions 1 and 4 you did pretty good. In fact, very good, at least in relation to other students. On those two questions you got the highest scores that I gave, namely, 5 and 6. On questions 2 and 3, however, not so good. You got only the average scores, namely 6.

Smith: So, I got the highest score given on half of the test, and the average score on the other half?

Teacher 1: Yes, that's correct.

Smith: And I still got only the average grade for the course?

Teacher 1: Yes, that's correct.

Adams then cuts in.

Adams: How did I do on the individual questions?

Teacher 1: Well, on all of the questions you got about the average scores, namely, 4, 7, 7 and 5.

Adams: So, I got the average score for the whole test?

Teacher 1: That's correct.

Adams: And I got a B, the average grade

Smith butts in.

Smith: Wait a minute. Adams got average scores for all of the questions and he got a B. I got average scores on two questions but I got the top score on two questions. And I also got a B. That's crazy.

Teacher 1: Well, you two got the same number of points, so I had to give you the same grade.

Smith: But you admit yourself that I wrote a much, much better exam than Adams

Teacher 1: I'm sorry. There's nothing I can do. Numbers don't lie.

Smith and Adams then leave, and Teacher 1 begins to take what she hopes is a long break from students. But in walk Johns and Diaz, also study partners. And the nightmare continues.

Teacher 1: Well, Johns, you received the lowest point total that I gave, namely 13 points. Hence, you got the lowest grade, F. Ms. Diaz, you got 15 points. Since this was not quite as bad as Johns, you got a D.

Johns: Tell me, if you will, about my individual scores.

Teacher 1: Well, Johns, your score on questions 1 was 4 and on question 4 was 5. Those scores were just slightly above the average. But your score for question 2 was 2 and for question 3 was 5. Those scores were well below average.

Johns: So on half the test I scored slightly above average and on half I scored below average.

Teacher 1: Yes, that's correct.

Johns: And yet I got the lowest grade in the class?

Teacher 1: Yes, that's correct.

Johns: Let me make sure I understand this. On half the test I scored above average and I still got the lowest grade you gave?

Diaz cuts in before the teacher can answer.

Diaz: What were my scores?

Teacher 1: Well, Diaz, your scores were 1, 4, 5 and 4, respectively. All of those scores were below average, some way below average.

Diaz: I guess that's why I got a D.

Teacher 1: Sure.

Johns suddenly interrupts.

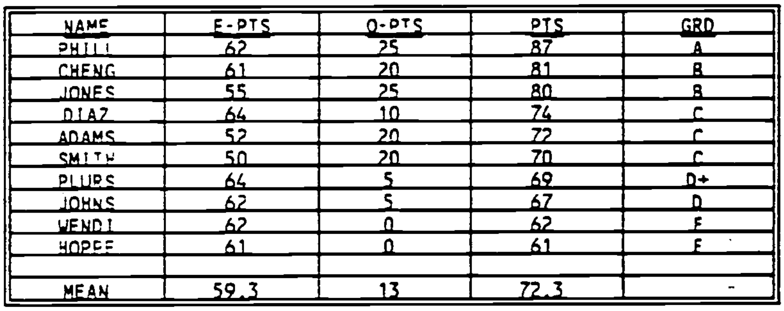
Johns: Wait a minute. I don't understand. Diaz was below average on every question, but I was below average on only two of the questions. I was above average on two of the four. But she got a D and I got an F. That doesn't make sense.

Teacher 1: I'm sorry, the numbers don't lie. And, I'm sorry, but our time is up.

2. Teacher 2

Regrettably, essay type exams are not the only kinds of exams that produce "weighting" problems such as those just described. Such problems also occur, perhaps even to a greater extent, when teachers administer exams that contain both essays and objective-type questions. Consider the following. Assume that the following figure contains the scores and grades that Teacher 2 gave on an exam that contained several essay questions, worth a total of 75 points, and a set of objective-type questions, worth, together, 25 points. (The term "E-Pts" in this figure, obviously, stands for essay points, and the term "O-Pts" stands for objective-question points.) This figure shows that Teacher 2, like most teachers who administer such exams, simply added up point totals and assigned grades accordingly.

Figure 2



NAME	E-PTS	O-PTS	PTS	GRD
PHILL	62	25	87	A
CHENG	61	20	81	B
JONES	55	25	80	B
DIAZ	64	10	74	C
ADAMS	52	20	72	C
SMITH	50	20	70	C
PLURS	64	5	69	D+
JOHNS	62	5	67	D
WENDI	62	0	62	F
HOPPE	61	0	61	F
MEAN	59.3	13	72.3	-

Imagine now that several of Teacher 2's students have appeared for routine grade conferences. Quickly, however, these grade conferences turn nasty.

Diaz: Both Adams and I got C's. But we would like to see how we did on the different parts of the exam.

Teacher 2: Sure, no problem. Diaz, you got 64 points on the essays and 10 points on the objectives.

Diaz: How's that in relation to other people?

Teacher 2: Good question. Well, a 64 on the essays was quite good. In fact, very good. That was the highest score that I gave. (You shared it with one other person.) But, the 10 on the objectives was well below average. That really pulled you down. So, you got a C.

Diaz: Let me make sure I understand this. I got the highest score that you gave on the essays. And they were worth 75% of the grade. I still only got a C for the course.

Teacher 2: That's correct. The total number of points you got, 74, was just about average, so you got the average grade, C.

Diaz: But how could I get the average number of points when my essays, which were worth 75%, were the best?

Teacher 2: I'm sorry, but the numbers don't lie.

Adams then takes over.

Adams: Can you tell me about my individual scores?

Teacher 2: Sure. You got 52 points on the essays. That was substantially below average.

Adams: How did I do on the objectives.

Teacher 2: You got 20 points on the objectives. That was quite. The objectives, in other words, pulled you up. Your total points, 72, put you right in the middle. Thus, you got a C.

Diaz: [Cutting in] Wait a minute, wait a minute. I'm confused. Let me get this straight. Adams did very poorly on 75% of the test but very well on 25%. And I did very well on 75% and very poorly on 25%.

Teacher 2: That's correct.

Diaz: But we got the same grade. How's that possible? I should have gotten a much better grade than him.

Teacher 2: I'm sorry. You two got just about the same total number of points. So I had to give you the same grade. The numbers don't lie.

Diaz: But, wait.........

Teacher 2: I'm sorry, I don't have anymore time. The numbers simply do not lie.

Assume now that Diaz and Adams leave. But in walk Plurs, Wendi and Jones.

Teacher 2: Let's get right to it. Plurs, you got 64 points on the essays and 5 points on the objectives for a total of 69. That total is very bad. So you got a D+

Plurs: What were those individual numbers like in comparison to others.

Teacher 2: 64 was quite good, in fact, excellent. On the essays you shared the highest score. 5 points on the objectives, however, was quite poor. So your total of 69 was, as I said, very poor. Frankly, a D+ was a gift.

Plurs: I don't get it. I did excellent on 75% of the exam and I got a D+?

Before the teacher can answer, Wendi cuts in.

Wendi: How did I do?

Teacher 2: Well, again you did well on the essays, namely 62 points. That was a bit above the average. But you did terribly on the objectives, namely 0 points. Your total, 62, was the second lowest. So, sadly, I had to fail you.

Wendi: Wait a minute. How's that possible. I got above average on 75% of the exam and I still failed.

Teacher 2: That's correct. The objectives killed you.

Wendi: But they were worth only 25%. That's a small part.

Teacher 2: I have to repeat. The numbers don't lie. You had the second lowest overall point total. I had to fail you.

Plurs jumps back in.

Plurs: Wendi's situations, in other words, is sort of like mine. I did excellent on the essays -- which you said were worth 75% -- and I still got a D+.

Hoppe then cuts off Plurs, anxious to find out what happened to her.

Hoppe: You failed me also. What happened with me?

Teacher 2: Well, Hoppe, you failed because your point total was very bad. In fact, that total, 61, was the lowest anybody got.

Hoppe: How did those points break down?

Teacher 2: Well, you did pretty good on the essays, namely, 61 points. That was a bit above average. But you blew the objectives. On them you got 0 points.

Hoppe: This sounds like what happened with Wendi. Above average on 75% of the exam; very poor on 25%.

Teacher 2: Correct.

Plurs: [Speaking to Wendi and Hoppe]. Well, you think you have it bad. I did excellent work on the essays, right at the top of the whole class. And I got a D+.

Teacher 2: I'm sorry but I have to run off to a faculty meeting. Just remember. Numbers cannot lie.

3. Combining Scores

The foregoing examples reveal, of course, that in fact numbers do lie, at least when it comes to grades and grading. Obviously, something went wrong as Teachers 1 and 2 combined the scores from different parts of the exam, and assigned grades. But, what was it that went wrong?

Experts in statistics and educational measurement know that sets of numbers can be widely spread out or tremendously compressed, or anything in between.[6] Thus, for example, if many of the students who take a test get very similar scores on one of the questions, the scores for that question will be very compressed. Conversely, if many of the students who take that test get very different scores on another question, the scores for that other question will be widely spread out. Although experts describe this sort of variability and compression in various different ways, perhaps the most common way of describing it involves use of the so-called "standard deviation." The standard deviation of a set of numbers is, simply, a number that describes how spread out or compressed the individual numbers in that set are. The higher the standard deviation, the larger the spread. And, of course, vice versa. (Note: Even people who know nothing about statistics can easily calculate standard deviations on any computerized spread sheet.)

Experts in statistics and educational measurement long ago noted the existence of a counter-intuitive fact about sets of numbers with <u>different</u> standard deviations. If such sets are simply added, these experts discovered, sets with high standard deviations (sets that are widely spread out) <u>inadvertently</u> may end up counting more than sets with low standard deviations (sets that are tightly compressed).

A simple illustration shows why this counter-intuitive thing happens. Imagine that somebody has been asked to sort a group of people from large to small and has been told to factor the height and weight of the people in the group equally while doing so. Imagine also, however, that most of the people in the group turn out to be quite similar in height but very, very different in weight. If this occurs, the similarity in height will cause height to be discounted as a sorting factor. Conversely, the differences in weight will cause weight to be exaggerated as a sorting factor.

Not surprisingly, the same thing can happen with sets of scores. Assume, for example, that a teacher plans to count each of two essays on a test equally.[7] Assume also, however, that scores on the first essay range from 80 to 100 and the scores on the second essay range from 20 to 100. Finally, assume that one student has the highest score on the first essay, 100 points, and the lowest score on the second, 20 points. And assume that another student has the lowest score on the first essay, 80 points, and the highest score on the second, 100 points. Since the performance of these two students was exactly the same -- highest on one, lowest on the other -- the grades they get should be the same. But, addition reveals that the first student will get 120 points and the second 180 points.

In many grading situations, of course, this differential weighting phenomenon will have no real world impact. But, consider what might happen if an individual student did well on a question that ended up being under-weighted and poorly on a question that ended up being over-weighted. Or consider what might happen if exactly the opposite happened for a different student. In these situations, major grading mistakes could occur. The first student described could easily get a significantly worse grade than deserved. And the second student could easily get a significantly better grade than deserved.

Not surprisingly, this "double whammy" phenomenon is exactly what happened to both Teacher 1 and Teacher 2. The scores on Questions 1 and 4 on Teacher 1's test are tightly compressed, with standard deviations of .82 and .63 respectively. Conversely, the scores on questions 2 and 3 on Teacher 1's exam are widely spread out, with standard deviations of 2.72. and 2.94 respectively. This means, of course, that when Teacher 1 added up the scores on her exam she inadvertently gave too much weight to questions 2 and 3, and too little weight to questions 1 and 4. The same thing happened to Teacher 2. The essay scores on Teacher 2's were tightly compressed, with a standard deviation of 5.06. Conversely, the objective scores were widely spread out, with a standard deviation of 10.06. Thus, when Teacher 2 added up point totals, he ended up distinctly undervaluing the essays ended and distinctly overvaluing the objectives.

Recall some specific examples. Smith, a participant in one of the conferences with Teacher 1, did very, very good work on Questions 1 and 4 but only average work on questions 2 and 3. Because of the double whammy phenomenon, therefore, his point total significantly under-rated his performance on the exam. Johns's situation with Teacher 1, however, is exactly the opposite. Johns did very poor work on questions 2 and 3 and average work on questions 1 and 4. Thus, Johns' point total significantly over-rated his performance. Consider also some of Teacher 2's students. Hoppe and Wendi and Plurs, it should be recalled, all did average or superior work on the essays -- worth 75% of the exam -- but terrible work on the objectives -- worth 25% of the exam. Nevertheless, all three got terrible grades. Obviously, the double whammy phenomenon caused this. The objective questions on Teacher 2's test _inadvertently_ ended up counting much more than 25%. And the essays on that test _inadvertently_ ended up counting much less than 75%.

Note now an important point. Some teachers might insist that the only thing that they care about is the total number of points earned on an exam, not the questions on which those points are earned. Further, other teachers might argue that questions that produce very compressed sets of scores are poorly designed questions and _ought_ to be undervalued and questions that produce widely spread out scores _ought_ to be over-valued. These teachers will then insist that the weighting problems just described will play no role in the grades that they assign. In fact, these teachers will be correct. If good questions _ought_ to be valued more than poor questions, and if total points earned is the _only_ thing that matters, then weighting effects will not matter.

Unfortunately, however, a fatal problem exists in the reasoning of these teachers. Teachers who really do not care where points are earned, or who mean to under-value and over-value certain questions, must say something like the following to their students:

> In response to your questions regarding the "weight" assigned to the different questions on this exam, I must, in all candor, say this. I simply don't know at the present time how much the different questions on the exam will end up being worth. Some of them may end up counting for a lot, and some of them may up counting for very little. Or, they may all end up counting the same. I won't know the specifics until I'm done grading. Further, since I do not now know how much the individual questions will be worth, I simply cannot now tell you how much time to spend on them. Perhaps, you should divide your time equally. Or, perhaps you should blow off one

or more of them and concentrate on the others.

Oh, and good luck to you all.

Obviously, teachers cannot actually say things like this to their students. But, if they mean for point totals to be the only thing that counts, and if they mean for under-valuing and over-valuing to occur, then they must say this to their students.

4. "Weighting" Solutions

Fortunately, teachers need not go in front of their students and throw bombs like the one just described in order to avoid differential weighting problems. Rather, teachers can use either of two simple techniques.[8] First, teachers can employ the same curve in connection with the scores or grades that they give on all of the different parts of an exam.[9] Thus, for example, a teacher can assign letter grades to all of the parts of an exam in light of, say, the following curve -- A's = 20%, B's = 30%, C's = 30%, D's = 20%. Once teachers do this, they create sets of scores that are the same in terms of compression or variability. And, if sets have the same degree of compression or variability, the differential weighting phenomenon does not occur.

Regrettably, teachers who employ a curving procedure like this can encounter problems of a different kind. For one thing, this approach forces teachers into an extra grading step, a step wherein the teachers convert the raw scores into curved scores. Further, use of this technique places teachers in a grading straight jacket. Scores on different questions, after all, will not naturally fall into the precise same pattern.

Fortunately, another technique for solving the differential weighting problem also exists, a technique that can easily be used by any teacher who has access to, or whose secretary has access to, a computerized spread sheet. This technique involves the conversion of "raw" scores into "standardized" scores. (Raw scores are the actual scores that teachers give to students when exams are graded.) Once raw scores are standardized, they can be added up, subtracted, averaged, combined and the like without any fear of weighting distortions.[10]

Standardized scores -- the most common being the so-called "Z-scores" or "T-scores" -- describe the distance of a raw score from the mean score in a group of scores.[11] Thus, a Z-score of -2.14 indicates that the raw score at issue is 2.14 standard deviations below the mean score in the group. (This, incidentally, is a very low score.) Conversely, a Z-score of [+]2.14 indicates that a particular raw score is 2.14 standard deviations above the mean score in the group of scores involved. Z-scores, which are more basic kind of standardized scores, are created by subtracting the mean score in a set of scores from the raw score involved, and then dividing the resulting number by the standard deviation of the set of scores involved.[12]

$$Z = \frac{\text{Raw Score - Mean Score of Group}}{\text{Standard Deviation of Group}}$$

All of this again brings the analysis back to the grades of Teachers' 1 and 2. The following figure shows the Z-scores, including total or average Z-scores for Teacher 1's test. This figure also shows what happens when the pertinent grading data is sorted, not by point totals, as Teacher 1 originally sorted the data, but by Z-totals. Finally, this figure shows what would have happened at Teacher 1 assigned grades by Z-scores rather than by point scores, <u>using the precise same overall grading "curve."</u>

FIGURE 3

NAME	#1-Z	#2-Z	#3-Z	#4-Z	PTS	Z-AVG	P-GRD	DR
SMITH	1.59	-0.15	0.07	1.9	23	0.8525	B	
JONES	0.37	1.32	1.09	0.32	28	0.775	A	
WENDI	-0.85	1.32	1.43	0.32	28	0.555	A	
HOPPE	0.37	0.96	-0.27	0.32	23	0.345	B	
ADAMS	0.37	0.22	0.41	0.32	23	0.33	B	
PHILL	0.37	-0.51	-0.95	0.32	17	-0.193	C	
PLURS	-0.85	-0.88	1.09	-1.27	20	-0.478	C+	
CHENG	0.37	0.22	-1.29	-1.27	17	-0.493	C	
JOHNS	0.37	-1.62	-1.29	0.32	13	-0.555	F	
DIAZ	-2.07	-0.88	-0.27	-1.27	15	-1.123	D	

Everything now makes a great deal more sense. Consider Smith. He, it should be recalled, did the best work in the class on two of the four questions, and about average on two others. Nobody else did such strong work on so much of the test. Point totals give him a B. Standardized scores give him an A. Which seems more appropriate? Or consider Johns. Admittedly, he did very poor work, however, it be scored. But, point totals give him the lowest grade in the class, and standardized scores give him the second lowest. Which seems more appropriate.

Z-score operations reveal similar problems -- albeit even more serious ones -- in connection with Teacher 2's grades. The following figure shows the Z-scores for the different parts of Teacher 2's exam for individual students. Further, this figure shows Z-score "totals." (Because Teacher 2 intended to assign 3 times as much weight to her essays as to her objective-type questions, the z-total reflects that kind of multiplication.) Finally, this figure shows all of the data sorted by Z-totals rather than by point totals. And, again, this figure shows what would have happened had Teacher 2 assigned letter grades, <u>using the same curve</u>, in light of z-scores rather than point totals.

Figure 4

NAME	E-PTS	Z	O-PTS	Z	PTS	Z-TTL	P-GRD	Z-GRD
PHILL	62	0.53	25	1.19	87	2.78	A	A
DIAZ	64	0.93	10	-0.3	74	2.49	C	B
PLURS	64	0.93	5	-0.8	69	1.99	D+	B
CHENG	61	0.34	20	0.7	81	1.72	B	C
JOHNS	62	0.53	5	-0.8	67	0.79	D	C
WENDI	62	0.53	0	-1.29	62	0.3	F	C
HOPPE	61	0.34	0	-1.29	61	-0.27	F	D+
JONES	55	-0.85	25	1.19	80	-1.36	B	D
ADAMS	52	-1.44	20	0.7	72	-3.62	C	F
SMITH	50	-1.84	20	0.7	70	-4.82	C	F
MEAN	59.3		13		72.3			
S.D.	5.06		10.06					

Consider Plurs. Plurs had the highest score given on 75% of the exam (the essays) but, he blew the objectives, worth 25% Point totals give him a "D+". Z-scores, however, give him a "B". Which grade makes more sense? Or consider Diaz. Diaz had the highest score given on the essays, and did just below average work on the objectives. Should she get a "C" (point scores) or an "A" (Z-scores)? And consider the other end of the scale. Adams and Smith did very, very poorly on the essays but well above average on the objectives. Conversely, Wendi and Hoppe did above average on the essays and very, very poorly on the objectives. If F's have to be given at all, who should get them? Wendi and Hoppe (point totals)? Or Adams and Smith (Z-scores)?

5. Real World Data

What then about real world data?

As noted at the outset here, 13 different sets of real grades were studied in connection with the preparation of this analysis. Of those 13, 10 involved combinations of scores. (3 sets of grades were from tests that were made up entirely of equally weighted objective-type questions.) Some of these 10, for example, involved combinations of scores on several different essays. Others involved combinations of scores on essays and scores on sets of objective-type questions. (And two of these 10 involved combinations of scores from different tests.) The results of Z-score analysis of these ten sets of grades are this:

1. On 10 out of 10 of these sets of grades, Z-score calculations produced different rankings for students than point total calculations. In other words, on 10 out of 10 sets of grades, point totals presented distorted pictures of at least some students' performance on the exams.

2. On 8 out of 10 of these sets of grades, z-score calculations would have produced different grades for at least some students. In other words, in 8 out of 10 of

these sets, students who actually did better (or worse) work on an exam than other students got lower (or higher) grades than those other students.

3. Most of the 8 sets of grades just described contained relatively minor grade distortions. On most of them, for example, a couple of students got B's when they should have gotten C+'s or B+'s. In several of those 8 sets, however, major grade distortions occurred.

Consider a real world example in connection with the last point just made. Professor B is a real teacher at a real law school. Teacher B, like Teacher 1 in the foregoing examples, gave a test made up of several essay questions. As with Teacher 1, scores for some of the essays on Teacher B's exam were rather compressed, and scores on other essays were rather spread out. Despite that fact, however, Professor B simply totaled up the points. The following figure shows Professor B's actual point totals, and the actual grades that he gave. The following figure also shows the results of z-score calculations and z-score "sorting." The student with the highest Z-score -- and hence the best actual performance on the exam -- is at the top of the list. Conversely, the student with the lowest z-score -- and hence the worst actual performance on the exam -- is at the bottom of the list. (Note: Underlined grades in the following figure, incidentally, are grades that were "adjusted" for things like class participation and the like. These grades, therefore, should be ignored in connection with the present analysis.)

Figure 5

I.D.	PTS	GRD	Z-ITL
36	80	B+	3.91
29	87	A	3.76
57	86	A	3.57
43	86	A	3.36
26	85	A	3.3
64	81	B+	2.97
67	81	B+	2.9
53	81	B+	2.74
34	79	B+	2.52
35	81	A	2.51
66	78	B+	2.46
31	82	B+	2.46
11	77	B+	2.13
15	72	B	1.97
18	76	B+	1.83
32	70	B	1.67
59	72	B	1.53
10	72	B	1.28
68	73	C	1.09
68	73	B	1.09
20	74	B	0.98
9	68	B	0.97
19	71	B	0.73
52	68	B+	0.68
14	69	B	0.68
55	67	C+	0.57
16	65	C+	0.42
30	68	B	0.42
13	65	C+	0.22
69	66	C+	0.19
4	63	C+	0.15
23	62	C+	0.09
17	63	C+	0.03
24	64	C+	-0.09
56	61	C+	-0.13
63	59	C	-0.23
60	61	C+	-0.32
54	62	C+	-0.33
33	60	D+	-0.34
44	58	C	-0.42
28	61	C+	-0.47
25	60	C+	-0.51
65	60	C+	-0.61
2	58	C	-0.7
8	53	C	-0.95
49	54	C	-1.08
12	56	C	-1.14
47	51	C	-1.19
61	54	C	-1.29
27	55	C	-1.34
51	53	C	-1.39
42	55	C	-1.44
50	52	C	-1.45
3	55	C	-1.58
37	53	C	-1.78
48	48	D+	-1.81
22	49	D+	-2.14
39	47	D+	-2.2
45	47	D+	-2.26
46	47	D+	-2.26
62	50	D+	-2.32
21	50	C	-2.44
38	46	D+	-2.67
5	49	D+	-2.67
6	45	D+	-2.69
40	46	D+	-2.76
7	40	D+	-2.95
1	45	D+	-3.09
41	35	D	-4.04

Note the discrepancies between point scores and standardized scores. Student 36 actually did the best work in the class. But she only got a B+. Five students who did poorer work got A's. Likewise, Student 68 got a C despite the fact that he did work that was roughly comparable to that done by students who generally got B's. Note particularly, however, Student 33. This student got a D+ for the course despite the fact that numerous students who did poorer work on the exam got C's and some who did poorer work even got C+'s. Conversely, note Student 21's incredible good luck. This student got a C despite the fact that his work was comparable to work that earned other students D+'s.

Consider also another real world example. Professor F, like Teacher 2 in the foregoing analysis, administered a test that combined several essays and a set of objective-type questions. Scores on F's essays, like Scores on Teacher 2's essays, tended to be very compressed. Conversely, scores on F's objectives, like scores on Teacher 2's essays, were widely spread out. Like Teacher 2, Professor F simply totaled up point scores. The following figure shows the actual point totals that Professor F calculated, and the actual grades that she gave. This figure also shows, however, the results of z-score calculations. The student with the highest Z-score -- hence the best performance on the exam -- is at the top of the list, and the student with the lowest z-score -- hence the worst performance on the exam -- is at the bottom of the list.

Figure 6

I.D.	PTS	GRD	Z-TTL
5	91	A	1.08
59	90	A	0.88
43	87	A	0.81
38	88	A	0.74
4	85	B+	0.72
28	87	A	0.66
61	84	B	0.64
48	86	B+	0.62
31	86	B+	0.61
15	87	A	0.6
35	82	B	0.57
13	83	B	0.53
44	85	B+	0.53
57	85	B+	0.52
32	81	B	0.39
1	81	B	0.37
22	81	B	0.37
16	81	B	0.35
52	81	B	0.32
30	81	B	0.31
6	78	C	0.3
63	79	C+	0.29
34	80	B	0.28
56	79	C+	0.27
72	79	C+	0.27
67	79	C+	0.24
14	79	C+	0.24
2	79	C+	0.17
12	79	C+	0.13
68	78	C	0.09
21	78	C	0.06
19	76	C	0.05
17	76	C	0.05
55	74	C	0.04
69	80	B	0.03
51	75	C	0.02
24	75	C	0.02
10	75	C	0.01
11	75	C	-0.01
25	72	C	-0.08
73	74	C	-0.11
49	69	D+	-0.14
39	74	C	-0.15
33	70	C	-0.15
3	72	C	-0.16
36	74	C	-0.17
54	70	C	-0.18
70	74	C	-0.2
64	70	C	-0.2
41	70	C	-0.2
29	69	D+	-0.24
53	69	D+	-0.28
45	67	D	-0.29
62	72	C	-0.3
60	72	C	-0.3
47	69	D+	-0.3
65	68	D	-0.31
23	69	D+	-0.33
8	67	D	-0.33
46	68	D	-0.34
7	66	D	-0.36
26	67	D	-0.39
27	70	C	-0.45
58	69	D+	-0.46
42	62	F	-0.51
9	68	D	-0.51
71	61	F	-0.62
20	66	D	-0.67
50	62	F	-0.71
37	57	F	-1
66	58	F	-1.01
18	57	F	-1.03
40	45	F	-1.68

Note the discrepancies. Student 61 got a B despite the fact that several students who did poorer work on the exam got A's and B+'s. Student 69 had better luck. He got a B though his work was roughly comparable to that of students who got C's. Likewise, Student 27 caught quite a break. He got a C despite the fact that his work was comparable to that done by students who generally got D's. Note finally the extraordinarily bad luck of the Students 42 and 71. Both of these students failed the course despite the fact that students who did poorer work obtained passing grades.

6. The Bar Examination

A brief digression from this analysis of law school tests and grading must now be made. This digression involves, however, a subject that is of considerable interest to law students, namely, the bar examination.

No one could dispute that bar examinations play an extraordinarily important role in the lives of law students and lawyers. People who cannot pass bar examinations generally cannot practice law. Given that fact, the possible existence of measurement error in the bar exam is a topic of considerable interest.

Fortunately, literature addressing bar exam issues indicates that bar examiners are aware of the weighting problems just described. Stephen Klein, for example, who has written extensively about measurement issues and the bar exam, addresses this weighting issue at considerable length.[13] Klein first notes that weighting problems can occur when the scores for the different essays on the essay portion of the exam are added up. This is comparable, of course, to Teacher 1's situation. Weighting problems can occur in this context, Klein notes, because the compression or variability of the scores on the different essays -- technically, the standard deviations -- can be very, very different. Klein notes in this context, incidentally, that these problems are "not trivial."[14] "I have seen," he continues "essay questions whose standard deviations were three times as great as the standard deviations of other questions on the same examination."[15] In other words, Klein has weighing problems on the essay portions of bar examinations that are just as bad as the problems described herein.

The second area in which Klein notes that weighting problems can occur is roughly comparable to what happened to Teacher 2 herein. On bar examination, essay portions of those exams must be combined with objective portions.[16] As soon as such combining is undertaken, however, weighting issues arise. Fortunately, the means bar examiners use to avoid this particular problem need not be described here. Rather, the only thing that needs to be said here is that bar examiners do not seem to make weighting mistakes when they combine scores from essay questions and objective-type questions.

These references to bar exam practice raise a final point. If bar examiners can do the things necessary to eliminate weighting problems, why cannot law school teachers also do those things. Admittedly, the score that students receive on the bar exam is much more important than the score that students get on any individual law school test. Nevertheless, real similarity exists between these situations. In both, a single test is used to make an important educational decision about an individual person.

B. The "Reliability" of Tests

Regrettably, weighting errors such as the ones just described are not the only kinds of measurement errors that law school teachers probably make in connection with law school exams. Most law school teachers seem to know little or nothing about the measurement concept of "reliability."[17] Thus, not surprisingly, many such teachers almost certainly make reliability errors.

1. Teacher 3

The figure below contains the scores that another hypothetical teacher, "Teacher 3," gave on an exam that consisted of six equally weighted essay questions.

Figure 7

NAME	#1	#2	#3	#4	#5	#6	TTL	GRD
JONES	6	8	7	9	6	6	42	A
ADAMS	9	7	8	6	7	5	42	A
WENDI	8	6	8	5	9	6	42	A
PLURS	8	6	9	7	5	3	38	B+
HOPPE	8	5	8	6	7	3	37	B
PHILL	5	9	3	7	4	8	36	C+
CHENG	6	7	5	7	6	5	36	C+
SMITH	6	5	5	4	4	7	31	C
DIAZ	2	8	3	7	8	2	30	C
JOHNS	4	4	6	4	6	5	29	D+
MEAN:	6.2	6.5	6.2	6.2	6.2	5	36.3	
S.D.:	2.15	1.58	2.15	1.55	1.62	1.89	4.97	

A quick look at these scores and grades reveals that the standard deviations for the scores for the different questions are somewhat different. In other words, scores on some of the questions were more compressed than scores on other questions. Thus, "weighting" problems such as those already described may well have occurred. This quick look at these scores and grades, however, almost certainly does not reveal the existence of another serious measurement error problem. But, such a problem exists. And it is a problem that is in many ways much more serious than the weighting problem.

Another grade conference from Hell introduces this problem. Diaz, Jones, Phill and Smith appear for a conference with Teacher 3.

Teacher 3: Well, let's get right to it. You three will recall that the test had six equally weighted questions, each worth a maximum of 10 points. They were, in my opinion, equally difficult. And they dealt with equally important areas of the law.

Diaz: (Impatient) How did I do?

Teacher 3: Well, Diaz, you got a total of 30 points. Way below average. In fact, you're lucky you got a C and not a D.

Diaz: Well, how did I do on the individual questions?

Teacher 3: I guess the best way to describe it is to say you really ran hot and cold. On about half of the questions you did very poor work. And on about half you did pretty good work. Very inconsistent.

Diaz: You know, that doesn't surprise me. I just kept losing my concentration during the exam.

Teacher 3: Well, that explains it.

Diaz: Well, don't you think that the school should do something about the air-conditioning in that room.

Teacher 3: What do you mean?

Diaz: You mean, you don't know about the room. That room is unbelievable. When the air conditioner comes on, it seem like the North Pole. When the machine shuts down, its like the Gobi desert.

Teacher 3: I didn't notice that at the front of the room.

Diaz: Not everybody can sit at the front of the room, Teacher. And, some people maybe had different clothes than I. My friend, for example, had a sweater and a light shirt. She just kept changing. I didn't think of that. I just had on a heavy sweatshirt. I suppose, I could have done like that student at Berkeley, the "naked guy." But, frankly, I didn't really think of that during the exam.

Jones: (Cutting in) Well, the temperature was fine where I was in the room. How did I do, Teacher?

Teacher 3: You got 42 points, Jones. That's a tie for the highest given. That's why you got an A.

Jones: How did I do on the individual questions?

Teacher 3: Well, let's see. You also ran hot and cold too. A couple of your scores were really high. And a couple were pretty low. Real inconsistent.

Jones: That's funny.

Teacher 3: What do you mean?

Jones: Well, our study group scheduled reviews for a couple of successive nights right before the exam. We devoted different nights to different topics. I missed about half of those sessions, the ones, I bet, that covered the stuff that threw me on the exam. Wow, what a lucky break that I did not miss another one, the one that dealt with the topic that you addressed in two questions.

Phill: [Interrupting] And me? How did I do?

Teacher 3: Well, Phill, you were just about in the middle of the pack total-wise, with 36 points. That put you in the middle of the pack. Hence, a C+. On the individual questions? You ran hot and cold. And warm. In fact, you got scores all over the map.

Phill: What about my handwriting?

Teacher 3: What are you talking about?

Phill: Well, I've been told that my handwriting is hard to read. So, I try to print answers as much as possible in the blue books. But, a lot of times I forget and slip back into using handwriting.

Teacher 3: Regardless of what you may be thinking, I pay absolutely no attention to handwriting when I grade exams. Handwriting plays no role in the grading.

Smith: [Cutting in]. You know, Teacher 3, that reminds me of something that I wanted to ask. Do you mind?

Teacher 3: Well, go ahead. We'll see.

Smith: You know how rumors go around, and I don't like to give them any credence, but let me ask you this. I've heard that you got a call from the <u>Harvard Law</u>

Review while you were grading the exams, a call accepting one of your papers.

Teacher 3: Yah, that's part of what happened.

Smith: I've also heard, however, that two days after that first call, the Harvard people called back and said that they'd made a mistake. They'd mixed up your article with somebody else's and they really meant to publish that other one.

Teacher 3: Yep. You cannot imagine how mad I was. I was tremendously excited for a couple days, and then absolutely heart broken.

Smith: Well, my question is this. When were you grading my paper. Were you grading it when you got the first call, or when you got the second call?

Teacher 3; Don't be ridiculous. Those sorts of things play no role whatsoever in grading.

2. Consistency in Measurement

Anybody who has tried to measure anything knows that the process of measurement itself often produces error. A trusted bathroom scale, for example, might show that we have achieved -- or exceeded -- a target weight. Five seconds later, however, when we step back on that scale for confirmation, the scale displays a different weight, a pound or two higher or lower. Obviously, weight was not gained or lost in those few seconds. Rather, the scale itself produced error in measurement. Likewise, consider what might happen if we used an old, old watch to time class sessions. On warm days, the watch might run a little fast, and on cold days a little slow. Thus, though we might intend the class always to be 55 minutes long, sometimes it actually runs for 54 minutes, and sometimes for 56. Again, what has happened here is clear. The process of measurement itself produced error.

The same thing occurs in connection with educational measurement. Tests themselves produce error. A teacher who allows handwriting to influence her judgment on essay answers, for example, introduces error into the measurement process itself. Likewise, if a particular student just happened not to study a topic that a teacher tests heavily, chance will play a significant role in the score that that student gets on a test. And chance, of course, is simply another word for measurement error. Further, if conditions of administration of a test -- air conditioning, for example -- play a role in the performance of some students, or many, error will necessarily infect test scores.

Not surprisingly, measurement experts have developed methods for measuring the amount of error that a measuring instrument itself produces. Measuring instruments that produce very little error are said to be high in "reliability." They then get number ratings that approach 1.00. (1.00 indicates perfect reliability.) Conversely, measuring instruments that produce a lot of error are said to be low in reliability and get number rating that approach 0.00. (0.00 indicates that an instrument measures nothing but error or chance.)

Consider reliability ratings for two different clocks. The old watch might have a reliability rating of, say, .8. This means, roughly, that this instrument pretty much of the time produces measurements that are pretty much alike. Conversely, a $10 million dollar clock that scientists use in connection with particle physics experiments might have a reliability rating of, say, .99999999. This means that this instrument virtually always produces measurements that are virtually identical.

Educational statisticians have developed a number of ways to measure the reliability of tests. Though the technical reasons for why these techniques work are beyond the scope of the present analysis, the techniques themselves -- at least some of them -- are relatively easy for non-experts to use. Teachers who wish to calculate the reliability of essay tests, for example, can simply plug the pertinent numbers from the tests that they give into the "co-efficient alpha" formula.[18] When the calculations are completed, the reliability of that test is described.

$$\text{Reliability} = \frac{\text{No. of Questions}}{1 - (\text{No. of Questions})} * 1 - \frac{\text{Sum of Individual Item Variances}}{\text{Variance of Total Scores}}$$

At first glance, this formula looks daunting. In fact, however, teachers who have access to a computerized spread sheet can very, very quickly learn how to use is. Teachers using this formula divide the number of essay questions on the test by (1 minus the number of essay questions on the test). Then teachers must multiply the resulting number by a number that is 1 minus (the sum of the "variances" of the scores of the different questions on the test divided by the variance of the total scores on the test). Spread sheets, incidentally, can instantaneously calculate the variance of a set of numbers.

The easiest-to-use formula for calculating the reliability of objective-type tests is called the "Kuder-Richardson Formula 21."[19]

$$\frac{\text{No. of Items}}{} * 1 - \frac{\text{Test Mean} * (\text{No. of Items} - \text{Test Mean})}{}$$

(No. of Items) - 1   No. of Items * (Variance of Test Scores)

Again, at least superficially, this formula looks daunting. But, again, teachers who have access to spread sheets can quickly master its use. Teachers using this formula simply take the number of questions on the test and divide that number by the number of questions minus 1. Then these teacher must multiply the resulting number by a number that is 1 minus (the test mean times (the number of questions minus the test mean) divided by the number of items times the variance of the total test scores.

Once reliability ratings for tests are calculated, the rest of the analysis is straight-forward. Although no hard and fast rules exist among experts in educational measurement regarding the degree of reliability that tests must display, some tentative standards have emerged over time. First, if tests are used to make judgments about groups of students, or if tests are going to be used as one of several factors contributing to a total score, then reliability ratings of somewhere between .50 and .60. are generally considered adequate.[20] (A rating of 1.00, it should be recalled shows perfection, while a rating of 0.00 shows pure chance.) However, if tests are to be the only thing that used to generate a score for individual students, then such tests should have reliability ratings of at least .85.[21]

Teacher 3's test now deserves further consideration. Teacher 3, like most law school teachers, uses a single test to determine grades for individual students. Thus, her test ought to have a reliability rating of at least .85. Unfortunately, however, her test nowhere near approaches that degree of reliability. Calculation of reliability for that test -- using the co-efficient alpha formula -- produces a reliability rating of .21. This means, of course, that this test is very, very low in reliability. In other words, if the students who took this test were to take it again, and if no "learning effects" occurred, those students probably would get widely different grades on the second taking. The reliability of this test, in short, is not only far, far below the generally accepted figure for tests that will be used as the sole determinant of individual students' grades, it also is significantly below accepted standards for tests that will make up only part of students' grades.

A bizarre joke about grading sometimes makes the rounds in law schools. Some teachers, it is said, grade exams simply by throwing the blue books down a set of stairs. Then, it is said, these teachers assign grades in light of the different steps upon which the blue books fall. A sad, sad fact must now be noted. The "stair method" of grading has a reliability rating of 0.00. In other words, pure chance is at work. But, as just noted, Teacher 3's test has a reliability rating of .21 Thus, not very much more than chance was involved in that test.

3. Real World Data

What then of the real world. As noted earlier, in connection with the preparation of the present analysis, 13 sets of grades were studied. Several comments about methodology, however, must initially be made. First, for some of the objective-only tests, reliability ratings were calculated using several different techniques; techniques that sometimes produce different ratings. All calculated ratings are listed. Second, technical problems made it impossible to calculate a reliability rating for two of the sets of grades submitted. Had it been possible to calculate reliability of these tests, however, technical reasons suggest that the ratings would have been quite low. Third, for technical reasons and because of insufficient data, the reliability of the separate parts of combination-type tests generally could not generally be calculated. Fourth, all of the teachers who administered combination-type tests here intended the <u>total</u> score on the objective-type questions to be the equivalent of a <u>single</u> score on the essay-type questions. Hence, when reliability for these combination-type tests was calculated, that same notion was employed.

1. Objective-Type Questions Only: <u>Single Test</u> determined course grade:

Teacher M:	.59 (co-efficient alpha);	.47 (split halves)
Teacher L:	.53 (co-efficient alpha);	.37 (split halves)
Teacher K:	.73 (co-efficient alpha);	.79 (split halves)

2. Objective-Type Questions Only: <u>Several Tests</u> combined to generate course grade:

Teacher J-#1:	.75 (KR-20)
Teacher J-#2:	.72 (KR-20)
Teacher J-#3:	.75 (KR-20)

3. Combination-Type Tests (Essays + Objectives): <u>Single Test</u> determined course grade:

Teacher E:	Impossible to calculate; technical reasons suggest low reliability
Teacher F:	.4 (co-efficient alpha)
Teacher H:	.52 (co-efficient alpha)
Teacher I:	.46 (co-efficient alpha)
Teacher G:	Impossible to calculate; technical reasons suggest low reliability

4. Essay-Type Questions Only: <u>Single Test</u> determined course grade.

Teacher B:	.3 (co-efficient alpha)
Teacher C:	.44 (co-efficient alpha)

Teacher D: .4 (co-efficient alpha)
Teacher A: .85 (co-efficient alpha)

One thing should now immediately be obvious. Although 13 sets of data were analyzed, only two of the teachers involved in this study clearly assigned grades in a manner that would satisfy generally accepted standards regarding test reliability. And only one additional teacher came close. Teacher J assigned grades in light of performance on three different tests, each of which had a reliability rating of more than .70. Since experts generally agree that tests that will be combined to generate grades are sufficiently reliable if they have ratings of around .50, Teacher J surely met that standard. Second, Teacher E assigned grades in light of performance on a single exam. But, that single exam had a reliability rating of .85. Since experts generally agree that tests that will be the sole determinant of individual students' grade should have reliability ratings of at least .85, Teacher J clearly meets the pertinent standards. Finally, Teacher K assigned grades in light of performance on a test that had a reliability rating -- according to one method of calculation -- that was .79. Thus, Teacher K came quite close to meeting the pertinent standards. All of the other teachers, however, assigned grades in light of performance on tests that measurement experts would universally agree were insufficiently reliable for that purpose.

4. The Bar Exam, Reprise

As noted earlier, one very important similarity exists between the bar exam and the exams that law students take in their regular classes. In both of these situations, performance on a single test is used to make an important educational decision about an individual person. Again, therefore, it seems appropriate to conclude this discussion of reliability issues in law school classroom testing with a brief comment about the bar exam.

Experts know that subjectivity in grading decreases the reliability of tests. Thus, essay type tests -- which are graded in a subjective fashion -- tend to be less reliable than objective-type tests. This notion, therefore, calls into serious question the reliability of the bar exam, or, better said, the reliability of the essay portion of the bar exam. Since this portion of this exam is graded subjectively, reliability is likely to be low.

Not surprisingly, early data regarding the bar exam supports this notion. Klein describes, for example, a study of the California bar examination that revealed that different graders of the same exam answers agreed regarding whether those answers should pass or fail only 67% of the time.[22] Thus, the inter-grader reliability of the scores for these essays was shockingly low. This same study, however, produced even more shocking news. When the same graders were asked to grade the same papers on different occasions, those graders agreed with themselves only 75% of the time. Later studies confirm the existence of this problem. For example, Klein notes that analysis of essay examinations in three states revealed reliability ratings in the low .70's.[23] (Experts generally agree, it should be recalled, that reliability ratings should be in the .85 range if tests are to be the sole determinants of important matters for individual people.) And, Gorfinkle and Klein conducted a study in which they wrote two different answers for the same essay question. The answers were substantively identical. However, one of the answers was significantly longer than the other. Trained bar examiners were asked to grade these two answers and to ignore things like spelling, length or answer and grammar. Nevertheless, the bar examiners consistently gave the longer answers the better score.[24]

It hardly need be said that bar examiners have gone to great lengths to reduce reliability problems with the bar exam. For one thing, most bar examiners now use the Multi-State Bar Exam, an exam that is graded objectively. Objectively graded exams, of course, tend to be more significantly more reliable than subjectively graded exams.[25] In addition, bar examiners now generally use very sophisticated techniques in connection with the grading of essay-type questions, techniques that produce surprisingly high degrees of inter-grader consistency.[26] Obviously, techniques that significantly decrease scorer variability significantly increase test reliability because score variability is one of the big problems on essay type tests.[27] Finally, most bar exams now have two distinct parts, an objective portion and an essay portion. Statistical analysis reveals that the performance of individuals on these different parts of these exams is remarkably consistent. (Students who do well on one part tend to do well on the other, and vice versa.) This consistency, in turn, suggests a high level of consistency in measurement for the exam as a whole.

The short of it is this. Though bar exam perhaps at one time suffered from serious reliability problems, a high likelihood exists that most such exams now are reliable enough to use them as the basis for making important decisions about individual students. Further, and perhaps more significantly, the fact that bar examiners have in recent years taken major steps to increase the reliability of bar exams suggests that the notion of reliability is not some purely academic exercise, something that classroom teachers need not address.

C. The Standard Error of Measurement

Interestingly, the notion of reliability does something other than help teachers (and schools) decide just how consistent the results are likely to be of tests that teachers use to generate the entire grade for a course that students will obtain. Reliability data also plays a role in determining how much "error" must be taken into account when scores from individual tests are evaluated.

This idea of accounting for error, in turn, is generally called the "standard error of measurement" in a test.[28]

1. Teacher A

Note quickly before the following "grade conference from Hell" that a major difference exists between the following conference and the ones present earlier. The following grade conference rests on data from a real teacher's course. In short, though conference itself described below is hypothetical, the grades discussed in it are real.

Student 29: I really have only one question, Teacher A. I got 84 points, and a B+. What was the cut off in point scores for A's? How many more points did I need to get an A?

Teacher A: Well, let's see. Student 7 got 85 points and I gave her an A. So, you missed an A by 1 point. Too bad.

Student 37: What about me? I got 62 points and a C. What was the cut off for C+'s?

Teacher A: Let me check. Well, I'll be darned. Student 15 got 63 and a C+. You missed a C+ by a single point.

Student 37: Are you sure that a one point score difference between Student 15 and me justified a different letter grade for the course for us.

Teacher A: Yes, of course. That's what the grades indicate.

Student 29: [Cutting in] Well, I'm sorry to be a pest. But, Student 7 made the law review because of that A, and I missed out on law review because of that B+. And you know how incredibly big a deal law review participation is. Are you sure that a single point difference reveals a real difference in our work?

Teacher A: I'm sorry. You just caught a bad break.

Johns: You know, it's funny that you say it that way. I think that is exactly what happened. I think that I just caught a bad break. I think that a single point difference on your test is just a function of luck. I think that such a point difference doesn't really indicate any difference in performance at all.

Teacher A: Don't be ridiculous. The numbers don't lie.

2. Measurement Error

Except in the most extraordinary circumstances, measurement instruments, including tests, contain at least some error. This is so even if the instruments have relatively high reliability ratings. Admittedly, the amount of error in an instrument with high reliability is going to be substantially less than the amount of error in an instrument with low reliability, but some error is likely to exist even reliable instruments.

Because experts in measurement know that most instruments, including highly reliable ones, contain at least some error, such experts have developed a number of methods for determining the amount of error that exists in any instrument. This amount of error, in turn, is generally called the "standard error of measurement" in an instrument. In effect, the standard error of measurement is the margin of error that must be considered when measurements are evaluated.

The work of public opinion pollsters provides perhaps the best known example of this concept at work. Pollsters know that measurements of public opinion always contain some error. Thus, when pollsters do their work, they first calculate the likely amount of error in their polls. Then, when they report results, they also report that margin of error. A pollster might report, for example, that 47% percent of the people in the United States approve of the President's actions in a certain context. This pollster might also report in this context, however, that a margin of error of 3 points exists in the poll used. What the pollster means, of course, is this. The reported approval rating of 47% is not necessarily the "true" approval rating. Rather, the true approval rating is likely to be in a "band" of numbers that extends from three points above the reported rating to three points below the reported rating. In other words, the "true" approval rating in this instance could be as high as 50% or as low 44%.

Educational measurement experts do a similar things when they report scores on tests. First, they determine the amount of error in a test itself. Then when they report scores on that test, they also report the margin of error in the test. A testing expert, for example, might report that a particular student obtained a score of 47 on a test. That expert might also report, however, that the test itself has a measurement error of 3 points. What this would mean, of course, is this. The reported score of 47 points is not necessarily the student's "true" score at all. Rather, the true score is likely to be in a band of numbers that extends from three points above the reported score to three points below the reported score. In other words, in the present example, the student's true score could be as high as 50 or as low as 44.

As with many other aspects of educational measurement, no hard and fast rules exist regarding the use that must be made of the standard error of measurement notion. Nevertheless, educational measurement experts generally agree on one thing. Reported scores cannot be considered actually different from each other unless the error bands around those the different scores do not overlap.[29] In other words, scores are only truly different if the plus side of one person's error band does not overlap with the minus side of another person's error band. In short, differences in scores of less than twice the standard error of measurement of a test cannot be assumed to reveal differences in performance.[30]

The standard error of measurement for a test can easily be calculated, at least once the reliability of the test has been calculated. Indeed, the formula is perfectly straight forward.

$$\text{Error} = \sqrt{\text{Variance of Total Scores} * (1 - \text{Reliability})}$$

In words, the standard error of measurement of a test is the square root of the variance of the total scores on the test times (1 minus the reliability of the test). Again, recall that spread sheets can instantaneously calculate the variance of a set of scores.

All of this brings this analysis back to Teacher A's exam, a real exam it should be recalled, given to real students, generating real grades. Calculation establishes that the standard error of measurement in that exam is 6 points. Consider, therefore, Teacher A's grades and scores.

Figure 8

I.D.	PTS	GRD
28	91	A
1	87	A
7	85	A
29	84	B+
25	83	B+
24	83	B+
13	80	B+
22	76	B
2	76	B
17	74	B
4	70	C+
35	69	C+
33	68	C+
5	67	C+
20	67	C+
26	66	C+
23	64	C+
15	63	C+
18	62	C
37	62	C
11	61	C
19	60	C
31	59	C
9	58	C
10	58	C
34	58	C
30	58	C
16	57	C
32	55	C
6	53	D+
21	52	D+
39	51	D+
38	48	D+
3	46	D
36	44	D
8	42	D
40	36	D
27	34	F
12	33	F
14	30	F
41	25	F

Teacher A, it should now be clear, does not deserve his name. Consider, for example, Student 7, the student who received the lowest A that this teacher gave. First, measurement experts would agree that 6 points must be subtracted from this student's reported score of 85 to account for the negative part of his own error band. Then, these experts would agree that another six points must be subtracted from this reported score to account for the positive part of the error bands of students who obtained lower reported scores. In short, twelve points must be subtracted from Student 7's reported score of 85 before any kind of assurance can be had that differences in performance really exist. Consider, however, what Teacher a actually did. Student 29, with a reported score only <u>one point</u> lower than 7's reported score, got a B+. This, unquestionably, was a mistake. Further, Students 2 and 22, with reported scores just 11 points lower than 7's, got B's. This may well have been a mistake.

And consider what happened at the bottom of Teacher A's class. Student 12, with 33 points, got the highest F that Teacher A gave and Student 40, with 36 points, got the lowest D that this teacher gave. Error measurement analysis reveals, however, that assignment of different grades to these different students simply was not justified. Differences in reported scores of 3 points simply cannot be considered to reveal differences in actual performance.

3. Real World Data

What then of real world data?

Recall again that in connection with the preparation of the present analysis, 13 sets of real grades were examined. Standard error of measurement analysis demonstrated a startling fact. Every single one of those sets displayed serious standard error problems. In other words, every set of grades examined showed that teachers gave different letter grades to students even though the test performances of those different students simply could not realistically be considered different.

Consider, just by way of additional example, the scores and grades of Professor C. C gave an exam made up of a series of essay-type questions to approximately 50 students. The top scoring student got 162 points on this test, and the bottom student got 120 points. And the rest of the students got just about every number of points in between those two extremes. The standard error of measurement for this test was 8 points. In other words, on this test students' "true" scores could be anywhere between a number 8 points above their reported scores and a number 8 points below their reported scores. But, C repeatedly gave different grades to students whose reported scores differed by only one or two points.

An extraordinarily important point must now be made. The standard law school examination system -- a single final exam made up predominantly of essay-type questions -- almost necessarily will produce standard error problems. This is so, of course, because that system almost inevitably will involve use of tests that contain fairly large error components. And it is so because this system will tend to produce sets of scores that move in small increments for top scores to bottom scores. These facts suggest, in turn, that that traditional system almost necessarily will cause many law school teachers to make serious measurement error in connection with the grades they assign.

Sadly, no easy solutions exist to this problem. Two possible approaches, however, come to mind. First, law teachers could move to a grading system that involved multiple components, several tests, for example, or several tests and papers and quizzes. Since measurement errors tend to cancel themselves out when multiple components make up grades, this approach would essentially solve standard error of measurement problems. Second, for teachers who do not wish to use multiple components -- and, frankly, few law teachers are likely to move to that approach -- error band calculations could be built into the grading scale itself. Teachers, for example, could openly admit to their students that half letter grade differences in scores almost certainly do not reflect real differences in performance. A B+, these teachers might candidly admit, could just as easily have been an A or a B. Once teachers made these admissions, they could simply make sure that appropriate error bands separated students with full letter grade differences in scores. And, these teachers could just let the fickle hand of luck cancel out standard error problems for the entire curriculum. "You might have gotten a half letter grade lower in my class than you deserved," these teachers might tell disgruntled students. "But, chances are that in other classes you got a half letter grade better than you deserved."

4. Legal Implications of Standard Error Issues

Interestingly, standard error of measurement problems have not for the most part been explicitly addressed in connection with the bar exam. Thus, sadly, no direct data from that exam can be added to this discussion of law school grading. Nevertheless, several points can yet be made. First, since the reliability of both the essay portion and the objective portion of the bar exam seems to be quite high, standard error problems should not be much of an issue on the bar exam. The standard error of measurement in a test, it should be recalled, is inversely proportional to the reliability of that test. In other words, if a test is high in reliability, its standard error of measurement is likely to be low. Second, as noted repeatedly, when testers combine scores from different exams, measurement errors tend to cancel themselves out. This is so, of course, because good luck on one exam is likely to be balanced by bad luck on another. Since the bar exam in most jurisdictions is made up of what clearly are two different exams -- the essay portion and the objective portion -- overall scores on that exam are less likely to be affected by error problems than scores on single exams.

One last point must yet be made in this context. Although law school teachers for the most part seem to know nothing about the standard error of measurement, and although bar examiners do not seem regularly to discuss this issue, the standard error of measurement is not simply an educational abstraction, something that nobody other than statisticians addresses. Rather, measurement error is a concept that has real world vitality, even in the courts. Craik v. Minnesota State University Board, for example, contains an extended discussion of this notion generally, an explanation of how it works.[31]

A much more significant case in this context is Georgia State Conference of Branches of NAACP v. State of Georgia.[32] This case involved a civil rights action brought by a group of African-

American school children. One aspect of that case involved the method used to assign children to "special" education classes. Not surprisingly, that method included the use of an IQ test. Students who scored below a certain point on that test, and who fit other criteria, were assigned to special classes. Conversely, children who scored above that point, and who otherwise qualified, went to regular classes. The question then arose as to how precise scores on that IQ test were. Not very precise at all, it turns out, as the district court ruled. The appellate court affirmed in language that seems to have great applicability in the present circumstances.

> The [district] court's construction of the I.Q. score regulation was based on the factual finding that including a standard error of measurement is sound and that the range suggested by the AAMD [American Association on Mental Deficiency] is professionally desirable. Although the state regulation does not explicitly refer to the standard error of measurement, a number of experts testified at the trial that inclusion of this amount of flexibility in considering I.Q. scores is necessary for a meaningful cutoff. See e.g. Record, Vol. 49 at 2126-28,2135 (testimony of Dr. Kicklighter) (standard error of measurement is an intrinsic part of I.Q. test). Furthermore, these is substantial evidence in the record supporting the view that the AAMD guidelines are acceptable professional tools.

Again, the overall point is clear. Measurement notions that most law school teachers might think of as obscure technicalities, in fact have a real world, judicial world existence. Real people's lives are affected by these notions, and the courts are aware of that fact. Or, better said, at least in some contexts courts are aware of that fact.

III. Grading Decisions and the Courts

If the empirical data just described presents a picture that is at all representative, and if empirical data discussed earlier also presents a representative picture, then a substantial likelihood exists that many law school teachers make a number of serious measurement errors in connection with the grading of law school exams and the assigning of letter grades pursuant to scores received on those exams. The question thus arises as to what, if anything, law students can do about these errors. Particularly, the question arises as to whether law students might find a sympathetic ear in court.

Two separate but parallel sets of cases and ideas address the legal rights of students who disagree with grading decisions. One of those sets, a set that is very well known to teachers, deals with the notion of "academic challenges." The other set of cases, a set that seems to be unknown to most teachers, involves what is sometimes referred to as "high stakes" testing. Interestingly, these two sets of cases seem to point in opposite directions when it comes to students' rights.

A. "Academic Challenge" Cases

"Susan M" is one of the most recent students to file what is sometime called an "academic challenge" lawsuit.[33] Susan, like other students who have filed these lawsuits, claimed that the grades she received in individual classes should be changed, and an expulsion decision made in light of classroom grades should be expunged. Susan M flunked out of law school in the late 1980's. Unfortunately for Susan, and students in comparable cases, a widely accepted rule exists for dealing with this sort of situation. This rule, described most vividly in the United States Supreme Court case of Board of Curators, University of Missouri v. Horowitz,[34] Horowitz states in no uncertain terms that teachers and schools have an enormous amount of discretion when it comes to grading / expulsion decisions. Indeed, this rule states that absent the most extraordinary circumstances, students simply have no judicial recourse whatsoever when it comes to grading / expulsion disputes.[35]

Seemingly, the Horowitz rule makes it pointless for students to file academic challenge cases. And, frankly, the cases preceding and following Horowitz confirm that idea. Thus, when it comes to classroom grading / expulsion decisions, the protective arm of discretion provides teachers and schools with almost complete protection. Having said that, however, an important caveat must be made. Two small but potentially significant possible exceptions to the Horowitz rule seem to exist.

In its most recent discussion of academic challenge issues, the Supreme Court specifically noted that at least one kind of situation exists in which students might prevail in an academic challenge case. The judgment of a teacher or a school can be overturned, the Court noted in Regents of the University of Michigan v. Ewing, if that judgement constitutes a "substantial departure from accepted academic norms."[36] What the just quoted phrase from Ewing means, of course, is not at all clear. Probably, however, this phrase reflects the Court's concern that unprincipled teachers and schools might try to use the cloak of grading discretion to protect themselves from well-founded claims of deprivation of civil rights. It is possible, for example, that an unprincipled teacher or a group of unprincipled teachers might dismiss an African-American student from school despite that fact that his academic work is no worse than that of white students.[37] Likewise, it is possible that an unprincipled teacher might fail a student who did acceptable academic work because that student made controversial political statements. Since in both of these situations, the teacher's grading judgment constitutes a substantial departure from accepted academic norms, in both of these situations an academic challenge might succeed.

The second possible exception to the Horowitz rule grows out of Maitland v. Wayne State University Medical School,[38] a case that is just about the only modern academic challenge case in which a student prevailed. Several problems occurred in connection with the "comprehensive" pass / fail test that Maitland took at the end of his second year of medical school. First, proctors of this test employed different procedures in the two different rooms in which the test was administered. Second, some sort of "error...in the grading process" initially occurred. When that error was corrected, Maitland got 20 points more than his original score. Third, the passing score for the "retake" exam was set at a higher figure than the passing score on the original exam. (If the passing score on the retake had been set at the same figure as the passing score on the original exam, Maitland's retake performance would have been a pass.)

Interestingly, something that herein is called "measurement error" played a critical role in the student's success in Maitland. During deliberations regarding grades given on the exam just described, the Chair of the pertinent faculty committee asked for a statistical analysis of the scores obtained in the two different rooms. The Chair sought that analysis, of course, to determine whether the different procedures followed in the different rooms affected scores obtained on the exam. Unfortunately, however, the Committee chose not to wait for the results of this statistical analysis before deciding Maitland's appeal. Thus, the committee ruled twice against Maitland prior to receipt of that analysis. This proved to be a fatal mistake. Though the statistical analysis ultimately showed that the different procedures used in the different rooms did not affect scores obtained, the court concluded that the Committee's failure to wait for the results of that analysis was educationally irresponsible. Hence, Maitland's challenge was accepted.

Note carefully now an important point. The statistical issue just described was only one of the reasons that the student prevailed in Maitland. Equally important in that case was the fact that the court also thought that Maitland's school had erred when it allowed people with lower scores than Maitland on the original test to retake the exam without filing formal appeals. This second reason for decision no longer holds up. Ewing, which was decided after Maitland, categorically states that the fact that Ewing's school allowed students with lower grades and scores than Ewing to continue in school despite expelling Ewing made no difference. Schools and teachers, the Court noted in Ewing, could weigh all sorts of intangible things when making grading decisions and still not fear judicial intervention.

The short of it is this. Given the fact that one of the grounds for decision in Maitland no longer is sound, it is entirely possible that Maitland itself would be decided differently now than it was before Ewing was decided. Nevertheless, the statistical analysis / measurement error point addressed in Maitland perhaps still is sound. Perhaps, in other words, the failure to consider statistical evidence of measurement error associated with tests is a "substantial departure from accepted academic norms."

B. "High Stakes" Testing

As noted earlier, academic challenge cases generally involve complaints that students make about the grades that they receive from classroom teachers, or about expulsion decisions that university officials make in light of classroom grades. Academic challenges like this almost always fail. Maitland, however, did not really involve classroom grades, nor an expulsion decision based on classroom grades. Rather, Maitland involved the grade received on a single, extraordinarily important test. Maitland, therefore, is not really an "academic challenge" case. Rather, it is something like a "high stakes" testing case.[39]

Debra P, like Susan M, was negatively affected by a testing decision.[40] Debra, however, was not concerned about classroom grades, nor an expulsion decision. Rather, Debra was concerned about a "minimum competency" test that the State of Florida had decided had to be past before high school diplomas could be granted. When Debra failed this test, she sued, claiming, among other things, that this test invidiously discriminated against members of minority groups.

High stakes testing, which was what was involved in Debra P's case, occurs when the score or grade that individuals obtain on a single test has enormous consequences for the individuals involved. The minimum competency test Debra P failed, of course, was a high stakes test. Failure to pass it meant that students did not get high school diplomas. The LSAT also is a high stakes test. Scores that individuals obtain on this single test can have life-changing impact. A very high score on this test, after all, might lead to admission to an "elite" law school. Conversely, a slightly lower score on this single test might limit admission to only a "national" school. And, lower scores yet might limit admission to "regional" or even "local" law schools. The LSAT, of course, is not the only example of a high stakes test. In fact, countless high stakes tests exist. Most tests associated with admission to educational institutions, for example, are high stakes tests. Thus, the SAT is a high stakes test, as is the GRE (Graduate Record Exam) and the MCAT (Medical College Admission Exam). Further, licensing and certification exams can be high stakes tests. Bar exam failure, after all, can have life changing consequences, as can failure of a teacher certification exam. Finally, tests that individual employers might use to screen potential employees, or tests that individual employers might use in connection with promotion practices, can be high stakes tests. The scores that people get on tests given to potential police or fire officers, for example, can have a life changing impact.

Countless people other than Debra P have filed law suits in connection with high stakes testing. For example, teachers who have failed re-certification exams have repeatedly claimed in court that the exams involved discriminated against members of minority groups. These kinds of law suits, obviously, are high stakes testing cases.[41] Or, again by way of example, a group of female students who had lost out on a major scholarship claims in court that the SAT discriminates against female students.[42] This again was a high stakes testing case. In addition, people who have scored poorly on employment examinations have repeatedly sued to set the test results aside.[43] Finally, students who have been assigned to special educational programs, or not assigned to such programs, because of performance on individual tests, have often sued to set the results of these tests aside.[44] Again, obviously, high stakes tests were involved in these cases.

Interestingly, courts that have dealt with high stakes testing claims[45] -- including the United States Supreme Court[46] -- have done exactly the opposite of what courts have done in the academic challenge cases. As noted earlier, in academic challenge cases courts routinely place the burden of proving substantial violations of generally accepted educational norms squarely on the student plaintiffs. Since student plaintiffs can virtually never do this, such plaintiffs almost always lose these cases. Conversely, in high stakes testing situations, the courts routinely place the burden of proving compliance with generally accepted educational norms on the proponents of a test or scoring methodology. In other words, in high stakes testing cases, the test-graders rather than the test-takers have the burden of proof.

Many cases illustrate this point, with Debra P's being perhaps the most important.[47] Of particular interest in the present context, however, is a 1981 case, Delgado v. McTighe[48] In this case, the plaintiff insisted that the Multi-State Bar exam, which he had failed, could not be used to make important decisions about individual students. In other words, in this case, the plaintiff raised a high-stakes testing issue. Once this was done, the burden shifted to the testers. But, the testers met the burden. The bar exam, the court ruled, was sufficiently reliabile for use in a high stakes testing situation.

It would be far too time-consuming to attempt to describe herein all of the things that courts have required high stakes testers to do in order to show the soundness of the tests involved. Nevertheless, two quick summary points can be made. First, several commentators have suggested in recent years that perhaps the best thing that high stakes testers can do is follow the "Standards for Educational and Psychological Testing," standards created by the American Psychological Association, the American Educational Research Association and the National Council on Measurement in Education. These standards reflect the best current wisdom regarding educational measurement issues.[49] Second, one commentator who has written extensively about the legal implications of high stakes testing -- S. E. Phillips -- concluded a recent essay with a list of recommendations for high stakes testers.[50] Several of these recommendations deserve quotation. High stakes testers, Phillips insists, should engage in a process that includes "collecting appropriate validity and reliability evidence, constructing and evaluating tasks according to professional standards, [and] setting passing scores based on professionally-acceptable methodology...." Later, Phillips notes that high stakes testers should also "obtain (and follow) the advice of a technical advisory committee composed of nationally recognized experts in psychometrics who have had experience with the particular testing application....."

Note carefully now two important final points about high stakes testing cases. First, since high stakes testing cases have in the past invariably involved claims of civil rights deprivations, it could perhaps be argued that the existence of civil rights claims is a pre-requisite to success these cases. In other words, it is possible that individuals can succeed in high stakes testing cases only if those individuals can show two things: First, successful claimants in these cases perhaps must be able to show that a problem exists in connection with an educational measurement issue. Second, successful claimants in these cases perhaps must be able to show that that measurement problem leads to a civil rights deprivation.

Careful analysis reveals that this in fact should not be the case. Consider the following. Assume that a teenage prodigy, Jeannie Genius, takes the SAT and scores 799 points out of 800. Assume also, however, that an Jeannie publishes a description of her high school science project in journal, Science. In this article, Jeannie demonstrates that the question that she missed on the SAT was mis-scored. Her answer, which was marked wrong, was actually the correct answer. In other words, in this article, Jeannie demonstrated that the Educational Testing Service (ETS), which drafts and scores the SAT, goofed. Finally, assume that ETS has now dug in its heels and refuses to change Jeannie's score. Thereafter, Jeannie sues. Question: Can Jeannie succeed only if she proves a violation of her civil rights? Must she prove, for example, that the SAT invidiously discriminates against geniuses? Or can Jeannie win simply by proving that ETS was wrong? Surely the latter. But if the latter, then high stakes testing cases generally need not necessarily include assertions of civil rights violations.

The point, of course, is this. Though claims of civil rights deprivations might be useful in connection with high stakes testing cases -- perhaps by helping claimants obtain statutory authority for their claims, the existence of educational measurement error should by itself be sufficient to gain a hearing in high stakes testing situations. Admittedly, deprivations of civil rights are serious wrongs. But, they are not the only kinds of wrongs that exist.

The second thing that must yet be said about high stakes testing cases involves an issue that was addressed in the Maitland case. Recall that in Maitland, which is just about the only example of an academic challenge case in which a student prevailed, the court was particularly bothered by the Committee's failure to wait for the results of a statistical analysis of the test results involved. Statistical analysis of test results, the judge seemed to think, is not something that interferes inordinately with the discretion of teachers and university officials. In other words, mere numbers and calculations do not seem to be things that are subject to discretion. The same thing can be said about most of the high stakes testing cases. Virtually all of these cases involve careful analysis of statistical evaluations of the tests involved. In other words, number crunching plays an important role in these cases.

C. Law School Grading: What is it?

All of this, finally, brings this analysis to a very brief comment about law school grading. Everyone familiar with legal education knows that most law school teachers base the entire grade for their courses on students' performance on a single final exam. Everyone familiar with education generally also knows that teachers in virtually all other kinds of institutions base the grades for their courses on a number of different tests, or on a number of tests and papers, or on a number of tests and papers and quizzes. These facts raise a series of interesting questions. Are the classroom grades that teachers in law school give akin to the grades that teachers in other kinds of educational institutions give? Alternatively, are the tests given in most law school classes "high stakes" tests. If the former, then law students who disagree with the class grades that they receive and law students who disagree with expulsion orders based on classroom grades, must deal with the Horowitz rule. Thus, for all practical purposes, such students should not even consider filing suit. But, if law school grading involves high stakes testing, then, perhaps, law students have considerable judicial recourse regarding classroom grades and expulsion decisions.

IV. Conclusion

A couple of summing up points must now be made. Some critics of the foregoing might note that few teachers actually do the kinds of calculations described herein. Few teachers, these critics might insist, actually calculate Z-scores -- and thus avoid "weighting" problems. Further, critics of the foregoing might insist that few teachers actually calculate the "reliability" of their classroom tests, and thus few teachers actually know whether they should or should not assign grades in light of those tests. Finally, critics might argue that few teachers actually calculate the standard error of measurement in their tests. Since most teachers do not do these things, these critics will insist, the failure to do them cannot possible constitute a substantial deviation from accepted academic norms.

Two things must quickly be noted about these criticisms. First, the failure by many teachers to do certain kinds of things surely is evidence of accepted academic norms. Thus, if law school grading cases are "academic challenge" cases, then, perhaps, students will lose. But, as noted repeatedly herein, law school grading cases may actually be "high stakes" testing cases rather than academic challenge cases. If this is true, then the burden of proving compliance with accepted academic norms shifts to the proponents of the test. Second, even if law school grading disputes are academic challenge cases -- and thus subject to the Horowitz / Ewing rule -- students still might have a chance. this is so, in turn, because educational practices outside of the law schools may not establish the standard of review for law school grading.

Consider again the criticism just noted, namely, the assertion that many teachers do not do the kinds of statistical analysis described herein. Even if that is true, it matters nothing. A significant difference exists between the grading practices of law school teachers and the grading practices of virtually all other teachers, a difference that plays a powerful role here. Law school teachers generally assign grades in light of performance on a single exam. Conversely, virtually all other kinds of assign grades in light of several tests, or several tests and some papers, or tests and papers and quizzes and class participation. In other words, the performance on individual tests has much, much less impact outside of the law schools than in the law schools. This is critically important. As teachers rely on more and more factors in connection with the assignment of grades, the individual impact of measurement errors such as those described herein tends to become less and less important. If a teacher gives grades in light of performance on five exams, for example, it probably matters not at all that none of the exams have reliability ratings of more than, say, .50. Bad luck on one test, after all, will tend to balance out good luck on other tests. Further, as teachers use more and more items in connection with the calculation of grades, differences between the performance of different students will become more and more pronounced. Thus, technical issues regarding error measurement will play less and less a role.

The point is this. Because the tests that students take in law school classes are so commonly used as the sole determinant of important matters for these students, these tests are not really comparable to the tests that students take in other kinds of classes. Rather, law school tests are roughly comparable to the LSAT, or the GRE, or the Multi-State Bar Exam. These other tests, like tests in law school classes, are the sole determinants of important matters. Thus, when grading practices in the law schools are considered, it matters not so much what teachers in other kinds of institutions do. Rather, it matter what is done by people who write and analyze tests like the GRE and the LSAT and the Multi-State Bar Exam. And, not surprisingly, people who work on these kinds of tests subject them to enormously rigorous scrutiny.

One last response to the foregoing criticism must yet be made. As suggested repeatedly herein, calculations of the kind described herein in fact are difficult to do if teachers do not have access to a computerized spread sheet. Thus, since many teachers outside of the law schools simply do not have access to computerized spread sheets, a ready explanation for the failure to do these things exist. Law school teachers, however, have no such excuse. At the present time, many law school teachers have powerful computers on their own desks, or have easy access to such computers in their schools’ libraries. Further, all law school teachers presently have ready access to secretaries who have powerful computers on their desks.

The bottom line -- to use "lawyer-speak" -- is this. Law school teachers who make the kinds of measurement errors described herein make those errors not because they choose to exercise discretion and do different things. And, law school teachers who make these kinds of errors do not make them because they do not have access to appropriate computer equipment. Rather, law school teachers who make these kinds of errors make them because they are lazy or ignorant. And think what kind of defense that would be in court.

These thoughts, in turn, bring this analysis to one last grade conference from Hell. This grade conference does not take place in the stronghold of a teacher, however. Rather, it takes place in a court room.

Students’ Attorney: Just to remind you quickly, Judge, I represent two students. One of them missed out on graduating Number 1 in her class because she got a half letter grade lower score than a classmate in one class. That classmate, incidentally, is now clerking for a judge on the United States Supreme Court. My other client got an F in a course and flunked out of school. Both of these students believe that their grades were significantly affected by "measurement error."

Law School Attorney: Judge, I'm sure I need not remind you that all grading decisions made by classroom teachers involve discretion. Just read the "academic challenge" cases. Read *Horowitz*. Read *Ewing*. Those cases clearly require this action to be dismissed outright.

Students' Attorney: Not so fast, Your Honor. Two points. First, the "high stakes" testing cases require testers to prove that tests are accurate measures. And law school grades, at least those given when the teacher uses only one test, are a type of or hybrid form of high stakes testing. Second, even if *Horowitz* and *Ewing* apply, the present situation -- and I think they do not -- these situations do not involve any kind of discretion. We're not saying here that the teachers should have given the students additional points on an essay, or that the number of letter grades given was wrong. Those kinds of things, which were involved in *Horowitz* and *Ewing* in fact are discretionary and should not be reviewed by courts. What we're saying is that the teachers here made essentially mathematical errors.

Law School's Attorney: Judge, we're talking here about obscure statistical ideas. Nobody pays any attention to this stuff.

Students' Attorney: No, Judge. We're talking here about human lives.

Grade Conferences From Hell: Measurement Error in Law School Grading

Notes

1. Gronlund and Linn explicitly state what many law teachers probably feel. N. Gronlund and R.L. Linn, Measurement and Evaluation in Teaching (6th ed. 1990) at 470: "No major educational decision should ever be based on a test score alone."

2. Numerous books on educational measurement exist. Three books, however, stand out as exceptionally useful to beginners, books that are simultaneously comprehensive but understandable. R.Ebel, and D. Frisbie, Essentials of Educational Measurement (5th Ed. 1991); W. Mehrens and I Lehmann, Measurement and Evaluation in Education and Psychology (4th Ed. 1991); G. Sax, Principles of Educational and Psychological Measurement and Evaluation (3rd ed. 1989). These books are the books upon which the present analysis principally relies.

3. Readers interested in more sophisticated discussions of statistics than the ones contained herein might consult: E.V.Glass and K.D. Hopkins, Statistical Methods in Education and Psychology (2d ed. 1984), or F.J. Gravetter and L. B. Wallnau, Statistics for the Behavioral Sciences (2d ed. 1988).

4. See, e.g. Descy, "Setting Standards and Cut Scores: Where Do We Draw the Pass / Fail Line?, 57(4) Bar Examiner 17 (1988); Kurdys, "Grading Essay Answers: The Issue of Reliability in Essay Scoring," 59(4) Bar Examiner 22 (1990); Lenel, "Issues in Equating and Combining MBE and Essay Scores, 61(2) Bar Examiner 6 (1992); Lenel, "Test Validation: What It Is and How It Should Be Done," 60(3) Bar Examiner 5 (1991); Lenel, The Essay Examination Part III: Grading the Essay Examination," 59(3) Bar Examiner 16 (1990).

5. For discussions of "weighting" issues generally, see Ebel, supra, at 276 et seq; Mehrens, supra, at 491 et seq.; Sax, supra, at 204 et seq, and 539 et seq: See also, P.W. Airasian, Classroom Assessment (1991) at 339-44: N.E. Gronlund and R.L. Linn, Measurement and Evaluation in Teaching (6th ed. 1990), at 437 - 39; K. Hopkins, J. Stanley, and B. Hopkins, Educational and Psychological Measurement and Evaluation (1990), at 331 et seq; W.J. Popham, Modern Educational Measurement: A Practitioner's Guide (2d Ed. 1990) at 378 et seq.

6. For discussions of "weighting" issues generally, see Ebel, supra, at 276 et seq; Mehrens, supra, at 491 et seq.; Sax, supra, at 204 et seq, and 539 et seq. See also, P.W. Airasian, Classroom Assessment (1991) at 339-44: N.E. Gronlund and R.L. Linn, Measurement and Evaluation in Teaching (6th ed. 1990), at 437 - 39; K. Hopkins, J. Stanley, and B. Hopkins, Educational and Psychological Measurement and Evaluation (1990), at 331 et seq;

W.J. Popham, Modern Educational Measurement: A Practitioner's Guide (2d Ed. 1990) at 378 et seq.

7. This example is drawn directly from Gronlund and Linn, supra, at 438.

8. Gronlund and Linn describe an additional method exists that allows teachers to combine scores without risking the weighting problems just described. N. Gronlund and R.L.Linn, Measurement and Evaluation in Teaching at 438-439 (6th Ed. 1990). Regrettably, however, this method only works when two components are to be combined and given equal weight. Teachers who wish to use this third technique must do four things. First, they must determine the range of scores on both of the two components. If, for example, scores on the first component range from 100 to 80, then the range on that component is 20 points. If scores on the second component range from 50 through 10, then the range on that second component is 40 points. Second, these teachers must divide the ranges by each other to generate a "weighting" factor. In this case, therefore, 40 divided by 20 is 2. So, the weighting factor is 2. Third, to equalize the scores, the scores on the component part with the lower range of scores is multiplied by the weighting factor. Fourth, and finally, the teacher then adds up the multiplied score from the one component and the raw score from the other.

9. This technique is suggested by Gronlund and Linn, supra, at 438-39.

10. It hardly need be said that this notion of "standardizing" scores is not widely known to lawyers and legal educators. Nevertheless, this notion has appeared from time to time in materials associated with the law. Merritt and Reskin used this notion, for example, in connection with their analysis of law school employment practices. Merritt and Reskin, "Double Minority: Empirical Evidence of a Double Standard in Law School Hirign of Minority Women," 65 S. Cal. L. Rev. 2299 (1992) at notes 56 - 57. See also, Garcia and Steele, "Mentally Retarded Offenders," 41 Ark. L. Rev. 809 (1988) at note 24. Perhaps the most thorough discussion of this concept in the law literature, however, and surely the most interesting in connection with the present analysi2, involves the Multi-State Bar Examination. Lenel, "Issues in Equating and Combining MBE and Essay Scores," 61(2) Bar Examiner 6 (1992).

11. Standardized scores are discussed in virtually all books on educational measurement. See, e.g. Ebel, supra, at 68

12. For one of countless discussions of this formula, see Ebel, supra, at 68. T-scores are a simple derivation of Z-scores. Instead of assigning the mean a value of zero, as Z-scores do, T-scores assign the mean the value of 50. Thus, a T-score above 50 is a score that is above the mean, and a T-score below 50 is a

score below the mean. T scores, which are useful because they do not contain negative numbers, are calculated by multiplying the pertinent Z-score by 10 and then adding 50 to the resulting number. The formula is: T = ((Z-score) * 10) + 50

13. Klein, "Are Your Test Scores Only Half Safe?" 48(1) Bar Examiner 137 (1979).

14. Id. at 139.

15. Id.

16. Id. at 142. Klein also discusses this issue in other works. See, e.g. Klein, "On Testing V: How to Answer the Critics," 55(1) Bar Examiner 16, 22 - 23 (1986). See also, Lenel, "Issues in Equating and Combining MBE and Essay Scores," 61(2) Bar Examiner 6 (1992).

17. See Ebel, supra, at p. 76 et seq.; Gronlund,supra, at 77 et seq, and 101 et seq; Hopkins, supra, at 113 et seq.; Tuchman, supra, at p. 146-47; Popham, supra at 121 et seq.; Sax, supra, at 279-81.

18. This formula is discussed, among other places, at Ebel, supra, at 85.

19. This formula is discussed, among many other places, at Ebel, supra, 84.

20. See Ebel, supra, at p. 86; N.E. Gronlund and R.L. Linn, Measurement and Evaluation in Teaching, (6th ed. 1990) at 77 et seq, and 101 et seq; Tuchman, supra, at p. 146-47. See also, W. J. Popham, Modern Educational Measurement: A Practitioner's Guide (2d ed. 1990) at 121 et seq.; Sax, supra, at 279-81.

21. Ebel, supra, at p. 86.

22. Klein, "Are Your Tests Only Half Safe?" 48(1) Bar Examiner 137 (1979).

23. Id. at 138.

24. Cited in Klein, "On Testing IV: Essay Grading Fictions, Facts and Forecasts," 54(3) Bar Examiner 23, 24 (1985).

25. Regrettably, the National Conference of Bar Examiners does not regularly publish data regarding the reliability of the MBE. Nevertheless, it appears that the MBE has a reliability rating that would satisfy generally accepted standards. See, Id. at 138. See also, Klein, "On Testing V: How to Respond to Critics," 55(1) Bar Examiner 16 (1986) (discussion of reliability studies of MBE).

26. For discussions of these techniques, see Lenel, "The Essay Examination Part III: Grading the Essay Examination," 59(3) Bar Examiner 16 (1990). See also, Klein, "Essay Grading: Fiction, Facts and Forecasts," 54(3) Bar Examiner 23 (1985); Kurdys, "Grading Essay Answers: The Issue of Reliability in Essay Scoring," 59(4) Bar Examiner 22 (1990).

27. It is surprisingly simple for teachers to determine how much inconsistency exists in their own grading of essay-type questions. The process is simple. First, teachers must grade a whole set of essays. Then these teachers must select a random sample of those essays, perhaps 25% in a class of 50 - 60 people, and simple regrade those essays. When doing this regrading, of course, teachers must not allow themselves to know what score they gave a paper the first time it was graded. Then after the sample has been graded, a simple correlation analysis is done of the grades given on the two separate occasions. (Correlation analysis can be done instantly by any major spread sheet program.) If the correlation between the first set of scores and the second is relatively high, with 1.00 being perfect, then chances are the teachers is grading the essays in a fairly consistent manner. Conversely, if the correlation between the first set of scores and the second set of scores is low, with 0.00 between pure chance, then chances are that the teacher is grading the essays in a fairly inconsistent manner. Consistent grading, of course, increases test reliability whereas inconsistent grading decreases test reliability.

For a discussion of this process in connection with bar exam, see Lenel, "The Essay Examination Part III: Grading the Essay Examination," 59(3) Bar Examiner 16, 23-24 (1990).

28. Ebel, supra, at 80 et seq.; Gronlund, supra, at 87 et seq.; Mehrens, supra at 251-257; Popham, supra at 136 et seq. and 152 et seq.; Sax, supra, at 275 et seq.

29. Mehrens, supra at p. 260.

30. Note now an important point. Critics of the foregoing analysis might note that the impact of luck will be such that any one student's reported score is likely to be the same distance above or below that student's true score as any other student's reported score is likely to be above or below that other student's true score. Thus, these critics will conclude, reported scores can safely be used as "stand ins" for true scores. In one important sense this argument has considerable merit. If teachers use a test to evaluate the performance of an entire group of students, then the reported scores on that test can in fact stand in for the true scores. For the group as a whole, luck will in fact cause the positive differences that exist between some students' true and reported scores to cancel out the negative differences that exist between other students' true and reported scores.

Unfortunately, this stand-in notion does not work when teachers use test to generate grades for individual students. When grades are assigned to individual students, after all, the good luck (or bad luck) of other students will not be able to counteract the bad luck (or good luck) of the student involved.

31. 731 F2d 465, 495 (8th Cir. 1983).

32. 775 F2d 1403 (11th Cir. 1985).

33. For discussions of Susan M, see, Weinberger and Schepard, "Judicial Review of Academic Student Evaluations," 77 Ed. L. Rep. 1089 (1992). See also, Rains, "The Cautionary Ballad of Susan M," 40 J. Leg. Ed. 485 (1990). For an exhaustive collection and analysis of academic challenge cases, see, Schweitzer, "Academic Challenge Cases," 41 Amer. U. L. Rev. 267 (1992). See also, Milam and Marshall, "Impact of Regents of the University of Michigan v. Ewing on Academic Dismissals from Graduate and Professional Schools," 13 J. Coll. Univ. L. 335 (1987).

34. Board of Curators, University of Missouri v. Horowitz, 435 U.S. 78 (1978).

35. An interesting corollary to the Horowitz rule arose in Tarka v. Cunningham, 917 F2d 890 (1990). In Tarka, a student brought suit pursuant to the Family Educational Rights and Privacy Act (FERPA) seeking to have a grade change. The court denied the request after considering the purposes of FERPA. Interestingly, the court did not cite any of the Horowitz line of cases. Rather, it simply relied on FERPA itself, and some legislative history thereto.

36. Ewing, supra, 474 U.S. at 225.

37. See Branum v. Clark, 927 F2d 698 (2d Cir. 1991) for an allegation of just this sort of thing.

38. 76 Mich App. 631, 257 NW2d 195 (1977).

39. An elaborate recent discussion of High Stakes testing, a discussion dealing with virtually all of the pertinent topics, is Phillips, "Legal Issues in Performance Assessment, 79 West. Ed. L. Rep. 709 (1993). See also, for additional discussions of high stakes testing, Ballew, "Courts Psychologists, and the EEOC's Uniform Guidelines," 36 Emory L. J. 203 (1987); Connor and Vargyas, "The Legal Implications of Gender Bias in Standardized Testing," 7 Berk Women's L. J. 13 (19--); Parkinson, "The Use of Competency Testing in the Evaluation of Public School Teachers," 39 U. Kan. L. Rev. 845 (1991); Kelman, "Concepts of Discrimination in "General Ability" Job Testing, " 104 Harv. L. Rev. 1157 (1991); Moore and Braswell, "Quotas and the Codification of the Disparate Impact Theory," 55 Alb. L. Rev. 459 (1991); Pullin and Zirkel, "Testing the Handicapped," 44 Ed. L. Rep. 1 (1988).

40. The most important of the various Debra P decisions is Debra P v. Turlington, 564 FSupp. 177 (1983), aff'd 730 F2d 1405 (11th Cir. 1984).

41. See Parkinson, "The Use of Competency Testing in the Evaluation of Public School Teachers," 39 U. Kan. L. Rev. 845 (1991)

42. Shariff v. New York State Education Department, 700 FSupp 345 (S.D.N.Y. 1989).

43. See, for a discussion of numerous cases, Phillips, "Legal Issues in Performance Assessment, 79 West. Ed. L. Rep. 709 (1993).

44. See. e.g. Georgia State Conference of Branches of NAACP v. State of Georgia, 775 F2d 1403 (1985).

45. The most important high stakes testing case clearly is Debra P. v. Darlington, 474 FSupp. 244 (M.D. Fla. 1979). This case appeared again and again in the federal district court and in the appellate court.

46. The most important of the United States Supreme Court cases in this context is Griggs v. Duke Power Company, 401 U. S. 424 (1971)

47. Debra P. v. Turlington, supra.

48. 522 FSupp. 886 (E.D. Penn. 1981).

49. Note carefully here two things. First, because these standards were developed for use in connection with "standardized" tests and not for use in connection with traditional classroom tests, they establish relatively tough rules regarding reliability, measurement error and the like. These tough rules are necessary, of course, because standardized tests are used alone to make important decisions about students. Grades on classroom tests, of course, generally make up only part of the grade for a course. Second, a somewhat comparable set of rules and standards, although a much, much less technical set, has been developed by the Joint Committee on Testing Practices. This set of rules is called the "Code of Fair Testing Practices in Education." W. Wiersma and J. Jurs, Educational Measurement and Testing (2d ed. 1990) at 375 et seq. (copy of this Code). Again this code was developed for use not by classroom teachers but by high stakes testers.

50.79 West Ed. L. Rep. 709 at n. 222 et seq.

(Student Name or Exam Number)

CONTRACTS

[Fall, 1992]

ESSAY QUESTIONNAIRE AND SCORE SHEET

Paul Wangerin
John Marshall Law School
Chicago, Illinois

This questionnaire and score sheet can be used to evaluate essays submitted in connection with Mr. Wangerin's Contracts exams. It poses a series of questions that can be answered with either a "Y" (yes) or a "N" (no) or a "Y/N," which score indicates the scorer's ambivalence. (The questionnaire also provides a series of examples of items that would be scored with a yes.) Each part of the questionnaire also contains a scoring item for the teacher's use only. This item is used to rank individual students against all other students in a given class. A number grade of 5 ["Excellent"] indicates that the material evaluated is in the top 12-15% of all student material evaluated. A number grade of 1 ["Unacceptable"] represents the bottom 12-15% of the total material evaluated. A number grade of 2 ["Minimally Acceptable"] represents work in the bottom 30% and a grade of 4 ["Good"] indicates work in the top 30%. A number grade of 3 ["Average"] places the work squarely in the middle of all of the student material evaluated. This ranking process is completely subjective. It does not involve the mere totaling up of yes or no answers. Rather, it reflects the teacher's general "feeling" about the material being evaluated.

[OUTLINE]

(Notes: (i)Students submitting exam essays in Mr. Wangerin's classes must submit <u>detailed</u> outlines for those essays along with the essays themselves. This is the first thing that Mr. Wangerin reads and something that plays a major role in his evaluation of student work. It is unlikely in the Contracts class that anything short of a two page outline, made up of perhaps 25 or 30 individual item entries, will be sufficiently detailed. Rarely should such an outline, however, contain more than three <u>major</u> topics for discussion. (ii) Outlines should reflect the various points described below for each major part of the essay itself. (iii) It may be wise for the outline specifically to "name" certain things. For example, in the introduction portion of the outline, the outline might state "[anecdote]" or "[theme]" or "[road map]" after the pertinent entry. (iv) Students in the process of actually writing essays, rather than just in the process of planning them, frequently think of things that they forgot when preparing the essay's outline. Obviously, those things should be added to the essay. However, students who make these subsequent discoveries of important ideas almost certainly should go back to their outlines and scribble in the extra idea <u>before</u> actually adding the idea to the essay itself. Frequently, such further reference to the outline will demonstrate that the subsequently discovered idea should go into the essay somewhere other than the place in the essay where the writer is working at the time when the new idea is suddenly discovered. In Mr. Wangerin's classes, at least, absolutely no penalty is imposed for interlineation, additions, cross outs or other changes in already prepared material.)

INTRODUCTION

(Note: Many law school teachers believe that they can predict the grade for an entire exam question by the end of the first page of that answer. This fact should encourage students to plan their initial words and paragraphs with considerable care.)

a. Opening Anecdote

1. _____ Does the introduction use some sort of eye-catching or amusing example, or some kind of anecdote or graphic illustration as a vehicle for introducing the topic and getting the reader's attention?

> Example: "Several days ago, Mr. Wangerin, a teacher at JMLS, embarrassed the entire Contracts class when he made an off-color joke about the title of Chapter nine in the Farnsworth and Young Casebook. [Details provided.]"

b. Topic, Theme, and Road Map (New Paragraph?)

2. _____ Does the introduction explain what the essay will generally be about, i.e. what the essay's "Topic" is?

> Example: "This essay is about Mr. Wangerin, particularly about his activities as a law school teacher.

3. _____ Does the introduction announce the "Theme" for the essay, i.e. the unifying idea that will connect all of the various parts of the discussion, and does

that theme involve a statement of the writer's position on the topic issue?

> Example: "He [Wangerin] is, without doubt, the worst teacher that I have had in law school." (This is a direct quotation from the first student evaluation that Mr. Wangerin ever read. This particular evaluation went on to say that Mr. Wangerin was "the most obnoxious human being that I have ever met.")

(Note: Students in Mr. Wangerin's classes frequently choose something related to one of the "policy" issues discussed in class as a theme for their essay. ("The Uniform Commercial Code is a product of Communist conspiracy!") In fact, many students choose to address all of the policy issues discussed in class. This later practice is, of course, foolish. In a short essay, no time exists for detailed analysis of more than one theme idea. Some students take a completely different approach. These students look to the structure of the course itself, or to the casebook's organization, for essay themes. Finally, some students move well beyond class discussion for themes.

4. _____ Does the introduction provide a brief, one, two or three sentence "road map" to the balance of the essay?

> Example: "After initially talking briefly about legal education in general, I will demonstrate that Mr.Wangerin provides students with absolutely no assistance in learning. Then I will argue that Mr. Wangerin is downright rude in class. Finally I will show that he himself knows nothing worthwhile about the law of Contracts. (By 'worthwhile' I mean things that real lawyers need to know.)"

For Teacher's Use Only: Rank ______

[BACKGROUND]

(NOTE: Good writers frequently begin their essays with a discussion of background ideas or topics. Thereafter, the writers provide specific examples illustrating how the general topic takes on concrete reality. In short, the essay moves from the general to the specific. In a sense, essays structured this way involve "deductive" reasoning. Another excellent approach, however, is just the reverse. Here, an essay begins with discussions of specific instances and moves from them to discussion of general topics. This is, of course, "inductive" reasoning. Students in Mr. Wangerin's classes are encouraged to use the first of those methods in their essays for the Contracts class, if only because consistency in format among students makes evaluation easier.)

1. _____ Does the essay begin with a discussion of background information about the topic?

> Example: "It seems to me that law school teachers in general have a number of important responsibilities. [Here would follow a lengthy discussion of general details.]"
>
> Example: "Free market economic theory establishes ...."
>
> Example: "Historically, a tremendous gap has existed between judges who believe that disputes should be resolved principally in light of predetermined rules and judges who believe that disputes should be resolved on a 'case by case' approach."

2. _____ Does the discussion of this background information assume that the reader has little or no familiarity with the topic or idea being discussed? In other words, does the discussion of the background information consistently explain the topic in language that could be understood by someone unfamiliar with the topic itself, a particular course in school, or the course's instructor?

3. _____ Does the essay avoid using catch words or phrases that would only be known to participants in the particular class itself? Alternatively, if such catch words or phrases are used, are they sufficiently defined and explained? (Examples of catch words or phrases in the Contracts class include among many others: "Hard In/Hard Out," "Efficiency," "Formalist/Realist," "consideration," and "conditions.")

> Example: "One of the unusual things about legal education is the so-called "Socratic" method, a method of classroom instruction used, or misused, by many first year course teachers in law school. The Socratic method involves [description of Socratic method]."

4. _____ Is this part of the essay sufficiently general or is it cluttered up with too many specific details?

> Example: "For the time being, I am not going to talk specifically about Mr. Wangerin himself. Rather I am going to talk generally about legal education."

5. _____ Does the essay provide pertinent authority for the background ideas that it discusses? (Comments made by classmates, or by the teacher outside of his or her published writings, will <u>not</u> be considered authority. Actual authority, perhaps from text books or scholarly articles, should be provided if at all possible. The notes and text in the casebook itself also provide a tremendous amount of authority for these kind of general discussions.)

> Example: "Numerous books and articles that generally talk about legal education argue against the practice of inflicting psychological abuse on students in the name of the Socratic method. [Citation of Authorities] "

6. _____ Is the discussion of the background information sufficiently long and detailed? (Many students in the Contracts class fail to provide sufficient analysis in this section of their essays. Rather, they simply substitute for detailed analysis a few sentences or paragraphs simply parroting what has been said in class.)

For Teacher' Use Only: Rank _____

<u>SPECIFIC TOPICS</u>

1. _____ Is there a smooth transition out of the discussion of the background information and into the specific topics?

2. _____ Do the introductions to the various specific topics discussed in the body of the essay contain reference to the essay's overall theme?

> Example: "Probably the most important reason for saying that Mr. Wangerin is the worst teacher that I have ever encountered is that he is terribly rude."

3. ______ Are the specific topics in the essay themselves divided up into subtopics.

> Example: "Mr. Wangerin's rudeness in class generally took one of three forms. The first of those was that he frequently made fun of students who asked him to repeat a question."

4. ______ Does the essay provide specific examples as support for the points made in the topics and subtopics.

> Example: "For example, on one day that I particularly remember, he called on me when I was furiously taking notes. When I asked him to repeat the question that he had asked of another student, he proceeded to ask me what was happening outside the windows that I thought was so fascinating."

(Note: The three following questions are all closely related, a fact that should indicate the importance of the following issue. Indeed, these questions deal with what is probably the single most important thing evaluated when Mr. Wangerin evaluates essays.)

5. ______ Does the discussion of the specific topics take the reader back and forth across several seemingly unrelated facets of the course?

> Example: "Although Mr. Wangerin's many different kinds of rude comments initially seemed to me to be unrelated to each other, I now realize that all of them reflect the same underlying attitude, an attitude of...."

6. ______ Does the discussion of the topics avoid simply taking the reader on a chronological tour of the course?

> Bad Example: "On the first day of class, Mr. Wangerin interrupted a student and stated that.... On the second day of class, he.... On the third day...." (Amazingly, this particular bad example is representative of perhaps 70% of the essays submitted in response to a very common type of question on Mr. Wangerin's exams, namely, "Discuss [a topic].")

7. ______ If a discussion simply follows the structure of the course itself, or of the book, does the essay provide a compelling reason for doing that rather

than for moving back and forth throughout the course?

(Note: It has been Mr. Wangerin's experience over the last few years that an astonishingly large number of students write essays that would generate "no" answers in connection with all three of the foregoing questions. This occurs despite the fact that class discussion in the Contracts class constantly requires students to think and talk about issues and topics discussed weeks or even months earlier. It also occurs despite the fact that the single most frequent subject of class discussion in the Contracts class is the possible relationships between seemingly unrelated ideas.)

8. _____ Does the essay acknowledge and deal with potential weaknesses in its own position.

> Example: "To be sure, he seems to have some sort of psychic power for telling when people are not paying attention, and frequently these are the people that ask him to repeat questions. But...."

9. _____ Does the essay contain smooth transitions between its various parts?

> Example: "Not only does Mr. Wangerin's rudeness reflect a fundamental psychological problem that this teacher has. Wangerin also uses rudeness to cover up the fact that he doesn't know anything useful to real lawyers about the law of contracts."

10. _____ Does the discussion reflect the potential reader's unfamiliarity with the topic?

11. _____ Does the discussion avoid "catch words" or phrases?

12. _____ Do the successive parts of the discussion reflect the fact that the writer adequately budgeted his or her writing time?

> Bad Example: "(Although I am almost out of time I will have to quickly say one last thing.)"

For teacher's Use Only: Rank _____

CONCLUSION

(Note: Frequently the conclusion to an essay is the mirror image of the introduction.)

a. Road Map, Theme and Topic

1. _____ Does the essay conclude with a brief restatement of all of the things discussed in the essay? In short, does the essay restate the "road map?"

2. _____ Does the essay conclude with a final reference to the essay's overall theme and topic?

b. Closing "Clincher"

3. _____ Does the essay end with some sort of dramatic "clincher."

Example: "The real problem here, however, is much larger than just a problem with one particular bad teacher. Bad teachers like Wangerin will always exist. The real problem is this: What can students do when they are forced to take a class with a very bad teacher? As of now, the answer is simply stated.'Nothing.'"

For Teacher's Use Only: Rank ____

Overall Rankings: Outline _______ Introduction _______

Background ______ Specific Topics ______ Conclusion _____

OVERALL RANK: ______

GRADE: _______

Essay Question: Contracts

[Spring, 1992]

Paul T. Wangerin
John Marshall Law School
Chicago, Illinois

****

This 50 minute long, open book, open note question is worth 1/2 of the grade for this exam.

Smith, a highly successful dealer in rare basketball cards, and Jones, also a dealer in such cards, but a dealer who is just barely managing to stay out of bankruptcy, separately agreed to buy highly sought-after Scottie Pippen "Tongue" cards from Davis for $10,000 a piece. These cards, which are part of a very small run of misprints, display a picture in which Michael Jordan's tongue is transposed onto Scottie Pippen's face. (Davis, who also is a dealer in these cards, obtained them by bribing an employee of the printing plant.) The contracts entered into by these parties included the following language, language which Jones strenuously but unsuccessfully attempted to have excluded:

> Buyer and Seller hereby agree that if the card at issue is ultimately determined by recognized experts to be a counterfeit, or if the Seller in any other way breaches this contract, the Seller will hold the Buyer harmless for any and all losses incurred by the Buyer in connection with the agreement regarding this Card. However, Buyer and Seller also hereby agree that under no circumstances shall such losses be considered to be greater than any difference that might exist between the contract price and the market price for these cards on the date this agreement is executed.

Shortly after the parties entered into these agreements, the price of the cards skyrocketed to $25,000. Davis then told Smith and Jones that she would not perform. Almost immediately after learning of the breach, however, Smith bought another of these cards for $19,000 from yet another dealer. Jones, however, despite strenuous efforts, could not find any other cards of this type. This turn of events proved to be the final straw for Jones' creditors who then forced him into bankruptcy.

Davis has now come to you seeking advice regarding lawsuits that she anticipates will be brought against her. Please discuss the various issues that might arise in these lawsuits.

Essauy Scoring Grid: "Basketball Cards"

Paul Wangerin
John Marshall Law School
Chicago, Illinois

[Spring, 1992]

****

The scoring grid for the "basketball cards" question does two things. First, it "groups" or "classifies" the various issues in the problem as follows: (1) illegalilty issues; (2) liquidated damages / penalties issues, and (3) general remedies issues. Second, the grid notes that three major points should be addressed in connection with each of those principle issues. Note carefully in this context, therefore, two things. First, the fact that this particular grouping is used in this should not be read as suggesting that scorers should deduct points if students actually answering the question used different groups. Rather, if this occurred, and it is quite likely that it did, scorers simply must look throughout the answer for pertinent discussions. Second, since the same number of sub-issues or sub-points has been listed in connection with each of the principle issues, the scores specified for each of the three principle issues are calculated in a roughly comparable way.

FIRST ISSUE (Illegality)

Good answers will address the following three points in connection with the illegality issue:

a. Illegality is one of several different kinds of defenses that allow parties to set aside contracts that otherwise satisfy the requirements for enforceability.

b. The buyers' reliance in this situation might negate the general rule regarding illegality.

c. The buyers' possible ignorance regarding the illegality of the cards, i.e. their status as possible BFP's, might completely negate the impact of the illegality issue.

The first issue on this exam, illegality, should be scored as follows on the separate Score Record:

0 Points:	Essay does not address this issue.
1 Point:	Essay contains a very, very brief reference to this issue, but nothing more than a sentence or two.

2 Points: Discussion of this issue is longer than one sentence or so but contains no discussion of any of the points listed above.

3 Points: Discussion of this issue addresses one, but only one, of the points listed above.

4 Points: Discussion of this issue addresses two, but only two, of the points listed above.

5 Points: Discussion of this issue addresses all of the points listed above.

SECOND ISSUE (Liquidated Damages / Penalty)

Good answers will address the following three points in connection with the liquidated damages / penalty issue.

a. The quoted language probably is a penalty rather than a liquidated damages clause. Thus, this clause probably does not limit either Smith's or Jones' damages.

b. Davis will insist, relying on Lake River, that the clause at issue, even if arguably a penalty, was in fact a part of the negotiated price. Had it not been for this clause, he will argue, he would have insisted on a higher sale price.

c. The different financial status of Smith and Jones is important. Because of his financial troubles, Jones might argue that the clause was unconscionable. (He had to sign it.) Smith, however, cannot make this argument.

The second issue on this exam, liquidated damages / penalities, should be scored as follows on the separate Score Record:

0 Points: Essay does not address this issue.

1 Point: Essay contains a very, very brief reference to this issue, but nothing more than a sentence or two.

2 Points: Discussion of this longer than one sentence or so but contains no discussion of any of the points listed above.

3 Points: Discussion of this issue addresses one, but only one, of the points listed above.

4 Points: Discussion of illegality issue addresses two, but only two, of the points listed above.

5 Points: Discussion of illegality issue addresses all of the points listed above.

__**THIRD ISSUE**__ (General Remedies for Smith and Jones)

Good answers will address the following three points in connection with the general remedies issue.

a. The bankruptcy of Jones may have been or may not have been forseeable to Davis. Thus, Jones' damages may be, or may not be, limited by the forseeability rule.

b. Except as noted belwo, the method for calculating the remedies for Smith and Jones is completely <u>different</u>. In Jones' case, the remedy would be calculated using one of the standard formulas. In Smith's case, however, the remedy probably would be calculated by looking at the difference between the contract price ($10,000) and the "substitute transaction" price ($19,000).

c. Smith should argue that he is a truly unusual entity, namely, a "lost volume <u>buyer</u>." He would do this by insisting that his second purchase was <u>not</u> a substitute transaction at all. Rather, he would argue that he would have bought the second card even if Davis had not breached. If this argument succeeds, Smith's and Jones' remedies will be same.

The third issue on this exam, the general remedies issue, should be scored as follows on the separate Score Record:

0 Points: Essay does not address this issue.

1 Point: Essay contains a very, very brief reference to this issue, but nothing more than a sentence or two.

2 Points: Discussion of this longer than one sentence or so but contains <u>no</u> discussion of <u>any</u> of the points listed above.

3 Points: Discussion of this issue addresses <u>one, but only one</u>, of the points listed above.

4 Points: Discussion of illegality issue addresses <u>two, but only two</u>, of the points listed above.

5 Points: Discussion of illegality issue addresses <u>all</u> of the points listed above.

(This material has been reproduced
with permission of the publisher.)

LOU M. CAREY
UNIVERSITY OF SOUTH FLORIDA

Measuring and Evaluating School Learning

ALLYN AND BACON, INC.
BOSTON LONDON SYDNEY TORONTO

Copyright © 1988 by Allyn and Bacon, Inc.
A Division of Simon & Schuster
7 Wells Avenue
Newton, Massachusetts 02159

Library of Congress Cataloging-in-Publication Data

Carey, Lou.
Measuring and evaluating school learning.

Includes bibliographies and index.
1. Educational tests and measurements—United States.
I. Title.
LB3051.C36 1988 371.2'6'0973 87-17569
ISBN 0-205-11109-2

Series editor: Mylan Jaixen
Production administrator: Annette Joseph
Production coordinator: Helyn Pultz
Editorial-production services: Wordsworth Associates/Grace Sheldrick
Cover administrator: Linda K. Dickinson
Cover designer: Susan Slovinsky

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1 92 91 90 89 88 87

H55

BRIEF CONTENTS

CHAPTER 8

Constructing and Using Essay and Product Development Tests

OBJECTIVES

1. Define essay and product development tests.
2. Describe the types of skills typically measured using essay and product development tests.
3. Discuss the positive and negative features of essay and product development tests.
4. Write essay questions and product development instructions.
5. Select and include relevant information in directions.
6. Define global scoring and describe its uses.
7. Define analytical scoring and describe its uses.
8. Develop a checklist for analytical scoring.
9. Develop a rating scale for analytical scoring.
10. Use a rating scale to score given products.
11. Describe typical errors associated with constructing and scoring essay and product development tests.
12. Summarize and analyze group performance on essay and product development tests.
13. Describe benefits of students using checklists and rating scales to evaluate their own work.

Curriculum guides often contain complex instructional goals that require students to demonstrate their ability to create a unique response. In constructing unique responses, students need to determine how they will approach a given problem, plan and organize their responses, and present their ideas. Objective tests do not require students to produce original pieces of work that demonstrate these capabilities.

Often these complex goals can be measured with essay questions, which typically require students to discuss, analyze, compare for similarities and differences, synthesize, or evaluate. For example, students may be asked to discuss a product, a procedure, an event, a political action, a natural phenomenon, historical characters, a scientific experiment, a piece of literature, or a philosophical viewpoint. They may be expected to analyze and describe the components of a concrete or defined concept, rule, or principle, or to analyze and describe the steps in a procedure or the components in a system. Essay questions also can be used to measure their ability to compare or contrast concepts, theories, systems, procedures, events, people, and numerous other subjects. They can measure students' abilities to synthesize pieces of literature, articles, events, or other phenomena. In addition, students may be asked to evaluate the quality of a product or event, the execution of some procedure or motor skill, or an instance of behavior.

Some instructional goals cannot be adequately measured by either objective or essay questions. To demonstrate their skill, students need to develop some type of product either with pencil and paper or by some other method. Such products include letters, themes, poems, abstracts, term papers, flow charts, computer programs, original songs, photographs, videotapes, maps, bookcases, blueprints, and drawings. Each student's product will be unique and will vary in complexity, originality, and accuracy. Even when students are following detailed instructions, their responses will vary.

Both essay and product development tests require students to synthesize many enabling skills in their responses. For example, students writing a summary of a book must form sentences and paragraphs, write legibly, select ideas and information, and organize their presentation. One advantage of these tests is the teacher's ability to separate and comment on particular elements of students' responses, such as their approach, organization, logic, and accuracy.

Unfortunately, this type of testing has several major drawbacks. First, a teacher must spend considerable time selecting the task, writing the instructions, scoring the products, analyzing students' work, and evaluating both individual and group performance. Second, more class time is usually required for essay and product exams than for objective-style exams, which often limits the breadth of the material tested. Third, because responses vary, only experts can make the fine distinctions that determine whether an answer is acceptable. Responses cannot be machine scored or given to an aide or student assistant for scoring. Finally, scoring is less reliable on these exams than on objective exams for several reasons. Teachers' standards may shift during scoring, and fatigue can cause lapses in concentration. Scoring bias is another serious problem. Some teachers tend to be lenient whereas others consistently give all the students average or below average marks. A teacher's perceptions of individual students also can bias the scores they assign. In addition, students with excellent verbal and organizational skills can bluff an answer and often receive better scores than less verbally skilled students whose answers contain superior content. Because of these limitations, you should use essay and other product development tests only when the prescribed skills cannot be measured with objective tests.

WRITING QUESTIONS AND INSTRUCTIONS

Many of the suggestions for writing objective test items also apply to essay questions and instructions for product development. Questions and instructions should match the behavior, content, and conditions specified in the behavioral objective; the vocabulary, complexity, and context should be appropriate for target students; and questions and instructions should be clearly written using correct grammar and punctuation.

An essay question or set of instructions should describe the type of response that you expect. For example, if students are to compare two things, the instructions should begin with the word *compare*. If they are to critique something, the instructions should begin with the word *critique*. Students should know the meaning of these words and the ways they can demonstrate these skills. If all or some students do not understand what these terms are, their responses may reflect their misunderstanding of what they are told to do.

Providing Guidance

The amount of guidance in the question or instructions depends on the skill being measured and the sophistication of the students. Related to the skill, instructions may require students to discuss or compare things using components they select themselves, or the instructions may specify which components they are to use. Students may have to select and use evaluation criteria, or they may be given the criteria to apply. Related to students' characteristics, older students and high achievers tend to work well with minimum guidance. However, younger students and average or below-average achievers need more structure in the directions.

Students' responses will depend on the amount of guidance in the instructions. Consider the amount of guidance provided in the three essay questions included in Table 8.1. The questions in the left-hand column have the content removed to illustrate the item format. Any content could be substituted for the letters X and Y and the numbers 1, 2, and 3. The questions in the right-hand column use the same format, but content is included in each.

Notice that the three sets of instructions vary considerably in the amount of guidance provided. The first question requires students to decide which aspects of the pretest and posttest they will compare and how they will com-

TABLE 8.1 Degrees of Guidance in Essay Questions

1. Compare X and Y.	1. Compare *pretests* and *posttests*.
2. Compare X and Y for the following: 1. 2. 3.	2. Compare *pretests* and *posttests* for the following: 1. the test's relationship to the goal framework 2. the time the test is administered 3. uses for test scores
3. Compare the similarities of and differences between X and Y for the following: 1. 2. 3.	3. Compare the similarities of and the differences between *pretests* and *posttests* for the following: 1. the tests' relationship to the goal framework 2. the time the test is administered 3. uses for test scores

pare them. This is the most complex question because students must determine all the elements in their response. The lack of guidance will produce greater variety in students' answers.

The second question is more restrictive. Students are told which aspects of pretests and posttests they are to compare. The third set directs students' responses even more by specifying that students compare similarities and differences. Each question is of value. The amount of guidance to include in an essay question depends on whether you want students to recall the facets to be compared or whether you want to provide them. In determining the appropriate amount of guidance, you should always consider what skills you want to measure and the skill level of students being tested.

The amount of guidance to include is also a consideration when writing instructions that require development of a product. For example, the instructions for producing a paragraph can vary considerably in the amount of guidance provided. Consider the following three sets of instructions:

1. Write a paragraph about a fire drill.
2. Write a paragraph about a fire drill. Your paragraph should include:
   a. A topic sentence
   b. At least four supporting sentences
   c. A concluding sentence
3. Write a paragraph that describes fire drill procedures. Your paragraph should include:
   a. A topic sentence
   b. At least four supporting sentences
   c. A concluding sentence

Each set of directions will undoubtedly produce very different paragraphs. The second set is more specific than the first and the third more specific than the second. The third set limits the topic to fire drill procedures and specifies the number and types of sentences to be included. To determine which set of instructions would be best, you would first need to decide how much guidance should be provided for a particular group of students and what skills you want to measure. Specific instructions like those in the third set would not be appropriate if you wanted to measure whether students remembered to include topic and concluding sentences in their paragraphs. However, they would be appropriate for measuring students' skill in ordering events in a prescribed procedure (such as a fire drill). Regardless of the test's purpose, the directions must include enough information to provide clear and unambiguous guidance for the task.

Providing Organizational Information

Besides specifying the nature of the task to be performed, instructions can include additional information to help students determine the relative importance of different questions or sections of the test. Information that can help students organize their work includes (1) the length and scope of the response sought, (2) the number of points each question is worth, and (3) the time available for completing the test or the recommended time students should spend on each question. This type of information is incorporated in the paragraph test in the following example:

1. Write a paragraph that describes fire drill procedures. Your paragraph should include six to ten sentences. Be sure to include:
   a. A topic sentence (5 points)
   b. At least four supporting sentences (8 points)
   c. A concluding sentence (4 points)
   You will have 20 minutes to write and revise your paragraph.

Essay tests can include several questions that require relatively brief answers, or they may contain only one or two questions that require lengthy responses. Whatever the number of questions, the entire class should be directed to answer all the questions included on the test. One relatively common practice in developing essay tests, which is *not* recommended by test specialists, is to provide several questions on an essay test and allow students to select a subset of questions to answer. This practice would be similar to letting students answer twenty of thirty objective items. Both essay and objective test items represent only a few of those a teacher could construct, and the same sample should be used to evaluate all students. When students' responses are not comparable, your ability to evaluate your instruction and students' performance is compromised.

DEVELOPING SCORING PROCEDURES

Teachers typically use two types of scoring procedures to evaluate the quality of students' responses to essay questions and products. With *global* scoring (also called *holistic* scoring) the teacher uses general impressions to judge the quality of an answer or product. A teacher using *analytical* scoring divides a response into its relevant components and evaluates each part separately.

Global Scoring of Essay and Product Tests

Global scoring is appropriate whenever a test is *not* used to provide corrective feedback to students or to evaluate instruction. Its purpose is to sort students' responses into categories that indicate quality. For example, global scoring would be sufficient for a writing test used to place students in an English class. When a teacher has many papers to score, and it is not important to communicate the nature of the errors, global scoring is adequate.

The procedure for global scoring consists of the following seven steps.

1. *Establish the scoring categories you will use.* For example, you may have responses that fall into such categories as pass or fail; good, adequate, and poor; or excellent, good, adequate, poor, and unacceptable. The number of categories selected depends on the purpose for the evaluation and on your ability to place similar responses consistently into the same category. If you create too many categories or do not carefully define each category, you will find scoring difficult.
2. *Characterize a response that fits each category.* If, for example, you were to use the three categories, excellent, adequate, and poor, you should describe the particular characteristics that a response should have to be classified into one of the categories. What characteristics should be present for a response to be considered excellent? What would be absent or present in an adequate response? What would be absent in a poor response? Listing these characteristics helps you classify responses more consistently.
3. *Read each response rapidly and form an overall impression.* During the reading, look for the characteristics you used to describe each rating category.
4. *Sort the responses into the designated categories.*
5. *Reread the papers that have been placed within a category.* After all papers have been classified, you should consider only those within a set for their comparability.
6. *Move any clearly superior or inferior responses to other categories.*
7. *Assign the same numerical score to all responses within a category.* For example, papers in the excellent category can be assigned a score of five, those in the adequate category can be assigned a score of three, and so forth.

BEST COPY AVAILABLE

Although global scoring is relatively fast and reliable, two limitations make it less appropriate for classroom tests. First, it does not provide students with adequate feedback about their work. Most students will want to know why they received the assigned score and will not be satisfied with a global rating. Second, global evaluations do not permit the teacher to analyze the responses and identify specific instructional problems. Thus, students do not have the information they need to correct their mistakes, and teachers cannot classify errors and relate them to problems in their lessons.

Analytical Scoring of Product Tests

Analytical scoring is more time consuming than global scoring, but it is a superior method for instructional purposes. This procedure helps a teacher focus on relevant aspects of students' responses and provides a systematic way to assign partial credit. Just as important, it allows students to see where they lost points. Using this method, teachers can summarize the group's performance on main components, analyze errors, and use the error analysis to evaluate and revise instruction.

At the same time, analytical scoring also has some disadvantages. First, constructing a scoring instrument and marking responses take considerable time. If the teacher has not identified and sequenced all the components desired in an answer in the order they are likely to appear in students' responses, searching for these components in students' work will take even longer. Second, developing a flexible scoring procedure that accommodates unanticipated responses can also be difficult. Teachers using analytical scoring may struggle with the problem of scoring correct, unusual answers that do not fit the structure and that cannot be compared to other students' responses.

Analytical scoring can be simplified by using a form that indicates the desired elements of a response and the number of points to be awarded for each. To construct such a form, first identify and order the desired components and subcomponents of each response. The major components you select should always be based on an instructional goal and its framework of subordinate skills. In addition to these skills, you may want to use other elements, such as the students' approach, organization, and originality, to judge their work. Some teachers believe that grammar, neatness, punctuation, and spelling should always count whereas others choose not to emphasize these elements in every task. If you choose to include elements not contained in the goal framework and not included in instruction for the unit, you should inform students of this intention in the test directions.

The process of selecting and ordering components can be demonstrated with an example. Figure 8.1 shows the framework of enabling skills for the goal, "Write a paragraph on a given topic that contains a topic sentence, supporting sentences, and a concluding sentence." Although most of the twenty-eight enabling skills can be measured with objective test items that require students either to write a short answer or select a response, the instructional goal requires that students write a paragraph. The goal framework indicates that students' paragraphs should contain (1) indentation, (2) a topic sentence, (3) supporting sentences, and (4) a concluding sentence. These four elements make up the main components included on the evaluation form. Both sequence and transition are considered within the supporting sentences component.

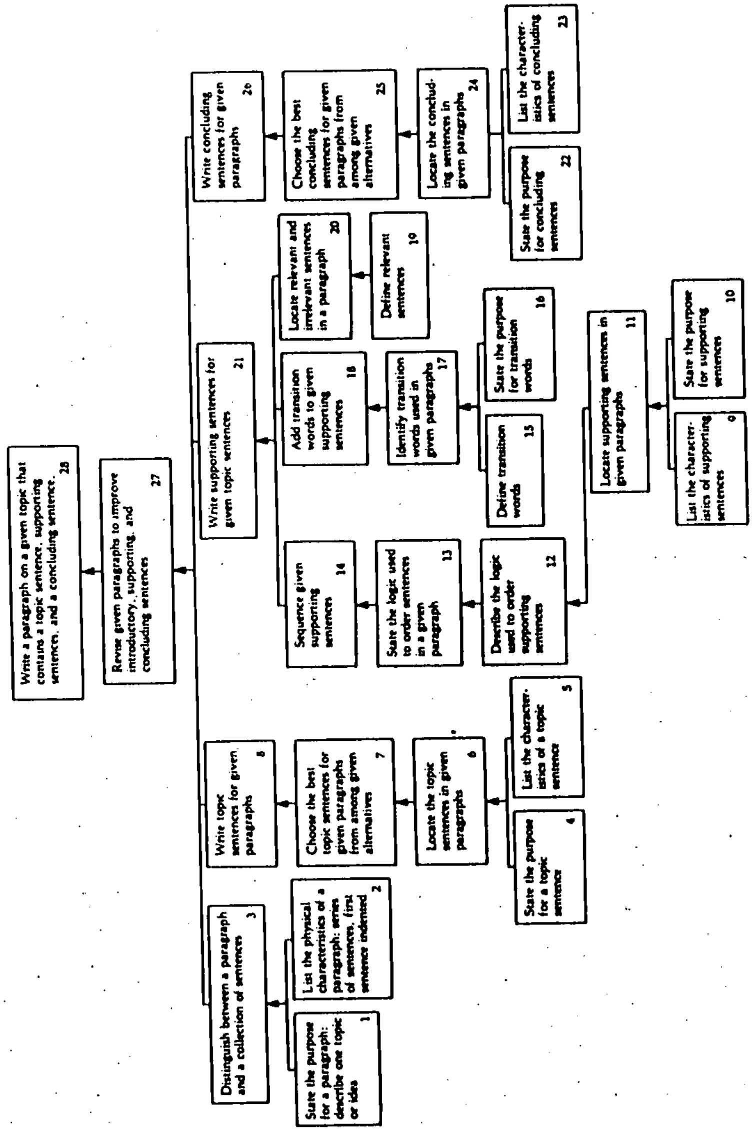
Next, you need to identify relevant subcomponents for each main component. Indentation is a simple skill that has no subcomponents. Writing a topic sentence is a more complex skill that includes at least two subcomponents. First, the topic sentence should be a general statement that introduces a topic, and second, it should be appropriately placed in the paragraph. Writing supporting sentences is even more complex so it has more subcomponents. The supporting sentences should all relate to the topic, should be logically sequenced, and should contain transition words. These three criteria are subcomponents for supporting sentences. Finally, you need to identify relevant subcomponents for a concluding sentence. Subcomponents could include whether the sentence adequately summarizes the topic and whether it is logically placed in the paragraph. Main components and their subcomponents should be sequenced in the order they are most likely to appear in the paragraphs.

Once you have selected and ordered the evaluation criteria from the goal framework, you must decide whether skills not included in the framework will also be judged. If you want to score prerequisite skills, such as complete sentences, correct verb tenses, spelling, and punctuation, you must add these to your list of criteria. Your completed outline of criteria taken from the goal framework would appear as follows:

I. Indentation
II. Topic Sentence
  A. Introduces topic
  B. Logically placed
III. Supporting Sentences
  A. Directly related to topic
  B. Logically sequenced
  C. Includes transition words
IV. Concluding Sentence
  A. Summarizes topic
  B. Logically placed

Criteria from the goal framework

FIGURE 8.1 Instructional Goal Framework for Writing a Paragraph


Write a paragraph on a given topic that contains a topic sentence, supporting sentences, and a concluding sentence. 28
Revise given paragraphs to improve introductory, supporting, and concluding sentences 27
Distinguish between a paragraph and a collection of sentences 3
State the purpose for a paragraph: describe one topic or idea 1
List the physical characteristics of a paragraph: series of sentences, first sentence indented 2
Write topic sentences for given paragraphs 8
Choose the best topic sentences for given paragraphs from among given alternatives 7
Locate the topic sentences in given paragraphs 6
State the purpose for a topic sentence 4
List the characteristics of a topic sentence 5
Write supporting sentences for given topic sentences 21
Sequence given supporting sentences 14
State the logic used to order sentences in a given paragraph 13
Describe the logic used to order supporting sentences 12
Locate supporting sentences in given paragraphs 11
List the characteristics of supporting sentences 9
State the purpose for supporting sentences 10
Add transition words to given supporting sentences 18
Identify transition words used in given paragraphs 17
Define transition words 15
State the purpose for transition words 16
Locate relevant and irrelevant sentences in a paragraph 20
Define relevant sentences 19
Write concluding sentences for given paragraphs 26
Choose the best concluding sentences for given paragraphs from among given alternatives 25
Locate the concluding sentences in given paragraphs 24
State the purpose for concluding sentences 22
List the characteristics of concluding sentences 23

V. Grammar and Spelling
   A. Complete sentences
   B. Correct spelling
   C. Correct punctuation

Criteria based on prerequisite skills for the goal framework

This outline of criteria is the basis for either a *checklist* or a *rating scale.*

Checklists A checklist permits you to judge only the presence or absence of each component. To score a written product that students developed, simply check each component that appears in a response and add up the checkmarks to obtain a total score. In the paragraph examples, the components in the outline become evaluation criteria on the checklist. A student receiving eight check marks receives a perfect score. Table 8.2 illustrates how a paragraph checklist might appear. Only criteria taken from the goal framework are included in the example.

TABLE 8.2 Checklist for Evaluating Students' Paragraphs

Name ________________ Date ________ Score ________
Total (8)

Criteria	*Yes*
I. Indentation (1 point)	
A. Paragraph is indented	________
II. Topic Sentence (2 points)	
A. Topic sentence is included	________
B. Topic sentence is appropriately placed	________
III. Supporting Sentences (3 points)	
A. They directly relate to the topic	________
B. They are logically sequenced	________
C. They contain transition words	________
IV. Concluding Sentence (2 points)	
A. Concluding sentence is included	________
B. Sentence is appropriately placed	________

Rating Scales The rating scale, an extension of the checklist, lets you judge not only the presence of a component but also its quality. A rating scale contains all the criteria on the checklist as well as a sequence of numbers that represent degrees of quality. Table 8.3 shows a rating scale for evaluating paragraphs. Each rating category on the scale includes both a number and a description of quality. A zero indicates the absence of a component or a subcomponent. Zeros should be used for rating only when the rated component can be absent. For example, they should not be used to rate such elements as spelling or organization because some degree of this type of element will be present. Such components as a topic sentence may be missing altogether, so using a zero to represent this absence would be appropriate.

TABLE 8.3 Rating Scale for Evaluating Paragraphs

Name ________________ Date ________ Score ________
Total (17)

I. Indentation (1 point)	Not Indented 0	Clearly Indented 1	
II. Topic Sentence (4 points)			
A. Content	Not Present 0	Vague 1	Clearly Introduces Topic 2
B. Location	Not Present 0	Misplaced 1	Logically Placed 2
III. Supporting Sentences (8 points)			
A. Relevance	Some Irrelevant Information 1	All Relevant Information 2	Thoroughly Develops Topic 3
B. Sequence	Illogical Order 1	Some Order 2	Logical Order 3
C. Transition words	No Transition 0	Some Transition 1	Smooth Transition 2
IV. Concluding Sentence (4 points)			
A. Content	Not Present 0	Not Comprehensive 1	Clearly Summarizes Topic 2
B. Location	Not Present 0	Misplaced 1	Logically Placed 2

The number of rating categories included for each element will depend on your ability to distinguish clearly between adjacent categories. A scale that includes too many categories will affect your rating consistency and thus the reliability of the scores. The words you use to describe each number will also aid the consistency of your rating. Descriptors, such as *inadequate, adequate,* and *good,* may be sufficient, but other terms related to the characteristics of the component measured may help you be more precise. Vague descriptors often reflect vague understanding of each number, and both affect the consistency of your ratings.

Look at the descriptors accompanying each number in the rating scale in Table 8.3. Rather than including vague terms, such as *inadequate, adequate,* and *good,* the descriptors relate to characteristics of the component to be rated. For example, the descriptors for the relevance of supporting sentences (III.A) indicate that students who include irrelevant information in their supporting sentences will be assigned a score of one. Those who include only relevant sentences will receive a score of two. Students whose supporting sentences are relevant and also develop the topic thoroughly will receive a maximum rating of three. Such descriptors for each score help you judge the quality of paragraphs consistently and clearly communicate each rating to students.

Checklists and rating scales may also be used to evaluate other types of student products, such as a bookcase or a picture frame constructed in an industrial arts class, a dress or meal produced in a home economics class, or a mobile constructed in either a math or art class. For example, students being tested on the goal "Construct a mobile" would need to construct mobiles to demonstrate their skill even though the principles involved could be measured at the knowledge and comprehension levels using an objective test. The instructional goal framework should be used as the basis for both objective tests and the product evaluation form. A goal framework that a math teacher might develop is illustrated in Figure 8.2. Table 8.4 is a corresponding checklist for evaluating students' mobiles. Notice that the criteria included on the checklist are observable and reflect the teacher's judgment of major components that should be included. An instructional goal framework and checklist created by an art teacher would undoubtedly be different than these examples because art teachers would need to include additional enabling skills related to the artistic quality of the mobile.

TABLE 8.4 Checklist for Evaluating Mobiles

Name ________________ Date ________ Score ________

Total (15)

I. Mobile with 1:1 weight ratio between two objects
________ 1. Objects correctly selected by weight
________ 2. Objects correctly positioned on crosswire
________ 3. Fulcrum found on crosswire

II. Mobile with 1:2 weight ratio between two objects
________ 1. Objects correctly selected by weight
________ 2. Objects correctly positioned on crosswire
________ 3. Fulcrum found on crosswire

III. Mobile with 1:3 weight ratio between two objects
________ 1. Objects correctly selected by weight
________ 2. Objects correctly positioned by crosswire
________ 3. Fulcrum found on crosswire

IV. Mobile having three objects of unequal weight and two crosswires
________ 1. Objects correctly selected by weight
________ 2. Wires correctly selected by length
________ 3. Objects attached to correct crosswires
________ 4. Objects positioned correctly on crosswires
________ 5. Fulcrum found on lower crosswire
________ 6. Fulcrum found on upper crosswire

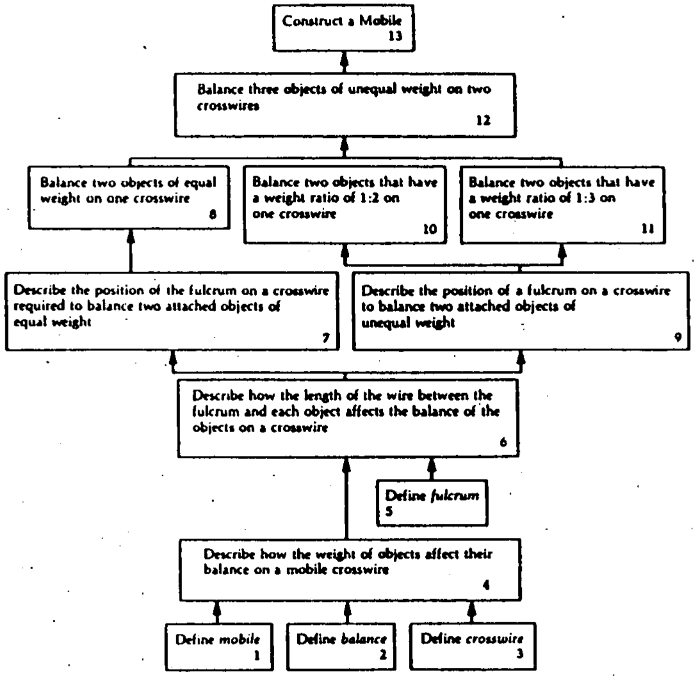

Construct a Mobile 13
Balance three objects of unequal weight on two crosswires 12
Balance two objects of equal weight on one crosswire 8
Balance two objects that have a weight ratio of 1:2 on one crosswire 10
Balance two objects that have a weight ratio of 1:3 on one crosswire 11
Describe the position of the fulcrum on a crosswire required to balance two attached objects of equal weight 7
Describe the position of a fulcrum on a crosswire to balance two attached objects of unequal weight 9
Describe how the length of the wire between the fulcrum and each object affects the balance of the objects on a crosswire 6
Define *fulcrum* 5
Describe how the weight of objects affect their balance on a mobile crosswire 4
Define *mobile* 1
Define *balance* 2
Define *crosswire* 3

FIGURE 8.2 Instructional Goal Framework for Constructing a Mobile

Plans for Weighting Components

After you have developed a checklist or rating scale for some type of product, you may need to weight the scores assigned for each criterion. You may want to add points to more important criteria or give more credit for complex or time-consuming skills. You might also want to weight skills that were emphasized during instruction more than skills such as spelling or neatness.

BEST COPY AVAILABLE

To weight different components, first look at the points each component receives on your checklist or rating scale. For example, Figure 8.3 shows the points already assigned in the checklist in Table 8.2 and the rating scale in Table 8.3.

FIGURE 8.3 Component Points

	Possible Points	
Component	*Checklist*	*Rating Scale*
I. Indentation	1	1
II. Topic Sentence	2	4
III. Supporting Sentences	3	8
IV. Concluding Sentence	2	4
	8	17

In both scoring plans, indentation receives the least amount of credit; topic and concluding sentences receive more credit than indentation; and supporting sentences receive the most credit. This balance may not reflect the values you would like each component to have. If you want to adjust the weighting, you need to determine the relative value of each component. For example, you might want the topic sentence, the supporting sentences, and the concluding sentences to be the same value and indentation to count much less than these three. In this case, you could multiply each of the sentence components by a factor that would give them equal weight in the total score.

Table 8.5 shows two different weighting plans. The first illustrates the weighting just described. Components II, III, and IV on the checklist are multiplied by a factor that makes each equal to six points; component I, indentation, is multiplied by 1, which keeps its total the same. The original 8 points have now become 19, with indentation assuming a much smaller portion of the total. On the rating scale, components II and IV are multiplied by a factor of 2, which makes each sentence component worth 8 points. The total score is now 25, with indentation counting only a very small percentage of the score.

The lower section of Table 8.5 shows a plan for the checklist and rating scale that weights the topic sentence and concluding sentence twice as much as the supporting sentences. On the checklist, components II and IV are weighted by a factor of 6, which gives each component a value of 12 points. Components I and III are weighted by a factor of 2, which gives them one-sixth and one-half the value of the other two components. On the rating scale, weightings of 4, 2, and 1 are used to weight indentation one-eighth and supporting sentences one-half the value of the topic and concluding sentences.

Table 8.6 shows a completed rating scale that weights topic, supporting, and concluding sentences equally. The scores for one student are inserted to illustrate how the form would be completed. To score the student's paragraph, you would circle the score assigned for each subcomponent and then sum these scores to obtain the original component score. This component score is placed in the Original Score column and then multiplied by the assigned weight. Finally, you would add the numbers in the Total column to determine the student's weighted score. As the scale indicates, this student received 21 of 26 possible points.

TABLE 8.5 Two Plans for Weighting Scores

Component	*Checklist*					*Rating Scale*				
	Original		Weight		Total	Original		Weight		Total
I. Indentation	1	×	1	=	1	1	×	1	=	1
II. Topic Sentence	2	×	3	=	6	4	×	2	=	8
III. Supporting Sentences	3	×	2	=	6	8	×	1	=	8
IV. Concluding Sentence	2	×	3	=	6	4	×	2	=	8
Total Points	8				19	17				25

Results: On the checklist, components II, III, and IV are of equal value and each counts six times as much as component I. On the rating scale, components II, III, and IV are equal and each counts eight times as much as component I.

Component	*Checklist*					*Rating Scale*				
	Original		Weight		Total	Original		Weight		Total
I. Indentation	1	×	2	=	2	1	×	2	=	2
II. Topic Sentence	2	×	6	=	12	4	×	4	=	16
III. Supporting Sentences	3	×	2	=	6	8	×	1	=	8
IV. Concluding Sentence	2	×	6	=	12	4	×	4	=	16
Total Points	8				32	17				42

Results: In both cases, components II and IV have equal value and count twice as much as component III and significantly more than component I.

Analytical Scoring of Essay Tests

The preceding examples illustrate the development of checklists and rating scales for scoring students' products. An extension of these same procedures is used to develop analytical scoring forms for essay tests. Scoring forms for essay tests differ somewhat because these tests usually contain several questions that represent broader content areas. The basic procedure for developing the scoring form for each question is the same as that previously described, that is (1) identify and sequence the main components and subcomponents for each question, (2) determine whether a checklist or rating scale would be most useful, and (3) develop the form weighting each component according

TABLE 8.6 Weighted Rating Scale for Evaluating Paragraphs

Name Katy Augustine Date 10/14 Score 21
Total (26)

Components	Quality			Original Score × Weight	Total
I. Indentation	Not Indented 0	Clearly Indented [✓ 1]		1 × 2 =	2
II. Topic Sentence A. Quality	Not Present 0	Vague 1	Clearly Introduces Topic [✓ 2]	4 × 2 =	8
B. Location	Not Present 0	Misplaced 1	Logically Placed [✓ 2]		
III. Supporting Sentences A. Content	Some Irrelevant Information 1	All Relevant Information [✓ 2]	Thoroughly Develops Topic 3	5 × 1 =	5
B. Sequence	Illogical Order 1	Some Order [✓ 2]	Logical Order 3		
C. Transition	No Transition 0	Some Transition [✓ 1]	Smooth Transition 2		
IV. Concluding Sentence A. Content	Not Present 0	Not Comprehensive 1	Clearly Summarizes Topic [✓ 2]	3 × 2 =	6
B. Location	Not Present 0	Misplaced [✓ 1]	Logically Placed 2		

Total 21

to a plan. In addition, you may need to assign relative weights to each question to ensure that it contributes the desired weight to the overall test score.

The rating form for an essay test might contain a rating scale for each question, a checklist for each question, or a combination of the two formats. In fact, the rating plan for one question might contain some components to be checked as present and others to be rated on a quality scale. Consider the hypothetical rating form in Table 8.7 for a four-question essay test. A checklist containing five components is used to score the first question; the second and third questions are scored using rating scales that contain 16 points each; and a checklist worth 6 points is used to score the fourth question. Using such a form, the topic of each question can be inserted beside the question number, and the components can be inserted beside the capital letters. The number of points a student earns can be inserted in the space beside each question and then multiplied by the desired weight, if additional weighting is needed to balance the questions. In the example, weights were chosen to increase the value of questions 1 and 4 to approximate the value of questions 2 and 3. Using such a rating form to score the questions on an essay test will help you be more objective during scoring, show students where they earned and lost points, and aid your analysis of the group's responses, which is necessary for evaluating instruction and planning review sessions.

TABLE 8.7 Hypothetical Rating Form for an Essay Test Containing Four Questions

Name ______ Date ______ Score 53.5
(62)

Question 1 (Checklist)

Present	Component	Score		Weight	=	Total
✓ A.	______					
✓ B.	______					
✓ C.	______	4/(5)	×	3/Wt.	=	12
___ D.	______					
✓ E.	______					

Question 2 (Rating Scale)

Components	Ratings (1)	(2)	(3)	(4)	Score		Weight	=	Total
A. ______	___	✓	___						
B. ______	___	___	✓		15/(16)	×	1/Wt.	=	15
C. ______	___	___	___	✓					
D. ______	___	___	✓						
E. ______	___	___	✓						

Question 3 (Rating Scale)

Components	Ratings (1)	(2)	(3)	(4)	Score		Weight	=	Total
A. ______	___	___	✓						
B. ______	___	✓	___		14/(16)	×	1/Wt.	=	14
C. ______	___	___	___	✓					
D. ______	___	✓	___						
E. ______	___	___	✓						

Question 4 (Checklist)

Present	Component	Score		Weight	=	Total
✓ A.	______					
___ B.	______					
✓ C.	______	5/(6)	×	2.5/Wt.	=	12.5
✓ D.	______					
✓ E.	______					
✓ F.	______					

Evaluating Checklists and Rating Scales

Before you use a checklist or rating scale to score students' work, you should evaluate the form and revise it if necessary. First, select two or three students' responses or products and rate their work using the form. Determine whether the components on the form are observable and whether they are listed on the form in the sequence they most frequently occur in the products. Delete components that are not observable, change the sequence of components if necessary, and add components observable in the products that you may have inadvertently overlooked. When adding components, be sure they are appropriate for the instructional goal measured and not an artifact of the particular subset of responses you have chosen. Adjust students' scores on the few products rated to reflect any changes made in the rating form. Second, score each paper again. It is good to allow some time to lapse between the two ratings so you forget the scores assigned the first time. Compare the two ratings you assigned each paper and locate any inconsistencies. Where inconsistent scores occur, either revise the form to include fewer categories of quality that need to be differentiated or write more descriptive titles for each of the categories.

You might also ask a colleague to rate the same papers using the form. Compare your ratings with those of your colleague; discuss any inconsistencies; locate ways in which the form can be changed to improve consistency; and revise the form as needed. Once you begin to score students' work, do not alter the form. Any changes at this point may produce inconsistent ratings.

AVOIDING COMMON ERRORS IN TEST DEVELOPMENT, SCORING, AND GRADING

Development Errors

In producing essay and product development tests, teachers tend to commit two kinds of errors. First, some teachers teach skills at the lower levels of learning and test them at the higher levels. You should not teach lessons at the knowledge and comprehension levels and then expect students to perform at an application level on the test. Likewise, you should not teach at the comprehension and application levels and then write tests to measure skills at the analysis, synthesis, or evaluation levels. If students are required to develop a product or analyze a piece of literature, they should have been instructed at these levels, and been given ample opportunity to practice these skills before the posttest. To avoid this problem, make sure that novel items used on the posttest mirror those used on practice tests.

Teachers also err in constructing essay and product development tests when the items or directions do not provide adequate guidance. One way to ensure that items and directions are clear and the task posed is feasible is to construct the evaluation form before you administer the test. Constructing the evaluation form will help ensure that you have a clear notion of what you are asking students to do. After the evaluation form is complete, recheck your questions and directions to see whether they provide adequate guidance for the responses you anticipate.

Scoring Errors

Two common scoring errors are inconsistency and bias. Even when teachers use checklists and rating scales they can score students' responses inconsistently. They may become stricter or more lenient as they work through a set of papers, and their attitudes can change when they become tired, hurried, or bored. You can avoid such inconsistencies in several ways. First, instead of marking all the essay questions on one student's paper, score all the students' responses to one question before going on to the next question. Complete the scoring of one question during one marking period rather than in several. After you have completed all the papers on each question, check your consistency by again scoring the first few papers. If a student's first and second scores are the same, your scoring was probably consistent for all students. Noticeable inconsistencies in marking might indicate that your attitude, rather than the quality of students' work, determined test scores. If the scores are different, you should rescore the tests.

Besides attitude, the inclusion of too many quality categories for each component can lead to inconsistent scoring. For example, on the paragraph rating scale in Table 8.6, the topic, supporting, and concluding sentence components were each assigned a weighted score of eight. However, this value was created by breaking down each component into subcomponents, limiting the number of quality categories for each, and using a weighting factor to arrive at a score of eight. A teacher who decided that this was too much work might have developed the scale shown in Figure 8.4. Although it would take much less time to develop such a scale, the probability of the teacher's marking paragraphs inconsistently is considerably greater. It would be more dif-

FIGURE 8.4 Rating Scale

Component	*Rating*
	Inadequate Excellent
I. Topic Sentence	1 2 3 4 5 6 7 8
II. Supporting Sentences	1 2 3 4 5 6 7 8
III. Concluding Sentence	1 2 3 4 5 6 7 8

ficult to identify distinct quality descriptors for all eight category numbers. It would also be more difficult to differentiate consistently between adjacent categories (e.g., determining when a student should receive a 3 instead of a 4 or a 6 instead of a 7). When inconsistencies are present, you might want to revise your rating scale besides using strategies to minimize potential attitude shifts before rescoring tests.

Scoring bias is another common problem. Some teachers give all the students high test scores regardless of quality. Others give all average or all very low marks. In your own marking, you should compare the range of scores you assign with the heterogeneity or homogeneity of the group. If you have a homogeneous group of students, a small range of high, average, or low scores may accurately reflect the quality of students' work. If you are assigning similar scores to a heterogeneous group, however, your marking may be biased.

Teachers' opinions of individual students can also bias their scoring. Teachers are often inclined to score particular students' papers more leniently than they score others, and disruptive or uncooperative students sometimes receive scores lower than they deserve. Students who are struggling can evoke a great deal of sympathy, and teachers who want to encourage their efforts may score these students' papers quite leniently. However you choose to justify unearned scores, keep in mind that they are, indeed, unearned.

Because teachers are often unaware of their bias, they may find it difficult to avoid. Test specialists recommend that teachers maintain objectivity by dissociating students' names from their responses. For example, you could instruct your students to put their names on the backs of their papers or you could cover their names with removable labels until you finished scoring the papers. Students could also write their names and an identifying number on a separate sheet of paper and indicate only the identifying number on their work. If you have difficulty scoring papers that do not identify the student, your impressions of students are probably influencing the way you score. If this is the case, you should keep in mind that you are scoring responses and not individuals.

A Grading Error

One common grading error is to count practice tests as posttests and include them in the term grade. This tends to happen more on essay and product development practice tests than on objective-style practice tests because of the time and effort required for essay and product tests by both the teacher and students. But the required effort is not what differentiates between practice exercises and posttests. The purposes of practice tests, regardless of the format, are to provide students with the opportunity to rehearse skills and to provide them with corrective feedback. Despite the degree of effort required, students' scores on essay and product development practice tests should not be considered when calculating term achievement grades.

SUMMARIZING TEST SCORES AND EVALUATING PROGRESS

The scores students receive on checklists and rating scales can help a teacher determine the quality of instruction. To pinpoint problems and instructional needs, however, you will require more detailed information than can be obtained by reviewing each student's scores separately. A summary of students' performance by component will help you analyze the data efficiently.

Summarizing Test Data

Table 8.8 shows a class performance form with all the students' scores listed by component. Students' names are listed in the left-hand column and the evaluation criteria appear across the top. The students' scores on each component have been entered in the appropriate cells. The far right-hand column records students' total scores on the test. With the data thus arranged, you can summarize the group's performance for each quality category within a component. One summary method is illustrated in Table 8.9. The components are listed in the left-hand column and the categories for each component appear across the top. Empty cells indicate categories not used on the rating

TABLE 8.8 Class Performance on the Paragraph Test

Students	I Indented	II Topic Sentence		III Supporting Sentences			IV Concluding Sentence		Total Points
		Content	*Loc.*	*Rel.*	*Seq.*	*Trans.*	*Content*	*Loc.*	(17)
Baker	1	2	2	3	2	1	2	1	14
Brown	1	2	2	3	3	1	0	0	12
Egan	1	1	2	2	1	1	1	1	10
Ford	1	0	0	2	1	1	2	1	8
Graham	1	2	2	3	2	1	2	2	15
Little	1	1	1	1	1	2	0	0	7
Martinez	1	2	2	3	1	0	1	2	12
Smith	1	0	0	1	1	0	0	0	3
Taylor	1	2	2	3	3	2	2	2	17
Williams	1	2	1	1	1	1	2	2	11

form. To complete this table, the teacher would refer to the class performance form, count the students who received each rating, and record that number in the appropriate cell. With this summary, the teacher can then assess the effectiveness of instruction for each component.

A teacher looking at the summary in Table 8.9 sees that all students indented their paragraphs properly. Most students remembered to include both a topic and a concluding sentence. The numbers for the third component, supporting sentences, indicate a problem. Although most students included relevant material in their paragraphs, eight of ten students did not properly sequence their sentences. Likewise, only two of the ten students effectively used transition words. Obviously, the instruction on sequencing sentences and using transition words was not as effective as it could have been.

TABLE 8.9 Summary of Students Receiving Each Rating

Class Second Period Date of Test 10–15 Number of Students 10

Component		Rating Categories *0*	*1*	*2*	*3*	*Total Students*
I. Indentation			10			10
II. Topic Sentence	A. Content	2	2	6		10
	B. Location	2	2	6		10
III. Supporting Sentences	A. Relevance		3	2	5	10
	B. Sequence		6	2	2	10
	C. Transition	2	6	2		10
IV. Concluding Sentence	A. Content	3	2	5		10
	B. Location	3	3	4		10

A class summary form like the one in Table 8.9 can also help a teacher detect problems with the rating scale. Blank cells on a completed form may indicate that one or two components had too many rating categories. Blanks might also mean that, on that particular occasion, no one submitted work that deserved those ratings. In using your own summary forms, note blank cells and check students' performances on subsequent tests. If you never use a particular rating, you should either redefine or remove it.

Evaluating Progress

Teachers can also summarize performance data across tests to evaluate students' progress over a period of time and to judge the effectiveness of instruction and review activities. Students' skills in the same area are often measured using pretests, practice tests, and posttests, and data can be summarized across all measures to observe progress over time. Language arts is one subject in which teachers administer a series of posttests to measure progress on the same skills, such as writing sentences, paragraphs, and themes.

Table 8.10 presents a class progress form for the paragraph skills already described. The components are again listed in the left-hand column, the rating

TABLE 8.10 Class Progress Form

Class ______________ Number of Students ______________

Components	*Dates Tested*	Rating *0*	*1*	*2*	*3*
I. Indentation	10–15		10	—	—
	10–22		10	—	—
	10–29		10	—	—
	11–5		10	—	—
II. Topic Sentence					
A. Content	10–15	2	2	6	—
	10–22	1	2	7	—
	10–29		3	7	—
	11–5		2	8	—
B. Location	10–15	2	2	6	—
	10–22		1	9	—
	10–29			10	—
	11–5			10	—
III. Supporting Sentences					
A. Relevance	10–15	—	3	2	5
	10–22	—	1	4	5
	10–29	—		4	6
	11–5	—		2	8
B. Sequence	10–15	—	6	2	2
	10–22	—	7	2	1
	10–29	—	2	6	2
	11–5	—	1	4	5
C. Transition	10–15	2	6	2	—
	10–22	5	1	4	—
	10–29	1	3	6	—
	11–5	1	2	7	—
IV. Concluding Sentences					
A. Content	10–15	3	2	5	—
	10–22	3	3	4	—
	10–29	3	3	4	—
	11–5		3	7	—
B. Location	10–15	3	3	4	—
	10–22	3	4	3	—
	10–29	3	3	4	—
	11–5		2	8	—

categories appear across the top, and unused values boxes contain a short horizontal line. Four different testing dates appear beside each component to indicate that students wrote paragraphs on October 15, 22, 29, and November 5. The data for component I show that all ten students received the total possible points on all four tests. After the first test, the lessons emphasized quality in topic sentences, with a resulting improvement of performance on the second test. This time, all ten students included a topic sentence, and only one student's sentence was too specific. Their skill in placing topic sentences also improved.

In writing supporting and concluding sentences, the students' skills changed very little. After the second test, the teacher emphasized supporting sentences and transition words. On the third test, the students kept their level of performance on topic sentences and wrote better supporting sentences. Their concluding sentences received ratings similar to those on the first two tests. After the third test, the teacher focused on summarizing ideas and on writing concluding sentences. The data for test four show that students improved their skills for this component. Even though many students still did not write perfect paragraphs after the fourth week, the teacher's data indicate that the lessons on paragraph writing were effective.

USING PRODUCT RATING FORMS AS AN INSTRUCTIONAL TOOL

Used as self-assessment tools, rating scales and checklists can help students identify and use appropriate criteria for developing their responses. During practice exercises, they can use an evaluation form to rate the quality of their work and to compare their ratings with yours. You can then discuss differences in ratings with students who have difficulty finding problems in their work. Students can also use the rating forms to evaluate example answers and products that you provide. They can discuss similarities and differences among their evaluations and between theirs and yours. Using the criteria when developing their responses for practice exercises and when reviewing work samples will help students focus on the important aspects of a skill. Such involvement should help them not only produce better work, but also feel a greater sense of responsibility for the quality of their work and the scores they receive.

SUMMARY

Essay tests and product development tests measure students' skills in producing original responses. Essay questions can ask students to discuss, explain, analyze, summarize, or evaluate

something. Product development tests can be used to measure students' skill in producing a variety of verbal products, such as paragraphs, book reports, and themes. Product tests can also be used to measure students' skill in creating such objects as mobiles, photographs, and maps.

Essay and product tests allow students to select their approach to a given problem, the information they will include, and the methods of organization and presentation they will use. (Such skills cannot be measured using objective tests.) These tests enable a teacher to determine both what and how students think. Test drawbacks include the time required to develop, analyze, and score them; difficulties in scoring responses consistently and without bias; and the advantage essay tests give to students with well-developed verbal skills.

The procedure for developing and using essay and product development tests consists of four phases: (1) writing questions and directions; (2) developing scoring procedures; (3) scoring students' responses; and (4) summarizing group performance.

The criteria for developing valid and reliable test items also apply to essay questions and directions for products. The behavior, content, and conditions included should match those in the objectives; the complexity, context, and vocabulary should be appropriate for students; and grammar, punctuation, and sentence structure should be correct. Good questions and directions also provide adequate guidance for responding. Questions and directions that provide little guidance require students to exercise considerable judgment in responding. A lot of guidance guarantees that students will address specific issues, but it also limits the approaches they can take. Considering the main purpose for your test and the skill level of students will help you select an appropriate amount of guidance.

Besides describing the task to be performed, you should provide information that can help students determine the scope of the anticipated responses. You might want to include such information as how long their response should be, the amount of time they should spend, and the number of points each question or section is worth. Developing an evaluation form before administering the test will help you evaluate the clarity and feasibility of your questions or directions.

Essay and product development tests can be scored using either global or analytical scoring methods. Global scoring provides information on the overall quality of students' responses. Analytical scoring provides information on the quality of specific components and subcomponents of a response. Although analytical scoring is more time consuming than is global scoring, it provides students with better feedback and teachers with specific information they can use to analyze and evaluate the quality of instruction.

Analytical scoring requires the use of either a checklist or rating scale. Both should list the components and, if needed, the subcomponents of a skill to direct your attention to particular aspects of a response. The components and subcomponents should be taken directly from the enabling and prerequisite skills in the instructional goal framework. A checklist allows you to mark the presence or absence of a component, whereas a rating scale enables you to differentiate various quality categories for each component using a scale. The number of quality categories included on the scale should not exceed the number you can consistently judge. Verbal descriptors should be written for each quality category. Once the checklist or rating scale is designed, review the number of points allocated to each component. If the value assigned does not contribute an appropriate amount to the total score, use weighting schemes to adjust the relative value of components.

When the evaluation form is complete, review the questions and directions to ensure that they provide students with the guidance necessary to produce the responses you anticipate. At this point, revisions can be made in the questions, directions, or evaluation form as needed. After the test is administered, the rating form should again be evaluated using a few tests. Determine whether the components listed are observable in students' responses; whether they are sequenced in the most efficient order on the form; and whether you can score responses consistently. Make necessary revisions before beginning to score students' work.

Teachers experience two problems when rating students' responses: inconsistent scoring and bias. Inconsistent scoring can be caused by shifts in attitude that may lead to stricter or more lenient scoring. It can also result when you include more categories of quality for each component than you can consistently judge or use vague descriptors for each category. Bias results from marking all students as good, average, or poor regardless of the quality of their work. Bias also results when you consider not only the nature of the response, but also the characteristics of the student who made the response. After scoring all papers, rescore the first two or three to see whether you were consistent. Also check to see whether the range of scores assigned is reasonable given the heterogeneity or homogeneity of the group. If inconsistency or bias is detectable, you should rescore the papers.

Information about a group's test performance can help you evaluate the instruction students received. After scoring is completed, make a table to summarize students' scores by component and subcomponent. Areas of poor performance will become obvious, and instruction for related components can be reviewed and revised. When multiple tests are administered to measure the same skills, make a table to summarize students' scores across measures. Such summaries enable you to monitor the group's progress.

PRACTICE EXERCISES

1. Which of the following skills could be measured only with essay or product development tests?
   a. Recall verbal information
   b. Design a solution to a problem
   c. Discriminate among examples and nonexamples of a concept
   d. Use a rule to solve a problem
   e. Decide how two or more things should be compared and make the comparisons
   f. Evaluate a rule, a principle, an object, or a product
   g. Synthesize facts and information
   h. Organize facts and information
   i. Follow rules to produce an original piece of work
   j. Recall, select, organize, and present information
2. Table 8.11 summarizes information and rules for outlining a passage. Use the table to:
   a. Write instructions for a test that measures students' skills in outlining given passages. Assume that the given passages each have three main topics, three secondary ideas, and no level-three or level-four ideas.
   b. Develop an evaluation form for scoring students' outlines. (See Table 8.15 in the Feedback section.)
   c. Develop a class performance form that includes each student's rating on each component listed on your rating scale. (See the top part of Table 8.12 for feedback.)
   d. Develop a class summary form that can be used to summarize the number of students who received each rating. (See the bottom part of Table 8.12 for feedback.)
3. Summarize the class performance data in Table 8.12.
   a. Count the students who received each rating and record the number on the summary form at the bottom of the table. (See Table 8.16 in the Feedback section.)
   b. Review your completed summary form and identify the components that require more instruction.

TABLE 8.11 Information and Rules for the Instructional Goal "Outline a Passage"

Components	*Define Levels*	*Paraphrase Entries*	*Sequence Entries within Level*	*Label Entries*	*Align Entries in Outline*
Main topics	Define main topics as key main ideas in total passage.	Use sentence or topic outline. Do not mix sentences and phrases in one outline.	Place entries in order defined by passage.	Use Roman numerals to label main ideas.	Place all main topics flush with left margin.
Secondary topics	Define secondary topics as key ideas within each main topic. There should be at least two entries.	Make all entries within a main topic parallel in structure. Structure may differ from that of main topics and from that of secondary topics for other main ideas.	Place secondary topics in order defined by passage.	Use capital letters to label secondary topics.	Indent secondary entries five spaces from left margin and align.
Level-three topics	Define level-three topics as key ideas within each secondary level topic. There should be at least two entries.	Make all entries within a secondary level topic parallel.	Place level-three topics in order defined by passage.	Label level-three topics with Arabic numerals.	Indent all level-three topics ten spaces from left margin and align.
Level-four topics	Define level-four topics as key ideas within each level-three topic. There should be at least two entries.	Make all entries within a level-three topic parallel.	Place level-four topics in order defined by passage.	Label level-four topics with lower case letters.	Indent all level-four topics fifteen spaces from left margin and align.

4. Score the three paragraphs entitled "Washing a Dog" in Table 8.13 using the paragraph rating form in Table 8.14. Assume the paragraphs were written by sixth-grade students. Record your scores for each paragraph on a separate sheet. When you have finished:
   a. Compare the scores you assigned with those of classmates.
   b. Discuss inconsistencies and try to determine why the scores were different.
   c. If you attribute differences to inadequacies in the rating scale, suggest ways to improve the form.
   d. Compare your scores to those in Table 8.17 (in Feedback section) that reflect the judgment of another rater.

Enrichment

5. Select a skill from your own field that can only be measured with an essay or product development test and do the following:

TABLE 8.12 Class Performance and Performance Summary Forms for the Outline Test

I. Class Performance Form

Students	Main Ideas A	B	C	D	E	F	Secondary Ideas A	B	C	D	E	F	G	Total Score
Allen	3	2	1	2	2	2	3	3	2	1	2	2	2	27
Baker	1	2	1	2	2	2	1	2	2	1	2	2	2	22
Carter	1	2	1	2	2	2	1	2	2	1	2	2	2	22
Doyle	2	2	1	2	2	2	1	2	2	1	2	2	2	23
Egan	3	2	1	2	2	2	3	3	2	1	2	2	2	27
Frank	1	1	1	1	2	2	1	1	1	1	1	2	2	17
Garcia	3	1	1	2	2	2	3	3	1	1	2	2	2	25
Howe	1	1	1	2	2	2	2	2	1	1	2	2	2	21
Jackson	3	2	1	2	2	2	3	2	2	1	2	2	2	26
Little	3	2	1	2	2	2	3	1	2	1	2	2	2	25

II. Class Summary Form

Class ____________ Date ________ Number of Students ________

	Rating Categories 0	1	2	3	Total Students
I. Main Ideas					
A. Main ideas identified	—				
B. Clearly paraphrased	—				
C. Parallel structure	—				
D. Consistent order	—				
E. Consistent labels	—			—	
F. Alignment	—			—	
II. Secondary Ideas					
A. Secondary ideas identified	—				
B. Related to main ideas	—				
C. Clearly paraphrased	—				
D. Parallel structure	—				
E. Consistent order	—				
F. Consistent labels	—			—	
G. Alignment				—	

a. Analyze the skill to identify its subordinate skills and create a goal framework that illustrates the relationship among the skills.
b. Write a behavioral objective for the skill; be sure to include relevant conditions.
c. Write an essay question or questions or the instructions for product development. Include adequate guidance for the skill being measured and for the sophistication of target students.
d. Develop a scoring checklist or rating scale.
e. If you have students' work available, score their work using your form.
f. Revise the rating form if necessary.
g. Develop a class performance form and summary form that will illustrate students' performance on each component of the test.

FEEDBACK

1. b, e, i, j
2. a. Instructions for a test of outlining skills.

Example 1
Read the article titled ______ . Make a topic outline of the information in the article. You have 20 minutes to complete the task.

Example 2
Make a sentence outline of the information in the article entitled ______ . You have 30 minutes to complete the following:

1. Read the article.
2. Review the rating form attached to the article that will be used to score your outline.
3. Select and organize the information you think should be included.
4. Write your outline.
5. Use the rating form to:
   a. review your outline and correct any problems you find.
   b. score your outline.
6. Turn in both your outline and your completed rating form when you have finished.

b. See Table 8.15.
c. See the top part of Table 8.12 (in Practice Exercises).
d. See the bottom part of Table 8.12 (in Practice Exercises).

3. a. See Table 8.16.
   b. You may have concluded that:
   1. Approximately half the class had difficulty differentiating between main and secondary ideas. Thus, you need to provide more instruction on the levels of ideas.
   2. All students had problems with parallel structure. They need more instruction on constructing parallel entries.
   3. Eight students had problems relating secondary ideas to the main idea. They need additional instruction in sorting secondary ideas by topic.
   4. Three students need more instruction on paraphrasing ideas for a topic outline.
4. Compare your scores with those of classmates and students included in Table 8.17.

TABLE 8.13 Three Paragraphs Entitled "Washing a Dog"

Paragraph 1

One of my chors on Saturday is to wash my dog Rover. I use warm water to get him wet all over. Then I rub him with a special soap that doesn't hurt his eyes and keeps his coat glossie. When he is clean, I rinse him twice to make sure all the soap is gone. After he shakes off most of the water, I rub him with a clean old towel. He thinks I do a good job.

Paragraph 2

Put a wash tub in the yard. Fill it with water and sope. Put the dog in the tub. Scrub the dog. Rinse the dog with a hose. Stand back when the dog gets out of the tub.

Paragraph 3

My dog hates to get a bath. When anyone gets out the tub she runs and hides. It is easy to find her because she always hides in the same place. You have to hold her the whole time or she will run away again. She shivers and barks the whole time. My cousin's dog likes to get a bath.

TABLE 8.14 Rating Scale for Evaluating Paragraphs

Components			
I. Indentation	Not Indented 0	Clearly Indented 1	
II. Topic Sentence			
A. Quality	Not Present 0	Vague 1	Clearly Introduces Topic 2
B. Location	Not Present 0	Misplaced 1	Logically Placed 2
III. Supporting Sentences			
A. Content	Some Irrelevant Information 1	All Relevant Information 2	Thoroughly Develops Topic 3
B. Sequence	Illogical Order 1	Some Order 2	Logical Order 3
C. Transition	No Transition 0	Some Transition 1	Smooth Transition 2
IV. Concluding Sentence			
A. Content	Not Present 0	Not Comprehensive 1	Clearly Summarizes Topic 2
B. Location	Not Present 0	Misplaced 1	Logically Placed 2

TABLE 8.15 Rating Scale for Evaluating Outlines

Name ________________ Date ________ Score ________

Total (35)

I. Main ideas			
A. Main ideas identified	Secondary ideas included 1	Some ideas missing 2	All main ideas included 3
B. Clearly paraphrased	Some meaning changed 1	Too brief/ wordy 2	Clearly paraphrased 3
C. Parallel structure	Mixed sentence/ phrase 1	All sentences or phrases 2	Consistent word format 3
D. Consistent order	Out of order 1	Most in order 2	All in order 3
E. Consistent labels	Inconsistent 1	Consistent 2	
F. Alignment	Flush with left margin 1	Aligned 2	
II. Secondary ideas			
A. Secondary ideas identified	Main ideas included 1	Some ideas missing 2	All secondary ideas included 3
B. Related to main idea	Scrambled placement 1	Most within right topic 2	All within right topic 3
C. Clearly paraphrased	Some meaning changed 1	Too brief/ wordy 2	Clearly paraphrased 3
D. Parallel structure	Mixed sentence/ phrase/word 1	All sentences, phrases, or words 2	Consistent word format 3
E. Consistent order	Out of order 1	Most in order 2	All in order 3
F. Consistent labels	Inconsistent 1	Consistent 2	
G. Alignment	Not indented 0	Indented five spaces 1	Aligned 2

TABLE 8.16 Class Summary Form for Outline Test

Class ________________ Date ________ Number of Students ________

	Rating Categories			
Component	0	1	2	3
I. Main ideas				
A. Main ideas identified	—	4	1	5
B. Clearly paraphrased	—	3	7	
C. Parallel structure	—	10		
D. Consistent order	—	1	9	
E. Consistent labels	—		10	—
F. Alignment	—		10	—
II. Secondary ideas				
A. Secondary ideas identified	—	4	1	5
B. Related to main idea	—	2	5	3
C. Clearly paraphrased	—	3	7	
D. Parallel structure	—	10		
E. Consistent order	—	1	9	
F. Consistent labels	—		10	—
G. Alignment			10	—

TABLE 8.17 Completed Ratings for Paragraphs Entitled "Washing a Dog"

				Scores		
Components				Parag. 1	Parag. 2	Parag. 3
I. Indentation	Not Indented 0	Clearly Indented 1		1	0	0
II. Topic Sentence						
A. Quality	Not Present 0	Vague 1	Clearly Introduces Topic 2	2	0	2
B. Location	Not Present 0	Misplaced 1	Logically Placed 2	2	0	2
III. Supporting Sentences						
A. Content	Some Irrelevant Information 1	All Relevant Information 2	Thoroughly Develops Topic 3	3	3	1
B. Sequence	Illogical Order 1	Some Order 2	Logical Order 3	3	3	2
C. Transition	No Transition 0	Some Transition 1	Smooth Transition 2	2	0	1

TABLE 8.17 *(continued)*

				Scores		
Components				*Parag. 1*	*Parag. 2*	*Parag. 3*
IV. Concluding Sentence						
	Not Present	Not Comprehensive	Clearly Summarizes Topic			
A. Content	0	1	2	2	0	0
	Not Present	Misplaced	Logically Placed			
B. Location	0	1	2	2	0	0
			Total Score (Out of 18)	17	6	8

SUGGESTED READING

Ebel, R. L., and Frisbie, D. A. (1986). *Essentials of educational measurement.* Englewood Cliffs, N.J.: Prentice-Hall, pp. 126–36.

Gronlund, N. E. (1985). *Measurement and evaluation in teaching.* New York: Macmillan Publishing Company, pp. 213–28, 383–405.

Mehrens, W. A., and Lehmann, I. J. (1984). *Measurement and evaluation in education and psychology.* New York: CBS College Publishing, pp. 94–124, 203–10.

Nitko, A. J. (1983). *Educational tests and measurement.* New York: Harcourt, Brace, Jovanovich, pp. 141–55, 243–79.

Popham, W. J. (1981). *Modern educational measurement.* Englewood Cliffs, N.J.: Prentice-Hall, pp. 274–84, 309–27.

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement (OERI)
Educational Resources Information Center (ERIC)

ERIC®

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: Lies; Damned Lies; Statistics & Law School Grades

Author(s): PAUL T. WANGERIN

Corporate Source: The Science & Art of Law Teaching

Publication Date: 1994

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic/optical media, and sold through the ERIC Document Reproduction Service (EDRS) or other ERIC vendors. Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce the identified document, please CHECK ONE of the following options and sign the release below.

☒ **Check here** Permitting microfiche (4"x 6" film), paper copy, electronic, and optical media reproduction	Sample sticker to be affixed to document "PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)." **Level 1**	Sample sticker to be affixed to document "PERMISSION TO REPRODUCE THIS MATERIAL IN OTHER THAN PAPER COPY HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)." **Level 2**	□ **or here** Permitting reproduction in other than paper copy.

Sign Here, Please

Documents will be processed as indicated provided reproduction quality permits. If permission to reproduce is granted, but neither box is checked, documents will be processed at Level 1.

"I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce this document as indicated above. Reproduction from the ERIC microfiche or electronic/optical media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries."

Signature:	Position: Associate Professor of Law
Printed Name: PAUL WANGERIN	Organization: John Marshall Law School
Address: John Marshall Law School 315 S. Plymouth Court Chicago, IL 60604	Telephone Number: (312) 427-2737
	Date:

OVER