Global Perspectives on Accounting Education
Volume 10, 2013, 61-80

DISPARATE METHODS OF COMBINING TEST AND ASSIGNMENT SCORES INTO COURSE GRADES

Daniel Tinkelman
Zarb School of Business
Hofstra University
Hempstead, New York
USA

Elizabeth Venuti
Zarb School of Business
Hofstra University
Hempstead, New York
USA

Linda Schain
Zarb School of Business
Hofstra University
Hempstead, New York
USA

ABSTRACT

Professors utilize a variety of assessments, such as tests, quizzes, homework, projects and oral presentations to evaluate student mastery of subject matter. The various assessment scores are aggregated to arrive at a course grade. Different methods of aggregation can result in different course grades. This paper outlines the nature of the differences using illustrative examples and demonstrates the disparities that result from applying different aggregation methods on actual data for more than 1,000 students in accounting classes. Possible modifications to grading schema that would

Acknowledgments: The authors would like to thank Ralph Polimeni, Jacqueline Burke, Cheryl Lehman and Nathan Slavin for sharing student data. They also would like to thank participants in the American Accounting Association 2010 Mid-Atlantic and Annual Conferences and a workshop hosted by Hofstra University's Center for Teaching and Scholarly Excellence for helpful comments.

reduce disparities and implications for course design, for student motivation and for faculty-student communication, are also discussed.

Key words: Assessment, grading methods, aggregation of scores

INTRODUCTION

Business school faculty members receive little guidance on grading. In some ways, the situation has changed little since Spence (1927, p. 2) wrote:

> An instructor comes to an institution to teach and in most cases finds out nothing more about grades than the manner in which they are to be reported – in letters or in numbers. Persons in charge do not realize there is any problem involved. Everyone gives grades; everybody must know how to give them. It is like reading or writing. To attempt to make any suggestions would be an affront.

Ekstrom and Villegas (1994) report that, in their study of 14 colleges, only one third of the responding department chairs reported having formal meetings to discuss grading with faculty.

The motivation for exploring the assignment of grades came from a discussion among faculty members in our department, which highlighted the impact differing grade aggregation methods could have on instruction and student motivation. Midway through the semester, some faculty who used the "total points" method to aggregate grades firmly counseled students who had done poorly on coursework thus far to withdraw without receiving a formal grade because the students could no longer earn enough total points to achieve a passing mark. Another professor, who converted each individual assessment score to a grade on the 4.0 scale before aggregating them, noted that in his class each of these students could still theoretically eke out a passing grade, with an A on the final.

Discussions with faculty and preliminary presentations revealed that many professors were unaware of the issues inherent in aggregating scores into course grades, or the possibilities of divergent grades. In looking at the issue further, it became clear that while there has been much theoretical discussion of issues of aggregating scores into grades, there has been little empirical work on how disparities in college course grades result from different aggregation techniques. Empirical evidence on the variability of student performance and the differential impact of grading methods is potentially important, because it affects both the perceived fairness of the grading system and the incentive effects of grades.

The purpose of this paper is to sensitize faculty as to how grade aggregation methods can impact course grades assigned. Illustrative examples are used to demonstrate the potential disparity in aggregation methods. Varying aggregation methods are then applied to real course data for 1,062 undergraduate students enrolled in 24 accounting classes at one northeastern U.S. university. The aggregation methods generally arrive at the same course grades for students who perform consistently on all coursework; however, there is a significant likelihood that the methods will give differing course grades for students with inconsistent performance.

The remainder of the paper is organized as follows. The next section briefly reviews applicable literature and describes three different baseline methods of combining scores: "total possible points"; "weighted average letter grade"; and the "median method." It also includes a variant of one method, "modified total possible points." The following section presents illustrative examples of the comparative impact of the methods and discusses modifications to the methods that might reduce the number and severity of grading disparities. After that, an empirical analysis of actual student data for 1,062 students in accounting courses is presented, and the final section of the paper includes discussion and recommendations.

AGGREGATION METHODS AND LITERATURE REVIEW

Aggregation Methods

Brookhart (1999) describes the grading methods that appear to be in the most common use: the "total possible points"; "weighted average letter grade"; and the "median method."[1] In addition to considering these three methods, which we refer to as "baseline" methods, a fourth method, "modified total possible points," is also considered in this paper.

"Total Possible Points" assumes that there are a certain number of possible points to be earned in the entire course, e.g., 600 points. Each assignment or test has an assigned maximum number of points, with the instructor assigning more points to those items considered more important. For example, there may be two projects, each worth 100 points, and two exams, each worth 200 points. A student achieving 500 points out of 600, which is 83.3% of the possible points, might thereby earn a B using the standards provided in Table 1.[2] This is mathematically identical to taking a weighted average of numeric scores.

Total Possible Points is consistent with how instructors normally combine the scores on the questions within a test to determine the overall test score. It has the virtue of simplicity but has various theoretical drawbacks discussed below. This method also makes it difficult for students to recover from poor performances early in the term. If a score of 360 out of 600 (60% of possible points) is needed to pass, any student earning fewer than 160 of the possible 400 from the first exam and the two assignments is mathematically incapable of passing the course, even with a perfect final exam score.

Instructors using the Weighted Average Letter Grade method convert the score on each assignment or test into the numerical equivalent of a letter grade, using the 4.0 scale.[3] The instructor then computes a weighted average of the various assignments and tests. This method of combining

[1] No empirical data on frequency of use was presented. Brookhart (1999) also considers a broad category of methods that are referred to collectively as "Holistic." This very general category includes many different aggregation methods. Grades can be combined using a judgmental rubric that bases grades on how well students have met various defined goals of the course. It may be very appropriate in situations when the instructor's goal is to have students achieve certain ending proficiency levels. In this case, the ending grade need not reflect performance on the individual assignments in a mechanistic manner. While such methods may be the most appropriate for certain goals, these holistic approaches are not readily susceptible to modeling, and are not discussed further in this study.

[2] Brumfeld (2004) reports survey data indicating that 96% of surveyed U. S. colleges and universities use some form of the four-point scale, and a majority of schools use pluses and minuses. The particular form of this scale used in Table 1 may differ from that used in some schools. Note that in addition to differences in grade aggregation techniques among professors, there may also be discrepancies in grading standards.

[3] See Waters (1979) for a similar approach, but using a zero to 11 scale.

TABLE 1

Grading Standards

Letter Grade	Percent Required	Required Value on 4.0 Scale
A = 4.0	92 or higher	3.71 to 4.00
A- = 3.7	90, 91	3.50 to 3.70
B+ = 3.3	88, 89	3.30 to 3.49
B = 3.0	82 to 87	2.71 to 3.29
B- = 2.7	80, 81	2.50 to 2.70
C+ = 2.3	78, 79	2.30 to 2.49
C = 2.0	72 to 77	1.71 to 2.29
C- = 1.7	70, 71	1.50 to 1.70
D+ = 1.3	68, 69	1.30 to 1.49
D = 1.0	60 to 67	0.70 to 1.29
F = 0.0	Under 60	Under 0.70

Notes: In situations where a median score falls between two grades, the higher grade is assigned.

assignments within a course is consistent with how schools combine grades across individual courses into overall student grade point averages. It also is consistent with McLachlan and Whitten's (2000) advice to convert scores into grades before aggregating them, but does involve a loss of information as compared to the previous method.

In the situation described above, assume the student received a 55 (F=0) on one assignment, an 85 (B=3) on the second assignment, a 95 (A=4) on one test and an 85 (B=3) on the second test. Since the tests have double weight, the instructor would average the following six numbers: 0, 3, 4, 4, 3, 3, resulting in 2.83. Using the standards in Table 1, the student has earned a B. In this example, the student has earned 500 out of 600 possible points, and the B is the same as under the Total Possible Points method. However, the results need not always be the same. If the student received a zero on the first assignment, instead of a 55, under the Weighted Average Letter Grade method the ending grade is unaffected, since both a 55 and a zero are F's. Under the Total Possible Points method, the student's performance has slipped from 500 points to 445, and the course grade would fall from B to C.

The Median method of aggregating assignment grades was recommended by McLachlan and Whitten (2000). Where different assignments have different weights, an instructor using this process would adjust the process to count the assignments with higher weights more heavily. In the example discussed previously, with two assignments and two tests, the student had an F (55) and a B (85) on the assignments, and a B (85) and an A (95) on the (double-weight) tests. The six scores would be F, B, B, B, A, A, or, alternatively, 55, 85, 85, 85, 95, 95. The median would be a B (or 85). This is the same result as the other two methods. Like the Weighted Average Letter Grade method, but unlike the Total Possible Points method, the result would still be a B even if the student had not submitted the first assignment.

In addition to the three methods outlined above, this paper also considers the Modified Total Possible Points method. This modification of Total Possible Points sets the minimum grade on any assignment at fifty, rather than zero. The purpose of this modification is to mitigate the disproportionate effect of grades below 50 on a student averages and to better align assignment grades with the traditional 4.0 letter grade scale. The disadvantage of this method is that it reduces the punitive consequences of a student failing to complete an assignment.

In the situation described above, the student with scores of 55, 85, 95 and 85 on two assignments and two exams, respectively, would earn the same course grade under both the total possible points and modified total possible points methods. However, if the score on one of the two assignments is zero, rather than 55, the Total Possible Points method would assign a C for the 445 points earned (74%), while under the Modified Total Possible Points method, the student would earn a B for the 500 points (83%). The Median and Weighted Average Letter Grade methods are indifferent to whether the assignment grade is zero or 55.

Literature Review

The theoretical issue of how to combine data from various assignments into a single overall grade or decision has been extensively discussed in the educational, psychometric, and statistical literature. French (1985) notes that the issue was first raised in a 1904 paper by Spearman. Similar mathematical issues arise in combining the candidates' scores on various parts of a single test, combining various test scores into a course grade, and determining whether a candidate's performance on various individual measures qualifies the candidate for professional licensure, or an honors degree in a U. K. university. See Simonite (2000) and Yorke, Bridges and Woolf (2000). This paper does not review the extensive related literature. Yorke (2008) summarizes the issues and research in the field of grading student achievement, and concludes that at present there is no general agreement or understanding of how individual scores should be aggregated. Some of the major problems are hereafter explained, together with some proposed solutions, where applicable.[4]

Simple addition (or weighted averaging), of component score, referred to herein as the "Total Possible Points Method," is problematic for several reasons. First, any single measure of multidimensional data involves a loss of information (cf. Yorke, p. 152). Also, scores may not meet

[4] A practical solution, from various sources (e.g. McLachlan and Whiten, 2000), is that examiners should carefully consider all borderline cases to ensure the grades meet intuitive standards of fairness. See also Rowntree, 1987; Cresswell, 1988; and Wiliam, 1995. Another suggestion is to recognize that each aggregation method is imperfect, and to choose the one that minimizes the loss function that applies to each specific application. See Biggins, Loynes and Walker (1986). Another suggestion, by Looney (2003), is to establish a definitional grading system by defining performance thresholds for each grading component (e.g., points earned on an exam, percentage of classes attended, number of times participated in class discussion, etc.). Providing there are enough components contributing to the overall grade, others suggest throwing out the highest and lowest scores (Walker, 2006).

the necessary mathematical requirements for mathematical computations based on percentages. Dalziel (1998), citing Otto Hölder, notes that for data to be quantitative, they must have both the properties of order and additivity. While, in normal arithmetic, the sum of 4/10 and 4/10 equals 8/10, instructors would not normally combine scores of 4 out of 10 points on two assignments into one score of 8/10. Dalziel cites work by the Ferguson committee of scientists from 1940, who concluded that variables then being studied by psychologists were not quantitative. See also McLachlan and Whitten (2000) who note these mathematical problems and, therefore, suggest that the median or the inter-quartile range would be better measures of overall performance than mean scores.

The planned weight of a component exam and its actual impact on the rank ordering of candidates may be very different. See Adams and Wilmot (1981) and McLachlan and Whiten (2000), among many others. Assignments may be scored out of varying amounts of points, and the mean, median and standard deviations of the scores will differ among assignments. Where the dispersion of performance among assignments differs, those assignments with higher variability will have disproportionate impacts on the final rank ordering of the candidates, unless steps are taken to standardize the scores. Frith and Macintosh (1984), Rowntree (1987) and Miller, Imrie and Cox (1998) suggest standardizing the distribution of scores for each assignment prior to combining. See also McLachlan and Whitten (2000). Rowntree (1987) advises stretching the less scattered distributions so that the standard deviation of scores matches that of the most scattered distribution. However, as Cresswell (1988) notes, standardizing scores involves a loss of information, which he believes on average to be relatively small, and this procedure may be less reliable than combining the raw marks.[5] Yorke (2008) also criticizes the idea of standardizing distributions, saying it introduces additional error into the grading process. Where components may vary in difficulty, various weighting schemes have been proposed. See, for example, French (1981).

Another problem with summing individual scores is that, when instructors use 100-point scales, scores of zero tend to have disproportionate impact. See Reeves (2004), "The Case Against Zero." He urges instructors to make the difference between the lowest D and the lowest F ten points, by making the bottom score fifty, rather than zero.[6] The Modified Total Points method adopts this suggestion. However, proponents of the use of zero in grading argue that there should be consequences for not doing work, and that assigning a score of 50 for work that was not completed rewards bad behavior (Butler, 2004).

Thus, prior theoretical literature supports the following key ideas. There is always a loss of information in summarizing multi-dimensional data into a single score. There is no one clearly superior method of aggregating scores. Some algorithms that are either in use or have been

[5] Creswell (1988) finds that increasing the number of independent grading components increases the reliability of the composite grade. However, as Looney (2003) points out, there are offsetting disadvantages. When the number of grading components increases, students will have difficulty understanding which course components and learning objectives are most important and most valued by the instructor. Also, the time to determine overall course grades will likely increase.

[6] Reeves (2009) argues that the use of zero is unduly punitive. "Even Dante's worst offenders were consigned to the ninth – not the 54th – circle of hell. Poets, it seems, understand interval data better than professors in the hard sciences do" (Reeves, 2009). The difference between two very low scores also may not measure differences in performance in a meaningful way. Consider, for example, that a score of 50% on a true-false test may be expected by guessing, and may therefore show the same amount of knowledge as a score of zero on a problem set.

recommended include summing (or averaging) all scores, a median method and converting scores to grades before aggregating. Different methods can in theory arrive at different grades, some of which may seem unfair. See, for examples, French (1981), Rowntree (1987), Cresswell (1988) and Reeves (2004). Aggregation problems are especially likely when individual assignment scores are widely dispersed, when the degree of dispersal varies among assignments, when scores of zero out of 100 are recorded, and when a student's scores vary widely across assignments. Such variation could reflect either differences in the difficulty of the assignments or differences in student effort. Student performance is more reliably measured by a larger number of assignments.

The literature does not, however, address the empirical question of how often, in real life, do the different methods actually result in differing grades? How often, in real situations, are individual assignment grades widely dispersed, or how often does a student get a zero on one assignment and a perfect score on another?

In contrast to the extensive theoretical literature, few studies have investigated student performance variability and the related impact of the different grade aggregation methods, using real data. None were found that dealt with U. S. college data. Simonite (2000) looked at U. K. data related to the awarding of honors degrees for 423 students at one university. The number of honors degrees that would be awarded in Simonite's (2000) data increased by 20% when using a method that considered only the 16 best scores, rather than all scores. In this data, using the median to measure performance resulted in more variability than did using means. Woolf and Turner (1997) find that 15% of U. K. degree classifications could change if different methods of aggregation than used by the candidates' home school were used. Dalziel (1998) presented simulation results for 1,039 students in an Australian psychology class, with 17 different assignments that were aggregated into course grades ranging from failures to "high distinction." Different aggregation methods (including a total points method, a method that varied the gaps between letter grades, a median-type approach, and a method adjusting for error in the individual scores) produced differing final grades for significant numbers of students. About 24% of the grades differed between the total points method and the median method. See also Wilson (2008), who created simulated data to compare five different proposed grading schemes in an Australian medical school, and found that failure rates ranged from 5% to 30%.

ILLUSTRATIONS

Illustrative Examples

Each of the examples in this section assumes that the instructor mechanistically follows one of the grading systems described in the previous section (Total Possible Points; Modified Total Possible Points; Weighted Average Letter Grade; and Median), and applies the grade definitions set forth in Table 1. The maximum grade is an A, with a numeric value of 4.0. All plus (minus) grades are 0.3 higher (lower) than the basic letter grade. There are no A+, D-, or F+ course grades. Some schools use variants of this process, so there may be differences between the results tabulated herein and the results that would apply in those institutions.

Table 2 indicates the letter grades that would be assigned to 20 students completing five equally weighted assessments by the four different methods. The examples are selected in order to highlight the potential for disparity in course grades and are not based on actual course data.

TABLE 2

Illustrative Examples of Course Grades with 5 Equally Weighted Assignments

	Scores on Assignments 1 to 5					Course Grading Method			
Example	#1	#2	#3	#4	#5	Total Possible Points	Modified Total Possible Points	Weighted Average Letter Grade	Median
1	100	100	100	100	75	A	A	***A-***	A
2	100	100	100	100	65	A	A	***B+****	A
3	100	100	100	100	0	***B-****	A-	B	A
4	95	95	95	95	95	A	A	A	A
5	95	95	95	95	75	***A-***	***A-***	***A-***	A
6	95	95	95	75	65	***B****	***B****	***B****	A
7	95	95	95	75	75	***B****	***B****	***B****	A
8	95	85	75	65	55	C	C	C	C
9	95	95	75	75	75	B	B	B	***C****
10	95	75	75	75	75	C+	C+	C+	***C***
11	95	75	75	75	55	C	C	C	C
12	95	75	75	0	0	***F****	D+	C-	C
13	85	75	75	55	55	***D+****	***D+****	***D+****	C
14	85	75	65	65	0	***F****	D+	D+	D
15	75	75	75	65	55	***D+****	***D+****	***D+****	C
16	75	75	75	55	0	***F****	D	D	C
17	75	75	65	55	45	D	D	D	D
18	75	65	65	55	40	D	D	D	D
19	85	65	55	50	45	D	D	D	***F****
20	100	65	65	50	0	***F****	D	D	D

This table indicates the letter grades that would be assigned by four different methods, following the standards outlined in Table 1, for students with the indicated scores on five equally weighted assignments. Items in ***bold italics*** are the lowest grade given. Asterisks (*) indicate the gap between top and bottom grade is at least two steps in the grading scale.

In Table 2, where the different methods arrive at different course grades, ***bold italics*** indicate the lowest grade. An * is used to indicate situations where the top and bottom grades differ by at least two steps on the grading scale, such as the difference between an A and a B+, or from C+ to C-. While sometimes the four methods arrive at the same course grade, in other situations they arrive at grades that differ significantly. Indeed, in one case (Table 2, example 16) the difference is from F to C, a difference of 2.0 on a 4.0 scale. The table also shows that no method is theoretically always more generous, or always least generous.

Some of the largest differences occur when a student receives a very low score on an assignment, approaching a zero. Under Total Possible Points, the difference between the lowest D and the lowest F is 60 points, which is 60% of the range from 0 to 100. Under the Weighted Average Letter Grade method, the difference is only 1.0 on a 4.0 scale. Under the Median method, all F's are identical. See Students 3, 12, 14, 16 and 20 on Table 2.

While usually, in statistics, the median is considered robust to minor changes in data, in our examples the median can lack robustness when most data points fall at extremes. Compare the situations of students # 7 and # 9 in Table 2. Student #7 has three 95's and two 75's. The median is an A. Student # 9 has a difference in only one test: this student has two 95's and three 75's. Student # 9's median grade is C, a difference of 2.0 on a 4.0 scale. Under the Total Possible Points, Modified Total Possible Points and Weighted Average Letter Grade methods, the three grades would all be B's (87%, 87% and 83% or 3.2, 3.2 and 2.8 out of 4.0).

Modifications of the Grading Methods

Of the 20 illustrative examples presented in Table 2, there were 15 where the four methods gave different answers. In 12, the resulting letter grades varied by at least two grading steps. As discussed above, the disparities arose from several causes: the impact of zero grades on the Total Possible Points computations; the failure of the Weighted Average Letter Grade method to differentiate between high and low scores within letter grade categories; and the impact on the Median method of changes in the middle score when the other four were at extreme levels. When excluding the Total Possible Points Method, the number of disparities among the remaining three methods dropped modestly, from 15 to 14 and the number of examples where results differed by two letter grades decreased from 12 to ten.

In theory, two additional modifications to the three methods also may reduce the disparities. First, the Weighted Average Letter Grade system can be improved to capture more information by expanding the grading framework to include, for example, an A+ with a value of 4.3 for scores of 99 and 100 (or higher if extra credit is given) and an F+ for scores just below 60. Second, the Median method can be modified to use the average of the three middle scores in any odd numbered series of assignments, instead of the middle letter score of any odd number of assignments.

The modification to the Median method (not tabulated herein) significantly reduces the number of large disparities, changing the letter grades in eight out of the 20 examples. In six examples, the change was greater than two grading steps. The changes were most significant for students where assignments one and two were very good and assignments four and five were very bad. For student #12, the effect of using the modified rather than the unmodified median method would be to change the grade from C to F. There was only a modest impact (not tabulated herein) by expanding the letter grades used to score each assignment in the Weighted Average Letter Grade method to include A+ and F+. In the three cases where students had scores of 100 (examples #1, #2 and #3), the grades using the modified Weighted Average Letter Grade method became higher, and more in line with the Median and Total Possible Points methods.

Many professors follow the practice of dropping the lowest assignment score before aggregating scores into a course grade. If the lowest assignment scores were dropped the illustrative examples in Table 2, there would be 17 upward changes in grade under the Weighted Average Letter Grade method; 14 under the Modified Total Possible Points method; and 13 under the Total Possible Points method. The Median method would be the least affected, with only five upward changes in grade in the 20 cases, although two would be full letter grade changes. Dropping the lowest grade also reduces disparities among the three methods, with both fewer total disparities and fewer large ones. The total number of examples where the four grading methods would differ would fall from 15 to 10, and the number with differences of two grading steps would fall from 12 to 5.

DATA AND ANALYSIS

In order to evaluate whether the hypothetical discrepancies discussed in Sections 2 and 3 are actually common, the four grading methods were applied to student data across several different accounting classes at one university. Data were obtained from seven instructors, including the authors, at Hofstra University, a northeastern university with dual AACSB business and accounting accreditation. The data include assignment scores for students in 24 classes in Introductory Financial Accounting, Introductory Managerial Accounting, or Intermediate Accounting 1 from 2007 to 2011. The instructors provided the weights for each type of assessment and the student scores on a 100-point scale for each of the assessments for 1,062 students who completed the courses.

Four limitations should be noted. First, this study does not include data on the 60 students who withdrew from these classes. Students at this school are permitted to withdraw until relatively late in the semester and receive a grade of "W," which will not factor into their grade point average. These students are likely to have been performing poorly. Second, variations in instructor grading practices across classes likely makes comparisons of raw component scores not meaningful. Third, the data do not contain any additional judgmental factors faculty used in assigning the actual course grades. Finally, student performance was presumably motivated by the aggregation method they expected instructors to use. All instructors (19 classes) used Total Possible Points or Modified Total Possible Points, except for one, who used the Weighted Average Letter Grade method for five classes.

Table 3 lists the classes, number of students enrolled, and assignment weights. Introductory Financial Accounting and Introductory Management Accounting are required three-credit courses for all business majors. The students are mostly sophomore business majors, of whom a minority plans to major in accounting. Intermediate Accounting I is a three-credit class required of accounting majors, but not other business students, so it is taken primarily by juniors majoring in accounting. These courses and the types of assignments are similar to those given in many other business schools. Class sizes varied from 20 to 71.

While faculty in each course are required to cover a departmentally approved list of topics, and the same textbook is used in all sections, instructors have considerable freedom in deciding how to evaluate students. In Table 3, the weights instructors placed on midterm and final exams, combined, varied from 60% to 90%; the weights on homework varied from 5% to 25%; and the

TABLE 3

Class Sizes and Grading Weights

			Weights Placed on Various Items (%)				
	# Finished						
Section	(Dropped)		Final	Test 1	Test 2	Homework	Other
Introductory Financial Accounting							
1	62	(3)	30	20	20	10	20
2	65	(4)	25	25	25	25	0
3	52	(0)	30	25	25	20	0
4	60	(2)	30	25	25	20	0
5	41	(2)	30	30	30	5	5
6	20	(5)	30	30	30	5	5
7	62	(7)	30	30	30	10	0
8	32	(1)	30	30	30	10	0
9	31	(6)	30	25	25	20	0
10	65	(4)	25	25	25	25	0
Subtotal	490	(34)					
Introductory Managerial Accounting							
1	48	(0)	30	25	25	20	0
2	35	(1)	30	25	25	20	0
3	16	(1)	30	25	25	20	0
4	35	(0)	30	30	30	10	0
5	71	(1)	30	30	30	10	0
6	58	(7)	30	20	20	20	10
7	59	(0)	30	20	20	20	10
Subtotal	322	(10)					
Intermediate Accounting 1							
1	21	(4)	30	25	25	20	0
2	20	(0)	30	25	25	20	0
3	56	(2)	25	20	20	15	20
4	24	(2)	25	20	20	15	20
5	28	(3)	25	20	20	15	20
6	39	(2)	20	20	20	15	25
7	62	(3)	20	20	20	15	25
Subtotal	250	(16)					
Total	1,062	(60)					

* The "Other" items may include attendance, quizzes, projects, and an additional exam.

weights on other assignments from zero to 25%. Homework, while comprised of numerous assignments, is here treated as a single assignment. For courses with more than one other assignment, their combined weight is shown in Table 3. No class used fewer than four different significant component assignments.

For each student, course grades were computed using the four methods described earlier: Total Possible Points; Weighted Average Letter Grade; Modified Total Points; and Median. In addition to the baseline computations, grades were computed in some tables with a modification that ignores the lowest score, which is a practice that Brookhart (1999) indicates that instructors choose to do for a variety of reasons.

The overall medians and mean grades using the four baseline methods are quite similar. The median grades using all four methods were the same, at 2.3, and the mean grades only varied across a narrow range, from 2.25 (Median) to 2.30 (Total Possible Points) to 2.35 (Weighted Average Letter Grade and Modified Total Possible Points). Runs tests (cf. Mood and Graybill, 1963) of the distributions of the grades for the overall sample rejected the hypothesis that the distributions of the Median, Total Possible Points, or Weighted Average Letter Grade methods were different at 1% significance levels. These three methods also have means and medians close to each other when the lowest grade is dropped, although student grades move markedly upward in all three methods.[7]

The first, baseline, part of Table 4, tabulates the frequency of the grades awarded by each method. It indicates that the mode of the Total Possible Points distribution is in the B- to B+ range, while the modes of the other three distributions are in the C- to C+ range. The number of A's awarded for the full sample was fairly similar across the methods, ranging from 180 for Weighted Average Letter Grade to 198 for Median. The number of F's varied more widely. The Weighted Average Letter Grade method only awarded 46 F's, Modified Total Possible Points awarded 59, Total Possible Points awarded 87, and Median method awarded 105.

The major difference between the Modified and regular Total Points Methods involves the smaller number of F's, and the greater numbers of C's and D's, in the Modified method. Much of this improvement is likely due to homework, since that was the area where very low scores were most common. In the classes in this sample, students rarely missed tests, and most tests were multiple choice tests, giving students who simply guessed an expectation of a 25% minimum score. In our sample, there were only a total of 5 scores below 25% out of the 3,186 scores on final exam and the first two midterms. There were 45 homework scores below 25%, representing 4.25% of the students.

When grades were recomputed, assuming the lowest score was dropped, all three main methods showed more A's and fewer F's. From 74 to 100 more students got A's, depending on the aggregation method. The number of F's fell from 105 to 48 for Median, from 46 to 29 for Weighted Average Letter Grade and from 87 to 53 for Total Possible Points.

Table 5 presents data on differences and similarities among three grading methods for the 1,062 students.[8] The first section indicates that, overall, the three baseline methods agree on grades

[7] Such an increase in grades is consistent with Simonite's (2000) findings on honors degrees, when only the best scores were considered.

[8] The Modified Total Points method is considered in the last column, as a modification to the baseline results.

TABLE 4

Frequency of Grades Awarded - Various Methods

Course	**A-, A**	**B-, B, B+**	**C-, C, C+**	**D, D+**	**F**
	Baseline Results - using actual grading methods (n = 1,062)				
Weighted Average Letter Grade	180	308	337	191	46
Total Possible Points	194	301	294	186	87
Median	198	278	291	190	105
Modified Total Possible Points	195	302	311	195	59
	Sensitivity Analysis - dropping the lowest score (n = 1,062)				
Weighted Average Letter Grade	254	391	262	126	29
Total Possible Points	288	366	244	111	53
Median	298	334	244	138	48

This table provides the grade distributions that result from mechanical application of various methods of combining exam and other scores described in the text.

for only 431 students (41%), but differ for the other 59%. If the lowest score is dropped, the three methods agree (disagree) 44% (56%) of the time, and if the lowest component score is limited to 50 out of 100, they would agree (disagree) 46% (54%) of the time. Thus in this population, both baseline and modified methods fail to agree for a majority of students.

If the comparison is restricted to the Total Possible Points and Weighted Average Letter Grade methods, they always disagree for at least a quarter of the students. In the baseline case they agree (disagree) for 68% (32%) of the grades. If the lowest score is dropped, the percentage of agreements (disagreements) changes to 70% (30%), and if the lowest score on any assignment is changed from zero to 50, the number of agreements (disagreements) is 74% (26%).

The second and third sections of Table 5 indicate that each of the three basic methods sometimes gave higher and lower grades than the other two methods. The Weighted Average Letter Grade method was least likely to be either highest or lowest, as it only gave the highest grades in 8% of the cases and the lowest grades in 5%. The Median method was most likely to differ from the others. It gave a higher grade 15% of the time, and a lower grade 23% of the time. The Median method remained the most likely to give highest and lowest grades when the methods were modified by dropping the lowest score and by changing zero scores to 50's. The high variability of this method is consistent with the findings of Simonite (2000). The Total Possible Points method gave the top

TABLE 5

Differences in Scoring Individual Students - Full Sample Results

	Baseline methods		Assume lowest score is dropped		Assume no item score < 50 out of 100	
Number and % of students for whom methods agree						
All three methods	431	41%	468	44%	450	46%
Weighted Average Letter Grade and Total Possible Points	726	68%	745	70%	789	74%
Total Possible Points and Median	606	57%	567	53%	632	60%
Weighted Average Letter Grade and Median	547	52%	585	55%	547	52%
Number and % of students for whom each method gives the highest grade						
Weighted Average Letter Grade	88	8%	57	5%	60	6%
Total Possible Points	107	10%	71	7%	107	10%
Median	156	15%	165	16%	156	15%
Number and % of Students for whom each method gives the lowest grade						
Weighted Average Letter Grade	51	5%	71	7%	51	5%
Total Possible Points	94	9%	82	8%	57	5%
Median	247	23%	205	19%	274	26%
Number and % of students for whome grading differences equal at least two grading steps						
Weighted Average Letter Grade and Total Possible Points	88	8%	49	5%	31	3%
Total Possible Points and Median	243	23%	192	18%	236	22%
Weighted Average Letter Grade and Median	215	20%	156	15%	215	20%
Largest differences noted between methods (out of 4.0)						
Weighted Average Letter Grade and Total Possible Points	1.0 (52 cases)		1.0 (52 cases)		1.0 (17 cases)	
Total Possible Points and Median	2.0 1.7	(1 case) (4 cases)	2.0 1.7	(1 case) (4 cases)	2.0 1.7	(1 case) (4 cases)
Weighted Average Letter Grade and Median	2.0 1.7	(1 case) (2 cases)	2.0 1.7	(1 case) (2 cases)	2.0 1.7	(1 case) (2 cases)

This table provides data regarding the differences in course grades resulting from mechanical application of the various methods of combining exam and other scores described in Table 3.

score 10%, and the bottom score 9%, of the times for the baseline methods. When scores of zero were treated as 50's, the number of times it gave the lowest grade changed from 9% to 5%.

The final two sections of Table 5 indicate the magnitude of the differences among methods. The fourth section shows the number and percentage of students for whom the grading differences were at least two grading steps, which we considered a significant difference. The Median method has numerous significant disagreements with the other two methods. Using the baseline computations, the Median method disagrees with the Total Possible Points and Weighted Average Letter Grade methods by at least two steps scale 23% and 20% of the time, respectively. The largest single difference, shown in the final section of the table, was 2.0. This student's assignment scores (weights) were, from smallest to largest score, 58 (30%), 58 (25%), 81 (25%), and 105 (20%). The weighted average of the scores was 73, for a C under the Total Possible Points method, but the median was a 58, or an F. Since the two low scores had a total weight of 55%, the two higher scores did not count.

There were fewer significant differences between the Weighted Average Letter Grade and Total Possible Points methods. While these two methods disagreed 32% of the time, differences of two or more steps only occurred in 8% of the cases. The final section of Table 5 indicates that the largest difference between those two methods was a difference of 1.0, a full letter grade, which occurred 52 times. Inspection of the data and a comparison to Table 4 indicates that most of these differences are between D and F. This likely occurs because the grading scale used has neither a D- nor an F+. If scores of zero were treated as 50's, the percentage of significant disagreements would fall from 8% to 3%, and the number of differences of 1.0 would fall from 52 to 17.

The final area explored in this study is the different motivational effect of the three baseline methods. Table 6 reports how many students could achieve each major grade level, based on their performance on all assignments except the final exam, with varying final exam scores. Thus, if all students received 95's on their final exams, the Weighted Average Letter Grade method would compute 291 A's, Total Possible Points would compute 314, and Median would compute 416. This compares to the figures of 180, 194, and 198 shown in Table 4, computed using actual final exam scores. (The mean actual final exam score was 72.) The implication of the large numbers of A's available to students who could score a 95 on the final exam is that the Median gives students a powerful incentive for a late study push. However, those students who had A's on early assignments may have no motivation to score well on the final, since the final section of Table 6 indicates that 160 students could receive A's under the Median method even if they score zero on the final. Table 6 also indicates that under the median method, the number of students who would receive A's is relatively constant for all the grades shown below a 95.

The Weighted Average Letter Grade method holds out more hope to poor students of achieving a minimal passing grade than do the other two methods. Under the Weighted Average Letter Grade method, if students score either 85 or 95 on the final, all could pass. In contrast, under the Total Possible Points method, 14 of the students who received 95 on the final would fail, and 31 students would fail under the Median method. There is no incentive for these students to make an effort at the end of the course. These students should rationally withdraw from the course before the final or, if school policy forbids such withdrawals, they should concentrate all their energies on their other courses. The differential effect may be more than reported herein, since the data analyzed exclude those students who withdrew before taking the final exams.

TABLE 6

Impact of Different Final Exam Scores on Grades Awarded, Using Three Different Methods (n = 1,062)

	A’s	B’s	C’s	D’s	F’s
If students score 95 on final exam					
Weighted Average Letter Grade	291	478	267	26	0
Total Possible Points	314	441	231	62	14
Median	416	323	210	82	31
If students score 85 on final exam					
Weighted Average Letter Grade	162	488	338	74	0
Total Possible Points	196	455	282	108	21
Median	165	543	238	85	31
If students score 75 on final exam					
Weighted Average Letter Grade	42	476	394	145	5
Total Possible Points	92	431	344	154	41
Median	160	261	506	104	31
If students score 65 on final exam					
Weighted Average Letter Grade	0	382	440	216	24
Total Possible Points	32	371	393	206	60
Median	160	245	289	332	36
If students score zero on final exam					
Weighted Average Letter Grade	0	244	470	279	69
Total Possible Points	0	0	94	387	581
Median	160	245	269	184	204

This table shows results using the baseline grading methods of different final exam grades. The mean and median scores students actually received were 72.3 and 74, respectively.

DISCUSSION AND RECOMMENDATIONS

The illustrations and data analysis suggest several factors for instructors to consider. The first is that our data indicate that the choice of aggregation methods is highly likely to affect individual students' grades. Only 41% of the students would have received the same course grade under all three basic methods. These tend to be students who perform consistently, at any given level of performance, on each assessment. None of the three methods compared was systematically more advantageous or disadvantageous from the students' perspective. The largest difference was for a student who earned a C under the Weighted Average Letter Grade and Total Possible Points methods, but an F under the Median method. While only one difference of that size occurred in the sample of 1,062, differences between the Weighted Average Letter Grade and Total Possible Points methods of at least two grading steps arose in 8% of the cases, and such differences between the Median method and, respectively, Total Possible Points and Weighted Average Letter Grade occurred 23% and 20% of the time. While an 8% frequency of differences of two grading steps may not seem large, it implies 3 students in a class of 40 whose grades would differ by two or more steps based solely on whether the instructor uses Total Possible Points or Weighted Average to aggregate scores.

Students' incentives differ across methods. The Median method, because it essentially ignores the most information, results in the most cases where a good student could afford to stop working after achieving A's in the early assignments, or where early failures make effort at the end of the course pointless, but there are also the most cases where a late push can be highly productive and make early poor grades irrelevant.

When comparing the other methods, the Total Possible Points method penalizes unsubmitted work and very low test grades more severely than does the Weighted Average Letter Grade method or, of course, the Modified Total Points method. More students would mathematically achieve an F, regardless of final exam score, under the Total Possible Points than the Weighted Average Letter Grade method. To put it another way, of the three basic methods, the Weighted Average Letter Grade method gives struggling students the most hope of eking out a passing grade with a (possibly miraculous) A or B on the final exam. Under the other methods, it would be rational for more students to drop the course early. The Weighted Average Letter Grade method is also the one least likely to reward a high final exam score with an A for the course.

Consistent with the theoretical literature, the data point to extreme inconsistency in student performance as the major cause of the larger disparities found among aggregation methods. As expected, a smaller number of assessments is associated with less reliable aggregations. The classes with the most disparities in this study placed very high weight (80% to 90%) on three exams, while those with the least disparities placed only 60% to 65% on exams and from 20% to 25% on projects. The instructor can design assignments with low levels of dispersion, and avoid ones with high variation in the data. Projects had low dispersion, whereas homework grades were widely dispersed, especially when extra credit made the top achievable grade over 100 points, and some students chose to do no homework. If homework is to have a strong weight in the course grading scheme, variation can be minimized by pushing students to submit assignments. Guskey (2000) suggests assigning a grade of Incomplete until all missing work is completed, ensuring that the grade is a more accurate reflection of what the student has learned. Mathematical ways to minimize the variation of homework grades include adopting the Weighted Average Letter Grade method or adopting Reeves' suggestion of setting 50, not zero, as the minimum score for any assignment.

The Median method, in the data, had the most differences with other methods, and the greatest number of both the highest and the lowest scores. Instructors interested in using this method should probably use as many assignments as possible, and should be careful in weighting assignments to avoid situations where minor differences in performance on one assignment can cause large shifts in grades.

Modifying the grading schemes by dropping the lowest score, or by assuming no item score is less than 50 points, are both ways of damping the impact of inconsistent performance. They had some modest effects in reducing the number of disparities between methods. If such modifications fit the instructors' goals and teaching philosophy, they may be worthwhile.

Grade aggregation techniques should be properly explained to students at the commencement of the course, with illustrations.[9] Involving students in the assessment process enriches their learning and creates a partnership between the professor and the student. As Elikai and Schuhmann (2010) demonstrate, if students understand a grading policy, however strict, to be attainable, they will be motivated to succeed. If the grading procedures are not explained, students may assume a different method of combining grades is being applied, and the student is likely to form unrealistic grade expectations, leading to awkward semester-end discussions.

Our study has various limitations. The data all came from accounting classes at a single university. Replication in other colleges and at other educational levels is needed to test the robustness of our findings. A second limitation of this study is that it does not look at the impact between aggregation methods and students' performance on individual assessments. The comparison of the impacts of the aggregation methods on final grades treated the student performance on individual assessments as exogenous, and simply noted what the differences in final grades would be between methods, given those individual assessments. No attempt was made to study how student effort levels would change depending on the aggregation method, although the analysis of Table 6 data indicates that student motivation should vary. Further research on the motivational impact of different aggregation methods is needed.

This study tested certain aggregation methods suggested by prior literature. However, we saw no empirical data on how frequently these aggregation methods are in fact used, or whether professors vary their aggregation methods based on the instructional requirements of the course.[10] When we presented our findings at a faculty seminar in our university, it was clear from the attendees' reactions that many faculty members are unaware of the various possible methods of aggregation, and therefore may not optimally match the aggregation technique to the learning goals of the course. For example, a professor teaching a language course might appropriately decide that

[9] As a result of this research one of the coauthors significantly increased the discussion of grading policies in his course outlines.

[10] While there appear to be no published studies on this topic, Reeves (2008) implies that a variety of methods are in use. Although there are no tables presented, he indicates that more than 10,000 faculty members were surveyed and asked to "calculate the final grade for a student whose 10 assignments during the semester had received the following marks: C, C, MA (missing assignment), D, C, B, MA, MA, B, A." Reeves (2008) states that, "The results include final grades that include F, D, C, B and A."

the full grade should be based on the knowledge shown at the end of the course, while a professor teaching an advanced accounting course with distinct topics like consolidations and governmental accounting might decide the grade needed to consider individual assessments of each important topic area. Further research could help measure faculty awareness of aggregation methods, the relative frequency with which they are used, and whether faculty choice of aggregation methods is correlated with factors suggested by the educational literature.

This study also did not address whether the aggregation methods used were understood and seen as fair by students. Studies of student understanding of, and attitudes towards, different methods of summarizing grades could help faculty design and communicate their aggregation methods.

There is no perfect theoretical answer to the issue of how to aggregate scores into grades. The data presented in this paper should sensitize instructors that the different grading policies can have a direct impact on both students' motivation and grades.

REFERENCES

Adams, R. M., and J. Wilmot. 1981. A Measure of the Weights of Examination Components, and Scaling to Adjust Them. *Journal of the Royal Statistical Society. Series D (The Statistician)*. (Vol. 30, No. 4) 263-269.

Biggins, J. D., R. M. Loynes, and A. N. Walker. 1986. Combining Examination Marks. *British Journal of Mathematical and Statistical Psychology* (Vol. 39) 150-167.

Brookhart, S. M. 1999. The Art and Science of Classroom Assessment: The Missing Part of Pedagogy. *ASHE-ERIC Higher Education Report* (Vol. 27, No. 1) (Washington, DC: The George Washington University, Graduate School of Education and Human Development.

Brumfeld, C. *Results of the 2004 AACRAO Survey*. (Washington, DC: American Association of Collegiate Registrars and Admissions Officers.

Butler, C. 2004. Are Students Getting a Free Ride? *New York Teacher* (June 2).

Cresswell, M. J. 1988. Combining Grades From Different Assessments: How Reliable is the Result? *Educational Review* (Vol. 40, No. 3) 361-382.

Dalziel, J. 1998. Using Marks to Assess Student Performance: Some Problems and Alternatives. *Assessment & Evaluation in Higher Education* (Vol. 23, No.4) 351-366.

Elikai, F., and P. W. Schuhmann. 2010. An Examination of the Impact of Grading Policies on Students' Achievement. *Issues in Accounting Education* (Vol. 25, No. 4) 677-693.

Ekstrom, R. B., and A. M. Villegas. 1994. *College Grades: An Exploratory Study of Policies and Practices.* (New York: College Entrance Examination Board).

French, S. 1981. Measurement Theory and Examinations. *British Journal of Mathematical and Statistical Psychology* (Vol. 34) 38-49.

_______. 1985. The Weighting of Examination Components. *Journal of the Royal Statistical Society. Series D (The Statistician)* (Vol. 34, No. 3) 265-280.

Frith, D. S., and H. G. Macintosh. 1984. *A Teacher's Guide to Assessment*. (Cheltenham, United Kingdom: Stanley Thornes Ltd).

Guskey, T. R. 2000. Grading Policies that Work Against Standards...and How to Fix Them. *National Association of Secondary School Principals. NASSP Bulletin* (Vol. 84, No. 620) 20-27.

Hölder, O. 1901. Die Axiome der Quantität und die Lehre vom Mass. *Berichte uber die Verhandlungen der Koeniglich Sachsischen Gesellschaft der Wissenschaften zu Leipzig, Mathematisch-Physikaliche Klasse* (Vol. 53) 1-46.

Looney, M. A. 2003. Facilitate Learning with a Definitional Grading System. *Measurement in Physical Education and Exercise Science* (Vol. 7, No. 4) 269-275.

McLachlan, J. C., and S. C. Whiten. 2000. Marks, Scores and Grades: Scaling and Aggregating Student Assessment Outcomes. *Medical Education* (Vol. 34) 788-797.

Miller, A. H., B. W. Imrie, and K. Cox. 1998. *Student Assessment in Higher Education*. (London: Kogan Page Limited).

Mood, A. M., and F. A. Graybill. 1963. *Introduction to the Theory of Statistics*. (New York: McGraw-Hill Book Company).

Reeves, D. B. 2004. The Case Against the Zero. *Phi Delta Kappan*. (Vol. 86).

_______. 2008. Leading to Change/Effective Student Grading Practices. *Educational Leadership* (Vol. 65, No. 5) 85-87.

_______. 2009. Remaking the Grade, From A to D. *The Chronicle of Higher Education* (Sept. 18, A64.

Rowntree, D. 1987. *Assessing Students – How Shall We Know Them?* (London: Harper and Row).

Simonite, V. 2000. The Effect of Aggregation Method and Variations in the Performance of Individual Students on Degree Classifications in Modular Degree Courses. *Studies in Higher Education* (Vol. 25, No. 2) 197:209.

Spence, R. B. 1927. The Improvement of College Marking Systems. *Teachers College, Columbia University Contributions to Education* (No. 252).

Walker, K. 2006. Research Brief: Role of Zero in Grading. The Principals' Partnership: A Program of Union Pacific Foundation. October 9, 2006. http://kin.psdschools.org/webfm_send/473.

Waters, M. 1979. Grading in a More Complex Learning Environment. *The Balance Sheet* (Vol. 61, No. 2) 74-77.

Wiliam, D. 1995. Combination, Aggregation and Reconciliation: Evidential and Consequential Bases. *Assessment in Education: Principles, Policy & Practice* (Vol. 2, No.1) 53-73.

Wilson, I. 2008. Combining Assessment Scores - A Variable Feast. *Medical Teacher*. (Vol. 30) 428-430.

Woolf, H., and D. Turner. 1997. Honours Classifications: The Need for Transparency. *New Academic.* (Autumn) 3.

Yorke, M. 2008. *Grading Student Achievement in Higher Education: Signals and Shortcomings*. (Obindgon, Oxon: Routledge).

_______, P. Bridges, and H. Woolf. 2000. Mark Distributions and Marking Practices in UK Higher Education: Some Challenging Issues. *Active Learning in Higher Education* (Vol. 1, No. 1) 7-27.