ISSUES IN ACCOUNTING EDUCATION
Vol. 25, No. 4
2010
pp. 677–693

American Accounting Association
DOI: 10.2308/iace.2010.25.4.677

An Examination of the Impact of Grading Policies on Students' Achievement

Fara Elikai and Peter W. Schuhmann

ABSTRACT: The strategy of evaluating students' achievement using a marking system is a common practice in higher education institutions. The result of a student's effort is usually communicated in the form of a letter grade or percentage correct on an exam or on the course as a whole. Although a vast majority of instructors use various grading policies and the impact of different grading policies on learning is a basis of considerable debate among academics, the empirical work regarding the impact of different grading policies on students' performance does not include applications to accounting, a discipline for which student learning is directly tied to success in passing professional examinations. Theoretically, one of the functions of a grading system is to motivate students to work harder and perform better. This study provides insight into the impact of a lenient grading scale versus a strict grading scale on students' achievement, where the level of "average" mastery in the latter category (the grade of C), is coincident with the minimum passing requirement of the professional accounting examinations. The results of this study support the notion that an attainable strict grading policy can be used as an important pedagogical technique to motivate students to study and may provide insight into grade scale decisions faced by accounting faculty seeking to prepare their students for the rigor of professional exams. Contrary to prior results in the literature, we find that when used in an upper-level undergraduate accounting course the stricter standard has a more profound effect on achievement for students at the lower end of the grade distribution.

INTRODUCTION

Evaluating students' achievement using a marking system is common practice in higher education institutions. The result of a student's effort is usually communicated in the form of a letter grade or percentage correct on an exam or on the course as a whole. Although a vast majority of instructors use various grading policies and the impact of different grading policies on learning is a basis of considerable debate among academics, the empirical work regarding the impact of different grading policies on students' performance does not include applications to accounting, a discipline for which student learning is directly tied to career achievement through the completion of professional examinations.

Recently, many universities have changed from traditional letter grades to a plus/minus grading system which adds or subtracts three-tenths of a grade point. This system is stricter than the

Fara Elikai is an Associate Professor and Peter W. Schuhmann is a Professor, both at The University of North Carolina at Wilmington.

Published Online: November 2010

traditional system for the lower scores within each grade category; and under this grading system students must perform better in classes to get the same letter grade. Furthermore, professional organizations in accounting have set the passing score at 75 percent or higher for professional examinations (see for example, http://www.nasba.org for Certified Public Accountants examination and http://www.theiia.org for Certified Internal Auditor). This implies that, in order to pass these professional examinations, a student must achieve at a level higher than C and D—levels that are typically acceptable as passing scores under the traditional grading system at the university level.

Predominantly, a grading system is based on the notion of an absolute index of achievement on a scale from 0 to 100. The percentage scale is what the instructors consider when they adopt a system such as A = 90–100, B = 80–89, C = 70–79, D = 60–69, and F = 59 or below (Milton et al. 1986). A survey of over 1,600 institutions from the American Association of College Registrars and Admissions Officers (AACRAO) indicated that about 92 percent of all institutions use letter grades and that the GPA used in about 95 percent of all institutions is a 4.00 point scale with A = 4.00 and F = 0.00 (Quann 1984). Traditionally, the categories of percent scores above 60 or 70 are considered as the levels of passing performance. This is due to a prevalent belief that students should demonstrate some mastery of the subject matter in order to be judged acceptable (Terwilliger 1971).

The purpose of this study was to examine whether the motivational aspect of a grading system on students' performance advanced in education and psychology literature (e.g., Adams and Torgerson 1964; Norman 1981; Walvoord and Anderson 1998) applies to students in university level accounting courses. We adopted an approach that is similar to previous studies in the literature, comparing the effects of a lenient grading scale and a strict grading scale on students' performance in a cost accounting course. This study extends prior research by providing current evidence for the important issue of whether grading scales can have a positive impact on students' achievement. It is the first study of this kind that applies the highest attainable grading scale to university level courses and the first to measure university-level grading scale effects across students' GPA categories. Importantly, this study was conducted within the accounting discipline; prior studies indicate that there are significant differences between accounting (and business) majors and nonbusiness majors regarding several key aspects such as attitude, self-discipline, achievement, independence, goal orientation, and other value types (Giacomino and Akers 1998; Ridener 1999). Numerous studies point to significant differences in personality types and learning styles between accounting students and students in other disciplines as measured by the Myers-Briggs Type Indicator (e.g., Booth and Winzar 1993; Wolk and Nikolai 1997; Briggs et al. 2007). Accounting students' tendency toward extrinsic motivation for learning (availability of jobs and/or good salaries) is evidenced repeatedly in the literature as is the related finding that a vast majority of undergraduate accounting students (up to 86 percent) plan additional education beyond the baccalaureate (Nelson and Deines 1995; Nelson and Vendrzyk 1996; Nelson et al. 2002). In addition, the accounting context is underrepresented in most educational journals. Notably, this void was recently acknowledged in the editorial essay in *Issues in Accounting Education* (St. Pierre et al. 2009).

> The importance of disciplinary differences remains largely unexplored (Neumann 2001). Scholars exploring the importance of differences across academic areas have found that, because knowledge within fields is defined differently, basic epistemological assumptions implicit within teaching practices differ. Educators structure their time and their teaching differently, and strive for different outcomes. The impact of disciplinary differences on students' learning has received insufficient attention. But even within business schools, one can easily see that Strategy and Operations Management courses are taught differently, and students must approach their studying and structure their efforts accordingly. While the literature on higher education offers important insights, key

> unanswered questions that arise within disciplines must be studied separately. A similar argument can be applied to justify research specifically aimed at the pedagogy of music or graphic art on the one hand (which lead on to professional practice in those fields), or to mathematics or philosophy on the other (which are less directly linked to specific fields of professional practice).

In addition, the lenient grading scale (or the traditional grading scale) differentiates letter grades based on the following percentage points: A = 90–100, B = 80–89, C = 70–79, D = 60–69, and F < 60 percent; the strict grading scale separates letter grades as: A = 93–100, B = 85–92, C = 75–84, D = 65–74, and F < 65 percent. Under the latter category, the level of "average" mastery (the grade of C) is coincident with the minimum passing requirement of the professional accounting examinations. When combined with the theory previously advanced in the literature, results of this study have important implications for grade scale decisions faced by accounting faculty seeking to prepare their students for the rigor of such exams.

In general, there are two central theoretical arguments related to the motivational aspect of grading policies. One argument is that a strict grading scale motivates students to put forth more effort to achieve the instructional objectives, which have been specified (e.g., Adams and Torgerson 1964; Norman 1981; Johnson and Beck 1988; Walvoord and Anderson 1998). The premise of this argument is that, like any standard setting in performance evaluation, the grading scale can function as a motivational tool to encourage students to exert greater effort toward their class performance as long as it is attainable. Consequently, this marginal effort may lead to a better understanding of the subject matter and students may achieve an even better grade under the stricter scale than they would have otherwise. Like a plus/minus grading system, a grading scale where the minimum required to earn a given grade is raised, may have a positive motivational effect on student performance. As noted in McClure and Spector (2004), smaller differences between any two given grades gives students a greater possibility of being able to improve their grade during the course of the semester. Validation of these arguments would support the notion that setting higher standards in accounting courses will increase student success in the classroom and subsequently lead to a higher probability of achievement on professional exams.

However, another argument is that the additional pressure induced by a stricter grading scale can be detrimental to students' performance. The pressure can force a disproportionate amount of preparation and resulting intellectual and emotional fatigue (e.g., Butler and Nissan 1986; Squires 1999). Additionally, one could argue that students view different grading standards as simple issues of scaling, hypothesizing that a given level of effort will result in the same course grade under different grading regimes. For example, students may believe that strong effort will result in an "A" grade regardless of how an "A" is defined. Compounding the difficulty inherent in an examination of the issue is the notion of reverse causation. In addition to standards affecting performance via the student motivation reasons noted above, student performance may also affect an instructor's choice of standards.

Of course, these explanations need not be mutually exclusive in a particular classroom. Indeed, one could argue that higher grading standards may motivate some students to work harder to achieve high grades and at the same time serve as a disincentive to other students who may now view certain grades as out of reach. As noted in Betts (1997), weaker students may be more apt to "give up" when faced with stricter standards. For other students, higher standards may have no impact on motivation at all. Thus, consistent with utility-maximization models put forth by Becker and Rosen (1992) and Betts (1997), and with Crooks' (1988) summary of 14 fields of research regarding learning strategies, motivation, and achievement, the effect of grading standard on performance is theoretically and often empirically ambiguous.

Despite this uncertainty, there is little doubt that incentives serve to motivate behavior. Basic economic behavior shows us that when marginal costs are high or marginal benefits are low, there is a lower incentive to perform an action. Similarly, higher benefits or lower costs at the margin

serve to motivate action. As noted above, this simple notion becomes quite complex when applied to student behavior in the classroom. While we can make the assumption that most students seek to earn passing grades, they respond to incentives differently, have different learning styles, and more generally they may have vastly disparate reasons for completing a particular course; for example, Gardner and Lambert's (1972) division between intrinsic and extrinsic motivational factors. The effect of higher grading standards on students in accounting (or other disciplines that require professional examinations) warrants particular attention, as there is little room for failure—these majors must attain a baseline level of understanding in order to succeed in their profession. Unlike other college courses, the prospect of "giving up" in the accounting classroom may be tantamount to abandoning their chosen field of study.

LITERATURE REVIEW

Adams and Torgerson (1964) have identified four functions of grades assigned by teachers: (1) administrative, (2) informational, (3) guidance, and (4) motivational. The first of these, the administrative function, suggests sorting procedures in which decisions are made about individual students, within the school as well as outside of the school setting, for selection for honors, graduation, and employment of applicants, etc. The informational aspect means transmitting information to the students concerning their progress toward certain educational goals. The guidance function refers to the use of grades to identify areas of strength and weakness so that students can plan their study agendas and their educational and vocational future. The motivational function suggests that grades are considered incentives for greater effort. Norman (1981) also suggested that students' performance and skill can be enhanced if the students are motivated to study. Employing appropriate pedagogical techniques, such as a grading system, can encourage students to do their assignments, thereby increasing their understanding of a subject matter and improving their class performance, as well as allowing the instructor to obtain feedback regarding the degree of understanding and performance of students. For instance, exams can be used as a feedback mechanism and can encourage students to put forth more effort to study if the score on exams is a significant part of the course grade. Moreover, Walvoord and Anderson (1998) suggested that grading can evaluate the quality of students' work as well as motivate and encourage them to study and become involved in the course.

Empirical research in education also supports the notion that incentives can have a significant effect on students' performance. In an early study, Clark (1969) found that graduate students who competed for grades attained higher test scores than students who did not. The result of his study indicated that grades can motivate students to learn. Studies of grades versus pass-fail college courses by Gold et al. (1971) and Hales et al. (1971) found substantially higher achievement in classes that assigned grades. Austin (1978) found that homework that was assigned and checked contributed more to students' achievement than homework that was assigned, but not checked. Cherry and Ellis (2005) found that a rank-order grading system (norm-reference grading) can have a positive and significant effect on student performance relative to traditional criterion-reference grading. Hence, stricter grading standards in accounting courses may serve the dual purpose of raising the level of student performance and learning while at the same time preparing them for the rigor of the professional exams.

This basic hypothesis is supported by extensive literature on college grading systems in the fields of education and psychology (for reviews see Duke, 1983; Geisinger, 1982; Milton et al. 1986), where empirical research on the impact of various grading standards on students' performance spans a variety of disciplines and levels of education. Goldberg (1965) compared undergraduate students' score on a midterm exam in a psychology course by applying five different grading systems: a strict scale, a lenient scale, a scale based on bimodal, normal, or rectangular

distributions. His analysis indicated some evidence that students who were evaluated based on the strict scale earned higher scores on exams than students graded with the lenient scale. He concluded that grading policies have a marginal effect on test performance.

Paschal et al. (1984) analyzed past research on the effects of homework on student motivation and achievement. Like Austin (1978), they concluded that graded homework has a more significant effect than homework that is assigned but not graded. Miller and Westmoreland (1998), however, find no significant differences in performance based on whether all homework assignments are graded or selectively graded.

Betts (1997) develops an education production model in which utility-maximizing students choose their effort level on schoolwork. A theoretical prediction of this model is that more stringent grading standards will increase effort by most students, thereby increasing their achievement over time. To empirically test this hypothesis, the author used data from a panel study of students in the 7th and 10th grades associated with the Longitudinal Study of American Youth. Results of this study suggest that the school's grading standards are more important determinants of student progress than other factors such as class size and teachers' level of education and experience, and that the effect of higher grading standards on performance in science classes is roughly twice as large as the effect in math classes. Betts also finds that grading standards have a more pronounced effect on stronger students than on weaker students, supporting the theoretical argument that weaker students may "give-up" when faced with stricter standards.

Betts and Grogger (2003) used estimated grading standards across 1,000 high schools to test the effect of grading standards on student performance on a standardized 12th grade math test. Their results also suggest that higher grading standards have a greater impact on students in the upper end of the grade distribution. Students with lower grades do respond favorably to the incentives provided by higher grading standards, but to a lesser degree than their higher achieving counterparts. Interestingly, they found no effect of higher standards on high school graduation rates or college attendance. Figlio and Lucas (2004) used data on third, fourth and fifth grade standardized test scores to examine the effects of changing grading standards on student performance over time. The authors found that higher standards benefit student performance, and that these effects are most pronounced for high-ability students. Interestingly, Figlio and Lucas also concluded that low-ability students benefit most from high standards when they are in classes with high-ability peers, and high-ability students benefit most from higher standards in classes with low-ability peers.

Johnson and Beck (1988) examined the relationship between strict and lenient grading scales, and students' performance on tests administered in an undergraduate Educational Psychology class. The subjects were 91 undergraduate students who enrolled in 11 sections of an Education Psychology course over a three-year period. They used a strict scale in six sections and the lenient scale in five. Eight percentage points separated letter grades on the strict scale (e.g., A = 93–100; B = 85–92), whereas 12 percentage points separated letter grades on the lenient scale (e.g., A = 89–100; B = 77–88). Their findings indicated that students graded with the strict scale obtained higher test scores than students graded with the lenient scale.

Dixon (2004) examined the effect of a plus/minus grading scale versus traditional straight letter grading scale in six sections of an Introduction to Programming course, which is a required course for a major or minor in the field of Computer Science. However, the vast majority of students in this course consisted of students who declared Computer Science as a minor and typically only wanted exposure to programming. In this experiment, the subjects were 224 students who had an option to choose a straight letter grade scale or plus/minus grade scale before any midterm exams were given. More than two-thirds of the students chose the straight letter option compared to the plus/minus option. The results of study indicated that students who chose the plus/minus scale obtained more pluses than minuses, and the significant difference in grades

between the two groups was in the number of Bs and Ds obtained within each group. In a similar study, McClure and Spector (2004) examined the effect of implementing a plus/minus grading scale on student motivation, and found no statistically significant difference in total grades earned during a semester between groups of students who chose to be graded under a plus/minus system and those graded using a traditional letter grading scale.

In summary, there is strong empirical support for the notion that higher grading standards will enhance student motivation and learning outcomes. Most studies find that the effects are more pronounced for high ability students. That the effects of standards are found to be dependent upon student ability levels and field of study suggests that results for one group of students may not generalize to others. To date, no studies have examined the relationship between standards and achievement in a tertiary level field of study such as accounting where a minimum level of accomplishment is required for continued success in the profession.

RESEARCH QUESTIONS AND METHODOLOGY

Hypotheses

In light of the complex nature of the relationship between grading scale and student performance, we attempt to address the following null hypotheses regarding whether changing the grading scale has any positive or negative impact on accounting students' performance:

H1a: There are no significant differences in accounting students' achievement, as measured by students' performance on exams, when using two different grading scales.

H1b: There are no significant differences in the effect of grading scales on accounting students' achievement across students of different ability levels.

Subjects

The experiment was conducted over two 16-week semesters during the 2004 academic year in a regional university with a population over 12,000 students. Two sections of undergraduate cost accounting were examined in each semester, with the students in the first semester serving as the control group, and the students in the second semester as the treatment group. All students in both sections were pursuing accounting majors. The course is required to complete the B.S. in Business Administration with a concentration in accounting. The class sizes and gender make-up in each group were comparable, and all sections met in the morning. In total, there were 91 students in the control group and 95 students in the treatment group. Students who were repeating the course were excluded from the experiment. In addition, students who dropped the course during the add/drop period (the first week of each semester) without receiving the letter grade of "W" were excluded.

As in previous studies (e.g., Baldwin, 1980; Vruwink and Otto, 1987; Oglesbee et al. 1988), the cumulative grade point average (GPA) of the students prior to the semester of the experiment was selected as a surrogate for measuring the overall ability of students attending each class. The intent was to determine whether there were any significant differences between students' academic performance in the control versus treatment groups. The findings of prior studies indicate that cumulative GPAs are a good predictor of students' continued performance in classes. Students who perform well on exams in previous classes usually continue to perform well in their current classes.

A t-test was used to test the null hypothesis that the students' cumulative mean GPAs were equal between the control and treatment groups. The test statistic was 0.078 indicating no significant differences between the cumulative mean GPAs of the control and treatment groups.

Materials, Instruction, and Procedures

All sections of the course had the same instructor who provided the same discussion of subjects using the same notes and PowerPoint files for both the control and treatment groups. In addition, the instructor made every effort to sustain a consistency of attitude and pedagogical techniques for both groups. Students used the same textbook and supplemental materials and were given the same assignments (different sets of the same homework problems). Since the essence of the topics in the cost accounting course are very quantitative and system oriented, like curriculums in production and statistic courses, the content of the subjects do not change often over short periods of time. This continuity allows instructors to develop their own supplemental materials and employ them consistently over several semesters. In fact, the same pedagogical technique had been used by the instructor for several semesters prior to the experiment.

Four exams (including a comprehensive final exam) were administered during each semester. All examinations given to the control and treatment groups were the same and were kept by the instructor. The instructor believes that the exams remained secured, were administered properly and that no student had prior knowledge of test items.[1] Each exam consisted of multiple choice questions related to theoretical concepts and terminology in each chapter, and problem-solving questions with blanks for answers. The questions and problems in exams were similar to those of professional accounting examinations (i.e., Certified Management Accountants and Certified Public Accountants exams). There was no partial credit given, nor any curving system used on the examinations. Students were informed in advance that any material assigned and discussed in class could be included in the tests. Exam scores were weighted based on their contribution toward the final course grade (700 total exam points are possible, with each of the three mid-terms worth 150 points and the final worth 250 points). The final letter grades for the control group were based on the following lenient percentage scale: A = 90–100; B = 80–89; C = 70–79; D = 60–69, and F $<$ 60 percent of all the possible points; whereas, the letter grades for the treatment group were based on the strict scale of: A = 93–100, B = 85–92, C = 75–84, D = 65–74 and F $<$ 65 percent.[2]

The instructor distributed a detailed syllabus of the course to each group of students at the beginning of each semester, which specified the grading policy, schedule of exams, homework assignments, etc. In addition, the instructor announced and described the grading policy in detail in classes and ensured that students were well informed of it. Except for the grading scale, all presentation of subjects, supplemental materials, examinations, etc., were identical for both groups.

It is important to note that, like Johnson and Beck (1988), our treatment group grading scale imposes two potential effects on student motivation as discussed in the introduction. First, the stricter scale requires that students put forth more effort to achieve a particular grade; the minimum score required for each grade is higher under the treatment scale than under the control scale. If students are aiming for a particular grade, or to simply pass the course, they must put forth more effort under the treatment scale. However, the higher minimum requirements may also serve to discourage weaker students, who may now view acceptable grades as out of reach. Second, the

[1] Because students in the second semester were given the same examinations as students in the prior semester, information spillover effects were possible. As the exams were not given back to the students, this possibility should be minimal. Further, the exams used were the same as those used in semesters prior to the experiment; hence the earlier term students would also have the benefit of such positive spillover effects, if they did indeed exist.

[2] The change between groups was a planned and permanent transition to stricter standards by the instructor. The control grading scale had been used for many years prior, and the treatment scale remains in effect today. Hence, we could theoretically expand the sizes of both control and treatment groups in this study. However, as this would require a greater gap in time between groups, the effects of differences in instructor quality, course coverage, and differences in exams would be exacerbated, potentially offsetting any statistical gains from a larger sample size.

range of values for A and B grades is narrower under the treatment scale than under the control scale, which gives students a greater possibility of being able to improve their grade for a given increase in effort. This may also motivate students to put forth additional effort, but should not in and of itself, discourage lower-ability students.

RESULTS AND DISCUSSION

Table 1 presents the final grade distribution of the four exams for each section of the control and treatment groups. As mentioned earlier, the lenient grading scale was employed for the control group, whereas the strict grading scale was used for the treatment group.

In order to provide a preliminary look at differences between the groups, Table 1 also presents the distribution of the grades of the treatment group based on the scale applied to the control group (lenient scale). Differences in the grade distribution between the groups indicate that there was a shift of grades toward As and Bs in the treatment group as compared with the control group. Further, students in the treatment group who were assessed based on the strict grading scale had higher exam scores than students in the control group who were evaluated with the lenient grading scale. Sixty-six percent of students in the treatment group scored As and Bs compared to 55 percent in the control group. Employing a simple one-tailed t-test we find that average exam scores in the treatment group were significantly higher than those in the control group (significance level of α far less than 0.01). Hence, we have initial confirmation of the commonly found result in the literature: higher standards improve performance.

This was even more evident when the grades of students in the treatment group were rescaled to the control group's grading scale. When applying the lenient scale to both groups, 75 percent of

TABLE 1

The Grade Distribution of Students in Each Experimental Group

	Letter Grades						
	A	**B**	**C**	**D**	**F**	**W**	**Total**
Control Group[a]							
Section One	13	11	10	5	—	7	46
Section Two	10	16	11	2	—	6	45
Total	23	27	21	7	—	13	91
Percentage of Total	25.3	29.7	23.1	7.7	—	14.3	100
Treatment Group[b]							
Section One	12	14	9	2	—	5	42
Section Two	16	21	11	1	—	4	53
Total	28	35	20	3	—	9	95
Percentage of Total	29.5	36.8	21.1	3.2	—	9.5	100
Treatment Group[c]							
Section One	17	14	5	1	—	5	42
Section Two	20	21	7	1	—	4	53
Total	37	35	12	2	—	9	95
Percentage of Total	38.9	36.8	12.6	2.1	—	9.5	100

[a] Based on the scale of A = 90–100%, B = 80–89%, C = 70–79%, D = 60–69%, and F < 60%.

[b] Based on the scale of A = 93–100%, B = 85–92%, C = 75–84%, D = 65–74%, and F < 65%.

[c] Based on the scale of A = 90–100%, B = 80–89%, C = 70–79%, D = 60–69%, and F < 60%.

the treatment group scored As and Bs compared to 55 percent in the control group. A decreasing percentage of students withdrawing from the course can also be observed in the treatment group compared to the control group. The percentage of withdraws in the control and treatment groups were 14 and 9 percent, respectively.

The two sets of raw exam scores, students' gender and GPA, and the distribution data from Table 1 were used to test the null hypothesis of no difference in achievement between the control and treatment groups. We employed both parametric and nonparametric procedures.

First, using the grade distributions presented in Table 1, the Kolmogorov-Smirnov two-sample test was employed to determine whether the changes in the grading scales had a significant impact on students' performance. This nonparametric test determines whether differences in samples constitute convincing evidence of a difference in the processes or examinations applied to them, and can be considered more powerful than t-tests for small samples (Dixon 1954). For our purposes, a two-sample, one-tailed test is concerned with the agreement between two sets of sample values, and is approximated by the Chi-square distribution with df = 2. In short, this test can examine whether the scores of the treatment group were statistically "better" than those of the control group under different conditions. For example, the frame of examination for differences can include a common scale applied to both groups, the actual scale applied to both groups, and the actual scales applied to a single group. The first of these allows us to test whether grades would have been similar if both groups faced a common scale, such as the lenient scale. The second frame of examination allows us to test for differences in scores between the two groups under the actual scales used, while the third allows us to test whether the treatment group would have preformed differently under the lenient scale.

After examining the grade distributions in this manner, we also analyzed the effect of the strict grading scale using regression analysis. Combining both classes into a single data set, we regressed exam and course grades on student GPA, a gender indicator variable, and an indicator variable for the strict grading scale. To test for distributional effects across student ability levels, we also interacted the GPA and treatment variables and enter the interaction terms as a regressor.

Table 2 summarizes the results of the Kolmogorov-Smirnov tests. The critical value of Chi-square was 8.078 for the control group versus the treatment group when including the grade of W (withdraw) in the computation. Without including the grade of W in the computation, the critical value of Chi-square was 6.297. The significance levels of α were less than 0.02 and 0.05 respectively. This result allows us to reject the first null hypothesis and indicates that students in the treatment group did perform better on exams than those in the control group. This finding suggests that a strict grading scale can have a positive impact on students' performance, and is consistent with most results found in the literature.

The Kolmogorov-Smirnov test was also employed to examine whether there was any significant difference between the grade distributions in the control group versus the treatment group based on their actual lenient and strict scales. Data from Table 1 were used to compare the grade distribution of the control group and the treatment group based on their grading scale. The critical value of Chi-square was 2.403 for the control group versus the treatment group when including the grade of W in the computation, and was 1.371 when the grade of W was not incorporated in the computation. The significance levels of α were less than 0.40 and 0.60 respectively. This implies that there were no reasonable significant differences between grade distributions of students in both groups based on their actual grading scale. In other words, we do not have evidence of significant movement between grades, as might result from smaller grade ranges.

Finally, we use the Kolmogorov-Smirnov test to examine whether there were any significant differences in the grade distributions within the treatment group when applying the two grading scales. The critical value of Chi-square was 1.705 for the control group when including the grade of W in the computation. Without including the grade of W in the computation, the critical value

TABLE 2

Results of the Kolmogorov-Smirnov Tests

Grading Scale[a]	Pairings	Critical Values[b]	p-value
Control Group	Control versus Treatment Including the grade of W	8.078	$p < 0.02$
	Control versus Treatment Excluding the grade of W	6.297	$p < 0.05$
Actual Scales	Control versus Treatment Including the grade of W	2.403	$p < 0.40$
	Control versus Treatment Excluding the grade of W	1.371	$p < 0.60$
Control and Treatment Scales	Treatment versus Treatment Including the grade of W	1.705	$p < 0.50$
	Treatment versus Treatment Excluding the grade of W	1.884	$p < 0.50$

[a] Control group grading scale: A = 90–100%, B = 80–89%, C = 70–79%, D = 60–69%, and F < 60%. Treatment group grading scale: A = 93–100%, B = 85–92%, C = 75–84%, D = 65–74%, and F < 65%.

[b] Critical values of the Kolmogorov-Smirnov one-tailed test are based on the Chi-square distribution with df = 2.

of Chi-square was 1.884. The significance levels of α were less than 0.50 for both tests, indicating that there was not any statistically significant difference in the grade distribution for the treatment group based on the two grading scales. This result reveals that a majority of students in fact would have received the same letter grades for their performance in class under either the lenient grading scale or strict scale. Again, we see no evidence of significant movement between grades.

To summarize the above results, we found statistical evidence of an overall shift toward higher scores on examinations when a strict scale was applied. The shift was significant when the test scores of students in the treatment group were recalculated with the lenient scale and compared with the students' test scores in the control group. However, there was only a marginal (insignificant) improvement in the grade distribution of the treatment group versus the control group when the test scores of each group were evaluated based on their actual grading scales.

While at first glance these results may appear contradictory, Betts (1997) suggests a plausible explanation. Students who strictly prefer one letter grade over all others will continue to strictly prefer that letter grade after a marginal increase in the grading standard. The opposite will be true for students who are initially indifferent between one letter grade and the next. Raising the standard required to obtain each of the two grades results in the student strictly preferring the lower grade. These students place more value on short-term benefits of avoiding the extra classroom effort than future gains from higher grades. Finally, students who prefer the option representing the lowest level of effort will continue to exert the lowest level of effort after a change in standards. Hence, if most students have strict preferences for higher grades, educational standards will lead to greater student effort and achievement. On the other hand, if students prefer a low level of effort, higher standards may not improve outcomes. Because our treatment scale combined higher standards and tighter grade ranges, we cannot empirically separate these two theoretical causes of student motivation. However, because we did not see significant changes in grade distributions, these results may indicate that the former has a more pronounced effect on student effort.

Next, we tested the second null hypothesis, whether the strict grading scales has a differential impact on the achievement of students with different ability levels. In this regard, Atkinson and Feather (1966) suggest that most students who are high achievers have a stronger motive to attain success than to avoid failure. This implies that while the strict scale may motivate this group of students to work harder and focus on their schoolwork, this group of students could attain grades they aspired to with either scale. Johnson and Beck (1988) also supported this inference and indicated that a lenient grading scale should have a particularly deleterious impact on the performance of students with lower academic ability and students will acquire higher test scores if they are evaluated with a strict grading scale. Further, as suggested by Figlio and Lucas (2004), this result may be more pronounced for lower ability students in high-student-ability classes—a situation that clearly applies to university students in an accounting curriculum. As the risk of non-passing grades increases—due to higher standards and/or high-ability cohorts—lower ability students may be more significantly motivated than higher ability students who do not perceive such risks. That is, the high-ability students know they will be fine under either standard.

To more thoroughly examine the relationship between grading scale and student achievement, and to test for differences in this relationship across students with different GPAs, we used regression analysis. The raw scores on each of the four exams and the final score for the course were used as dependent variables. Combining both classes we first regressed exam and course scores on an indicator for the treatment grading scale (strict scale = 1, lenient scale = 0), a gender indicator variable (male = 1, female = 0), and student GPA. For the final course score, we also included an interaction term between GPA and the treatment indicator. These results are shown in Table 3, and show that while controlling for GPA and student gender, the treatment grading scale has a positive and highly significant effect on grades for all grades except those received on exam 2, where the coefficient on the treatment variable is negative and significant. For the overall course grade, being in the treatment group produced an average increase of approximately 29 points out

TABLE 3

Regression Results for Full Sample
(n = 164)

	Coefficient (Standard Error)					
	Total Score 700 Points	**Total Score 700 Points**	**Exam 1 150 Points**	**Exam 2 150 Points**	**Exam 3 150 Points**	**Exam 4 250 Points**
Intercept	120.78***	88.44***	56.38***	52.67***	19.73**	−8.00
	(21.78)	(28.92)	(9.08)	(10.14)	(9.46)	(15.32)
Treatment	28.88***	100.09**	10.61***	−4.35**	7.14***	15.48***
	(4.35)	(42.43)	(1.82)	(2.03)	(1.89)	(3.06)
Gender	−1.91	−1.86	1.07	1.04	−1.21	−2.81
	(4.38)	(4.35)	(1.83)	(2.04)	(1.90)	(3.08)
GPA	144.74***	154.81***	21.44***	24.38***	33.26***	65.66***
	(6.62)	(8.88)	(2.76)	(3.08)	(2.87)	(4.66)
GPA * Treatment		−22.19*				
		(13.16)				
R^2	0.7672	0.7713	0.3703	0.2968	0.4853	0.5886

*,**,*** Indicates significance at the 10 percent, 5 percent, and 1 percent levels, respectively.

of a possible 700 (roughly 4.1 percent average improvement). As expected, the coefficient on student GPA is consistently positive and highly significant. The coefficient on gender is insignificant in all cases. Notably, with the final course score as the dependent variable, the interaction between GPA and the treatment variable is negative and statistically significant. This suggests differential effects across students with different ability levels, but in contrast to results previously found in the literature, we find more pronounced effects for students at the lower end of the GPA spectrum.[3]

To further examine the notion that the strict grading scale had a larger effect on students in the lower end of the grade distribution, and to perhaps shed light on the scores received for exam 2, we partitioned the sample of 164 students into four subsamples of 41 students corresponding to ranked GPA categories and again regressed scores on the treatment and gender indicator variables and GPA. Subsample A had the 41 students with the lowest GPAs (2.473–2.92), subsample B had the next 41 students (2.945–3.204), subsample C had the second highest GPAs (3.208–3.444), and subsample D had the highest GPAs (3.448–4.0)[4]. Results for the total course score are shown in Table 4. Results from the individual exams are not reported here for brevity, but are available from the authors.

TABLE 4

Regression Results for GPA Subsamples for Total Course Score[a]
(n = 41)

	Coefficient (Standard Error)			
	Subsample A[b] (lowest)	**Subsample B[c]**	**Subsample C[d]**	**Subsample D[e] (highest)**
Intercept	431.57***	−15.12	474.13**	336.10***
	(155.57)	(113.64)	(178.01)	(71.03)
Treatment	41.707***	44.23***	28.87***	−1.24
	(11.64)	(5.50)	(7.68)	(5.62)
Gender	−3.80	−2.80	−6.32	−3.95
	(11.95)	(5.53)	(7.437)	(5.83)
GPA	27.87	186.61***	42.57	88.39***
	(55.68)	(36.54)	(53.01)	(19.57)
R^2	0.2941	0.7011	0.2935	0.3585

*,**,*** Indicates significance at the 10 percent, 5 percent, and 1 percent level, respectively.

[a] Total course score = 700 points.

[b] Subsample A GPA range = 2.473–2.920.

[c] Subsample B GPA range = 2.945–3.204.

[d] Subsample C GPA range = 3.208–3.444.

[e] Subsample D GPA range = 3.448–4.00.

[3] We also estimated the full model with the GPA/treatment interaction term for each of the exams. The interaction term is negative in all cases, but only statistically significant for exam 3.

[4] Similar results to those described below were found with alternative groupings of students by GPA. For example, we also divided students into two even groups and analyzed subsamples of different sizes with letter grade categories (i.e., 2.5–2.99, 3.0–3.49, 3.5–4.0). We chose to present the results with four evenly sized subsamples for ease of comparison and statistical equivalence. Other results are available from the authors upon request.

For the total course score (Table 4), we see that the coefficient on the treatment grading scale is positive and highly significant for all groups except those with the highest GPAs (subsample D). The insignificant coefficient on the treatment variable indicates that the treatment has no effect on the grades for the students with the highest GPAs. This result contrasts with other results in the literature (e.g., Betts (1997); Betts and Grogger (2003)), and suggests that students in an upper-level accounting course may react to grading standards differently than other students. Interestingly, the magnitude of the treatment coefficient is considerably lower for subsample C (second highest GPA category) compared to subsamples A and B, indicating that the strict scale has a smaller effect for these students than for their lower-GPA counterparts.

For each of the exam scores, the regression results by GPA category offer a slightly different perspective. As with the overall course score, for the first and third exams, the treatment grading scale has a positive and significant effect on all students except those in the highest GPA category (category D). Notably, GPA only has a significant effect on exam 1 scores for this group. This may be due to the fact that category D contains a wider range of GPAs than other categories, introducing more variation. But we can also hypothesize that these high-achieving students are behaving as described by Atkinson and Feather (1966); they will do what it takes to earn the highest grade, no matter how it is defined. However, within the group, the best students have the aptitude and ability to distinguish themselves even further, regardless of the scale applied.

The regression results for exam 2 tell a different story. Indeed, for this exam, none of our independent variables were significant determinants of grades except the treatment variable for the highest GPA category, where the effect was negative. The overall goodness of fit for these models was far below that for any of the other models estimated, indicating a notable deviation from the otherwise strong explanatory power of our independent variables. We have no formal or testable explanation for this anomaly, but can hypothesize that something caused students to behave differently on this test than on the others. It may be the case that students who performed poorly on the first exam enhanced their efforts going into the second exam, while students who did well on the first exam may have been overconfident. Hence, students performed in such a way that the explanatory power of the model was countered relative to the other exams and overall score. For the final exam, the treatment grading scale has a positive and significant effect on students' scores in all GPA categories, though the magnitude of the effect is again considerably higher for students in the lower GPA categories.

CONCLUSION

One of the primary missions of schools is generally considered to be the development of students' skills and talents, and assessing and communicating their achievements. Assigning grades to students' performance is a traditional and widespread means of documenting student achievement at universities. Theoretically, it has been suggested in the education literature that a grading system can also play an important role as a motivational tool to encourage students to exert greater effort toward their class work (e.g., Adams and Torgerson 1964; Norman 1981; Walvoord and Anderson 1998). We tested the applicability of this notion to accounting students by examining the impact of a lenient grading scale and a strict grading scale on students' performance in a cost accounting course. Under the strict grading scale, the level of "average" mastery (the grade of C) is also coincident with the minimum passing requirement of the professional accounting examinations. We hypothesized that most students who aspire to become professional accountants will likely have preferences for grades on the higher end of the distribution, and will therefore be motivated by higher standards. Unlike disciplines or classes where a minimum level of understanding is not required for advancement into the profession, we also hypothesized that all

students will have preferences for passing grades. Hence the distributional effects of higher standards on performance of students with different ability levels may be different than those previously found in the literature.

Our findings support the theory of the motivational function of grading scales, which suggests that applying a strict scale can encourage accounting students to put forth greater effort in their classroom performance, and thereby acquire better skills and understanding of the subject matter in preparation for professional exams. Nonparametric tests and regression results indicate that students in the treatment group who were evaluated with a strict scale obtained higher overall test scores than students in the control group who were evaluated with a lenient scale. Unlike other results in the literature, we found that this effect is especially pronounced for students in lower GPA categories. This indicates that a strict grading scale may be an important tool for preparing otherwise marginal students for the rigors of the professional accounting exams, where grades are pass/fail. Accounting students may understand that failure to master the material means the inability to advance through the professional exams. Hence, by anchoring student's expectations of the requirements for these exams, the strict grading scale further motivates effort. Interestingly, and also in contrast with results found in the literature, the stricter standard did not significantly affect the grades received by students with higher GPAs. We can speculate that these high achievers will do the work necessary to earn the top grade, regardless of how that grade is defined. We did find that variation in GPA within the group of students in the top GPA category has a significant effect on grades of these students. This indicates that even in the smartest group of students, the best students are able to perform significantly above their peers.

Our results also indicate that fewer students in the treatment group dropped the course. One might speculate that, since under the strict scale better performance is required to pass the course than under the lenient scale, the strict scale could also provide a better early warning of failing the course for the students who were not performing well. For instance, if a student has an average of 63 percent, his/her grade is considered to be an F based on the strict scale (since it is less than 65 percent) than a passing grade of a D (which is between 60 to 69 percent), based on the lenient scale. Likewise, there is more pressure on a student who has an average of 55 percent. Under the strict scale, his/her grade is 10 percent below the passing grade of 65 percent rather than a marginal 5 percent below the passing grade of 60 percent under the lenient scale.

In addition, of particular interest is the students' informal reactions to the strict scale. Through written and verbal comments to the instructor, students in the treatment group indicated that the strict scale was more challenging, made them work harder and study more in this course than other courses. They also indicated that as a result of the hard work they better understood the subjects. However, students' formal evaluations of the instructor were comparable for both groups.[5] Hence, both empirical and anecdotal evidence support the notion that a stricter grading standard may in fact motivate students to work harder and thus gain a better understanding of content.

The results of this study have important implications for accounting programs seeking to prepare their students for the rigor and high standard of pass/fail professional accounting examinations. As noted in Ingram and Howard (1998) in order for course objectives (such as the mastery of a given level of course content) to be met, instructors must create grading methods that are consistent with those objectives. In addition to the ubiquitous objective of greater understanding of course content, we can assume that preparing students for the high standard of professional exams is a likely objective of most accounting courses above the introductory level. The results of this study suggest that the motivational incentive created by increasing the minimum score required to

[5] On anonymous end-of-semester student evaluations the average student ratings of the overall effectiveness of the instructor were 4.51 and 4.57 (five-point scale) in the control and treatment groups respectively.

earn a given grade can help in meeting these objectives, especially so for otherwise marginal students. Notably, these results can be achieved without suffering significant losses in enrollment or instructor ratings.

Our study has limitations. First, students were not randomly selected from their respective populations. The participation of students in the experiments was a function of the university registration process and there exists an implicit self-selection bias. Consequently, it was not possible to accomplish a true random selection of students in both the control and the treatment groups. Furthermore, the experiment was conducted only in a cost accounting course, which dealt with a relatively small number of students in small classes. In effect, the findings may apply only to courses with similar content, setting, and class size. In addition, the course was designated only for accounting majors. Hence, the results presented here may not be applicable to classes with a diversity of student majors or courses that are not required such as accounting principles courses. In addition, this study examined a pragmatic strict scale rather than extremely strict scale, which make it almost impossible for students to attain grades of C or better. Finally, our stricter scale contained both higher minimum requirements for each grade and a smaller range of value for grades of A and B, both of which may influence student motivation. Hence, while we can use our results to hypothesize that higher minimum values were the primary motivation for additional effort, we cannot empirically attribute our results to one cause or the other.

These and other considerations may also warrant further inquiry. "Test anxiety" literature (e.g., Sarason et al. 1952; Alpert and Haber 1960) suggests that, strict grading scales can create more anxiety than lenient grading scales and the directional effect of anxiety depends upon each person's response to that anxiety. Students who are high in facilitating anxiety and low in debilitating anxiety are expected to perform better when evaluated on a strict scale. Conversely, students who are low in facilitating anxiety and high in debilitating anxiety should perform better when evaluated on a lenient scale.

Although a vast majority of instructors use various grading policies and the impact of different grading policies on learning is a basis of considerable debate among academics, the empirical work on this subject suggests that results are context dependent. This study provides an insight into the impact of a strict and a lenient grading policy on accounting students' achievement and supports the notion that an attainable strict grading policy can be used as an important pedagogical technique to motivate students to exert greater effort toward their class work. In addition, future empirical research is warranted to examine the impact of various grading policies such as plus or minus grading scales, as a motivational vehicle, on students' performance for different course settings.

REFERENCES

Adams, G. S., and T. L. Torgerson. 1964. *Measurement and Evaluation in Education, Psychology, and Guidance*. New York, NY: Holt, Rinehart & Winston, Inc.

Alpert, R., and R. N. Haber. 1960. Anxiety in academic situations. *Journal of Abnormal and Social Psychology* 61: 207–215.

Atkinson, J. W., and N. T. Feather, eds. 1966. *A Theory of Achievement Motivation*. New York, NY: Wiley.

Austin, J. D. 1978. Homework research in mathematics. *School Science and Mathematics* 79: 115–122.

Baldwin, B. A. 1980. On positioning the quiz: An empirical analysis. *The Accounting Review* 4: 664–671.

Becker, W., and S. Rosen. 1992. The learning effect of assessment and evaluation in high school. *Economics of Education Review* 11 (2): 107–118.

Betts, J. R. 1997. Do grading standards affect the incentive to learn? Working paper, University of California, San Diego.

——, and J. Grogger. 2003. The impact of grading standards on student achievement, educational attainment, and entry-level earnings. *Economics of Education Review* 22: 343–352.

Booth, P., and H. Winzar. 1993. Personality biases of accounting students: Some implications for learning style preferences. *Accounting & Finance* 33 (2): 109–120.

Briggs, S. P., S. Copeland, and D. Haynes. 2007. Accountants for the 21st Century, where are you? A five-year study of accounting students' personality preferences. *Critical Perspectives in Accounting* 18 (5): 511–537.

Butler, R., and M. Nissan. 1986. Effects of no feedback, task-related comments, and grades on intrinsic motivation and performance. *Journal of Educational Psychology* 78: 210–216.

Cherry, T. L., and L. V. Ellis. 2005. Does rank-order grading improve student performance? Evidence from a classroom experiment. *International Review of Economic Education* 4 (1): 9–19.

Clark, D. C. 1969. Competition for grades and graduate student performance. *The Journal of Educational Research* 62: 351–354.

Crooks, T. J. 1988. The impact of classroom evaluation practices on students. *Review of Educational Research* 58 (4): 438–481.

Dixon, C. 2004. Plus/minus grading: If given a choice. *College Student Journal* 38 (i2): 280–284.

Dixon, W. J. 1954. Under normality of several nonparametric tests. *Annals of Mathematical Statistics* 25 (3): 610–614.

Duke, J. D. 1983. Disparities in grading practice, some resulting inequities, and a proposed new index of academic achievement. *Psychological Reports* 53: 1023–1080.

Figlio, David N. and M. E. Lucas. 2004. Do high grading standards affect student performance? *Journal of Public Economics* 88 (9–10): 1815–1834.

Gardner, R. C., and W. E. Lambert. 1972. *Attitudes and Motivation in SL Learning*. Rowley, MA: Newbury House.

Geisinger, K. F. 1982. Marking systems. In *Encyclopedia of Educational Research*, 3. New York: The Free Press: 1139–1149.

Giacomino, D. E., and M. D. Akers. 1998. An examination of the differences between personal values and value types of female and male accounting and nonaccounting majors. *Issues in Accounting Education* 13 (August).

Gold, R. M., A. Reilly, R. Silberman, and R. Lehr. 1971. Academic achievement declines under pass-fail grading. *Journal of Experimental Education* 39 (3): 17–21.

Goldberg, L. R. 1965. Grades as motivants. *Psychology in the Schools* 2: 17–23.

Hales, I. W., P. T. Bain, and L. P. Rand. 1971. *An Investigation of Some Aspects of the Pass-Fail Grading System*. Annual Meeting of the American Educational Research Association, New York.

Ingram, R. W., and T. P. Howard. 1998. The association between course objectives and grading methods in introductory accounting courses. *Issues in Accounting Education* 13 (4): 815–832.

Johnson, B. G., and H. P. Beck. 1988. Strict and lenient grading scales: How do they affect the performance of college students with high and low SAT Scores. *Teaching of Psychology* 15 (3): 127–131.

McClure, J. E., and L. C. Spector. (2004). Plus/minus grading and motivation: An empirical study of student choice and performance. Working paper No 200401, Ball State University.

Miller, E., and G. Westmoreland. 1998. Student response to selective grading in college economics courses. *The Journal of Economic Education* 29 (3): 195–201.

Milton, D., H. Pollio, and J. A. Eison. 1986. Making sense of college grades. Available at: http://www.nabsa.org.

Nelson, I. T., and D. S. Deines. 1995. Accounting student characteristics: Results of the 1993 and 1994 Federation of School of Accountancy (FSA) surveys. *Journal of Accounting Education* 13 (4): 393–411.

———, and V. P. Vendrzyk. 1996. Trends in accounting student characteristics: A longitudinal study at FSA schools, 1991–1995. *Journal of Accounting Education* 14 (4): 453–457.

———, ———, J. J. Quirin, and R. D. Allen. 2002. No, the sky is not falling: Evidence of accounting student characteristics at FSA schools, 1995–2000. *Issues in Accounting Education* 17 (3): 269–287.

Neumann, R., 2001. Disciplinary differences and university teaching. *Studies in Higher Education* 26 (2): 135–146.

Norman, D.A. 1981. What is cognitive science? In *Perspectives on Cognitive Science*, 265–295. Norwood, NJ: Ablex Publishing Company and Erlbaum.

Oglesbee, T. W., L. N. Bitner, and G. B. Wright. 1988. Measurement of incremental benefits in computer enhanced instruction. *Issues in Accounting Education* 3 (2): 365–377.

Paschal, R. A., T. Weinstein, and H. J. Walberg. 1984. The effects of homework on learning: A quantitative synthesis. *The Journal of Educational Research* 78 (2): 97–104.

Quann, C. J. 1984. *Grade and Grading: Historical Perspectives and the 1982 AACRAO Study*. Washington, D.C.: American Association of Collegiate Registrars and Admission Officers.

Ridener, L. R. 1999. Effect of college major on ecological worldviews: A comparison of business, science, and other students. *Journal of Education for Business* (September/October): 15–21.

Sarason, S. B., G. Mandler, and P. G. Craighill. 1952. The effect of differential instructions on anxiety and learning. *Journal of Abnormal and Social Psychology* 47: 561–565.

Squires, B. 1999. Conventional grading and the delusions of academe. *Education for Health: Change in Training and Practice* 12 (March): 73–78.

St. Pierre, K., R. M. S. Wilson, S. P. Ravenscroft, and J. E. Rebele. 2009. The role of accounting education research in our discipline—An editorial. *Issues in Accounting* 24 (2): 123–130.

Terwilliger, J. S. 1971. Assigning grades to students. *Scott, Foresman and Company*. The Institute of Internal Auditors. Available at: http//www.theiia.org.

Vruwink, D. R., and J. R. Otto. 1987. Evaluation of teaching techniques for introductory accounting courses. *The Accounting Review* LXII (2): 402–408.

Walvoord, B., and V. J. Anderson. 1998. *Effective Grading: A Tool for Learning and Assessment*. San Francisco, CA: Jossey-Bass.

Wolk, C., and L. A. Nikolai. 1997. Personality types of accounting students and faculty: Comparisons and implications. *Journal of Accounting Education* 15 (1): 1–17.