The Influence of Grading Bias on Reinforced Concrete Exam Scores at Three Different Universities

Dr. Benjamin Z. Dymond, University of Minnesota, Duluth

Ben Dymond obtained his B.S. and M.S. degrees in Civil Engineering at Virginia Tech before obtaining his Ph.D. in Civil Engineering at the University of Minnesota Twin Cities. Ben is currently an assistant professor of structural engineering at the University of Minnesota Duluth.

Dr. Matthew Swenty P.E., Virginia Military Institute

Matthew (Matt) Swenty obtained his Bachelors and Masters degrees in Civil Engineering from Missouri S&T and then worked as a bridge designer at the Missouri Department of Transportation. He obtained his Ph.D. in Civil Engineering at Virginia Tech and worked at the Turner-Fairbank Highway Research Center on concrete bridge research. He is currently an associate professor of Civil Engineering at the Virginia Military Institute (VMI). He teaches engineering mechanics and structural engineering courses at VMI and enjoys working with the students on bridge related research projects and with the ASCE student chapter.

Dr. Chris Carroll, Saint Louis University

Dr. Carroll is an Assistant Professor in the Department of Civil Engineering at Saint Louis University. His experimental research interests focus on reinforced and prestressed concrete, while his engineering education research interests focus on experiential learning at both the university and K-12 levels. Dr. Carroll serves as a voting member on ACI Committee S802 - Teaching Methods and Educational Materials and is Chair of the Career Guidance Committee for the ASCE - St. Louis Section. He has eight years of formal experience with K-12 engineering education.

The Influence of Grading Bias on Reinforced Concrete Exam Scores at Three Different Universities

Introduction

Grading student exams fairly and effectively remains a challenge for many professors. Maintaining consistency among students on the same exam can be accomplished by using grading rubrics, grading the same question for all students at the same time, and giving similar questions each semester. However, there are still natural tendencies and preferences that affect how an individual professor grades. The objective of this research was to quantitatively assess how professor grading biases influenced exam scores in the same upper level course offered at multiple universities.

The course selected for analysis was an introduction to the design of reinforced concrete structures, a common course in many civil engineering curricula. Three professors at three different universities taught similar topics using their unique teaching styles and methods. During the semester, the same exam questions were posed to the students at each university. To understand how grading biases propagated throughout the exam questions, each of the professors re-graded the questions from all three universities at the conclusion of the course after the student identifiers were removed. A comparative study was then performed to determine if there were patterns in the grading results from each professor.

Background and Literature Review

Synthesis

Providing feedback on engineering exams is an important phase of the learning process for both the professor and student [1]. For the student, this leads to grades and a permanent record of achievement, which influences their attitude toward the course and even profession [2], [3]. For professors, student assignments provide them with feedback on how well students are learning and allow the professors to determine if the students have mastered the subject well enough to pass the class. Both students and professors have demonstrated skepticism whether assessment methods are applied uniformly and fair, but assessment continues to be an integral part of the learning process and an important result that students and professors routinely reference [3], [4], [5].

One of the most common types of assessed work in engineering is an exam. These vary in length, complexity, and type. Creating exams with multiple choice, fill in the blank, or true/false questions tend to simplify the questions, but the grades are objective. However, they may lack the depth of more complex problems [6]. Solely depending on this type of objective problem makes asking detailed analysis and design questions with multiple steps and correct approaches very difficult. Most engineering professors have limited background knowledge or practice in making exams, therefore the learning curve can be steep [2]. Many engineers continue to rely on open-ended problems because that is how they were taught and those types of questions mimic the approach to problem solving prevalent in the industry. While this may be a common way to test students, it results in significantly more work to grade the questions.

Open-ended exam questions are difficult because they allow for significant variation in the approach and there is often a range of correct answers. Well written questions can still be complex, which results in significant differences on the problem solving approach, especially in a large class. Besides the time commitment to grade these problems, there is also the issue of grading consistency [7], [8]. Some professors attempt to grade consistently by ensuring there are no student identifying marks to keep exams anonymous or by grading one problem at a time in an attempt to grade the exams uniformly. In settings when there are multiple sections of the same course with different instructors, the instructors may split up the problems to grade in order to have more uniform results among the sections. There have even been attempts to use computers to assist with grading complex design assignments, although this has not been proven universally effective at eliminating biases [1]. Sometimes scaling or curving the scores will be used to help account for bias. However, using these methods to adjust grades has the negative connotation that a professor does not know what they are doing and many students do not appreciate when grades are changed either up or down [2]. While all of these methods have been shown to eliminate or adjust for some forms of bias, there are still natural variations that can occur when grading exams.

For many years, professors have attempted to overcome grading differences by using a grading rubric. Various types of rubrics have been used in the past with some degrees of success [1], [9]. There is no universal consensus that rubrics create more uniform grading results. Rubrics can be written in many different ways with or without flexibility, detail, and sub-categories. Research has shown that highly prescriptive rubrics leave less flexibility and may result in abnormally high grades, while those with less description or fewer categories may result in lower grades [10]. Creation of a highly effective rubric depends on who is grading the exams and the degree of complexity and variation in the problems. Even the best rubrics will still leave some results open to interpretation. In some cases, reducing this interpretation may be overcome with training and careful implementation.

Ultimately, there are many causes of variation in grading that are hard to overcome, even with a detailed rubric that includes instructions and training. Because there is still a human element in grading, various forms of bias will affect the final scores [3], [11], [12]. The Halo effect describes many of the biases that a professor might bring into a grading session from grading previous work [13], [14]. There are other forms of bias that have been reported, which include attachment toward your students, gender, personality, work ethic, and personal bias [11], [14]. One research project focused on investigating 30 professors in the same department that graded undergraduate psychology papers. Significant differences were seen in the grading results of professors who graded their own class and professors who graded students outside their class. The exact form of bias was not identified, but there was a clear pattern in the final grades regardless of the grader [13]. A similar study was conducted on 90 undergraduate engineering projects that were independently graded by their project supervisor and a non-supervisor. Even though the supervisor worked closely with the student groups, there was no measurable grading bias [15]. Another study was performed by professors who graded undergraduate research projects in classes they taught and in classes taught by their colleagues. On average, the professors tended to grade their student's projects half of a letter grade higher [14]. An additional study focused on coordinating the efforts made by two professors who taught the same senior

level course at the same school during the same semester. All assignments and exams were made and graded together. The effort required was much greater than leading a course independently and the initial exam grades had large differences. The professors agreed on a grade for each student after discussing their grading methods and making compromises [7].

While many of these human biases may be alleviated by using anonymous grading systems, there are still some biases that are difficult to overcome. Other forms of bias may be due to a person's professional work experiences, the year they went to engineering school, and personal compassion. There have been reported differences in how full-time professors grade versus full-time engineers working in the industry. A study reported that when external judges and faculty judges both graded the same capstone project, the external judges gave higher grades [16]. Other studies have conflicting reports about whether tenured versus non-tenured faculty graded easier [5], [7], [8]. Additionally, a study showed that bias might occur if grading is extended over a long period and breaks are taken between grading sessions, but the bias is not predictable nor significant [17]. This is a type of interrater reliability that occurs when a grader is not consistent in his or her grading over time [18].

Knowledge Gap Filled

There are many factors that can affect how exams are graded for civil engineering students, especially in design-based courses. The research study described herein focused on eliminating many of the human grading biases and exploring differences in grading based on personal grading preferences and style. This research was different from previous studies investigating grading biases in four unique ways:

1. The study focused on an upper level civil engineering design course. Most previous studies were performed in humanities courses with open-ended essay style questions or in basic engineering mechanics classes taught as analysis courses (i.e., statics or dynamics)
2. In this reinforced concrete design course, many of the questions were analysis and design related and required a series of complex steps to complete the problem correctly. Some problems had a range of correct answers, and in many cases, multiple correct approaches.
3. During this study, some human biases were eliminated among graders. All of the questions were written together as a team and the student identifiers were removed prior to grading the exams. The teaching styles and learning environment were not identifiable on the exams.
4. Previous studies have focused on multiple professors teaching the same class at the same university during the same or different semesters. In this case, the same course was taught at three different universities by three different professors during the same semester.

Methods

Professor Portraits

The three professors in this study taught with their preferred methods and organized the semester to fit the schedule at their university. The only point of similarity was that all of the exam questions that were compared in this study were the same and were made together prior to administering exams. While the professors had been aware of each other's teaching methods and

styles, there was no attempt to unify any of the classes. This was left to the prerogative of each professor. Additionally, all of the professors were untenured at their university when this study was performed. All of the graders attended graduate school within a similar period, had previously been practicing engineers with at least two years of design experience, and had previously taught a reinforced concrete course. Each person had developed their own unique opinions about how to apply grading procedures to a structural design class.

University A

University A is a small, public, liberal arts school in the south-Atlantic region (Carnegie Classification, Baccalaureate Colleges: Arts & Sciences Focus). The school only has undergraduate engineering programs, approximately one fourth of the student population majors in engineering, and the civil engineering graduating class averages approximately 60 students per year. The civil engineering degree is a general degree, which means that all students are required to take courses in at least seven different subareas of civil engineering. Within this structure, reinforced concrete design is a required course. The majority of students take reinforced concrete the second semester of their junior year. During the semester when this study was performed, there were two sections of reinforced concrete, one with 13 students and the other with 14 students (i.e., 27 students total).

At University A, the course was taught primarily with skeleton style notes. The students were provided a rough outline of the material in the notes and attending class was the only way to gather the critical information to complete the notes. The students were routinely required to work together in small groups and solve problems during class. The semester consisted of 25 lectures (each lecture was 75 minutes in length), three 75 minute exams administered during class time (50% of the final grade), and a 3 hour comprehensive final exam (30% of the final grade). All exams were worth a total of 100 points each. During exams, the students were only allowed to use their personal copy of the ACI 318-14 code [19]. Homework was required and contributed to 10% of the final grade, but the scores were based on participation. A one day concrete beam testing demonstration was included at the end of the semester to show under reinforced and unreinforced beam behavior. A semester long group design project (design the main components of a parking garage) was included that contributed to 10% of the final grade. Prior to each exam, a one-hour, optional evening review session was provided, which the majority of the students attended.

The grading style at University A focused on content and method. A scoring rubric was made prior to grading each exam question. Extensive partial credit was given for providing the correct thought process and writing down the correct steps in solving the problem. Deductions made for mistakes were not carried through the problem. Limited points were taken off for minor math errors or units.

University B

University B is a midsized, public, master's university in a medium density city in the West North Central Region (Carnegie Classification, M1). The university has six colleges/schools, a medical school branch, and a graduate school, which primarily offers M.S.

degrees (although Ph.D. degrees can be obtained in some majors or cross-disciplinary programs). Within the engineering college, there are five types of engineering disciplines that offer a B.S. degree. The civil engineering graduating class averages approximately 70 students per year. The civil engineering degree has four optional tracks (environmental/water resources, geotechnical, structural, and transportation), but the reinforced concrete design course is required for all students. The majority of students enroll in reinforced concrete within one year of graduation. During the semester when this study was performed, 22 students were enrolled in one section of the course.

At University B, the course was taught primarily with skeleton style notes with material provided by the professor via tablet. The students frequently worked in small groups on in-class examples. The course material was presented in 50 minute intervals over the course of 41 lectures. Two midterm exams were administered during the 50 minute class time (40% of the final grade) and a two hour comprehensive final exam was given at the end of the semester (25% of the final grade). All exams were worth a total of 100 points each. During each exam, the students were allowed to use their personal copy of the ACI 318-14 building code [19] and one 3 in. by 5 in. notecard. During exams, students were also asked to sit with one empty chair between themselves and their neighbor. No review of course material was conducted prior to each exam. Homework was required and contributed to 25% of the final grade. A multi-day design, construction, and laboratory testing experience was implemented for students to gain a deeper knowledge related to the bending and shear capacity of reinforced concrete beams. This experience was part of the course homework grade. A design project was completed over the course of the semester, which contributed to 10% of the final grade.

The grading style at University B focused on the problem solving process rather than the numerical result. A scoring rubric was made for each problem prior to grading each exam question. Partial credit was given for providing the correct thought process and writing down the correct steps in solving the problem. Deductions made for mistakes were not carried through the problem. Points were taken off for math errors or units.

University C

University C is a large, private, not-for-profit doctoral university in a dense city in the West North Central United States (Carnegie Classification, R2). The university has twelve colleges/schools, including a law school and school of medicine. The College of Engineering has six ABET accredited undergraduate programs and offers both M.S. and Ph.D. degrees. The civil engineering program has approximately 70 students enrolled and has a graduating class of about 15-20 students per year. The departmental curriculum requires that all students take courses in each subarea of civil engineering. Of the required courses, students take an introduction to structural design course during the second semester of their junior year, which combines reinforced concrete and steel design. During the semester in which this study took place, the class was unusually small with eight students.

At University C, the first half of the introduction to structural design course focused on reinforced concrete. The material was presented with skeleton file notes and a tablet, and the course incorporated various active learning strategies where students frequently worked in small

groups to solve in-class problems. The course content was presented through two 75 minute class periods (3 credit hour course) and a two hour lab (1 credit hour course) each week. Fourteen lectures and half of the lab sections focused on reinforced concrete design. The lecture portion of the course included four non-comprehensive exams, each worth 20% of the final grade. Two of the exams were focused on reinforced concrete. The exams were given during the lab section of the course to ensure enough time, and each exam was worth 100 points. During the exams, students were allowed to use one page of hand-written notes and a Styrofoam block provided in class by the professor to explain the equivalent rectangular stress block. No review of course material was conducted prior to each exam. The remaining percentage of the course grade was made up of homework (15%) and quizzes (5%) and half of each were focused on reinforced concrete. The reinforced concrete portion of the lab used a project similar to the Egg Protection Device competition held biannually at the American Concrete Institute Convention and Exposition [20], [21]. Students designed and fabricated two small concrete frames and tested one under static load and the other under impact load. The project was worth 50% of the lab grade.

The grading style at University C focused on the steps used to solve the problem using a rubric with weighted point values based on the complexity of each step. Point values were further divided within more complex steps to provide partial credit across all exam questions. Deductions made for mistakes were not carried through the problem. Additionally, points were deducted for math and unit errors.

Creation and Administration of Exams

All of the exam questions were written together as a team and the student identifiers were removed prior to grading the questions. No joint, formal grading rubric was created, but each professor used an individual rubric in an attempt to provide consistent grades. The teaching styles and learning environment were not identifiable on the exams. A total of 35 questions were created, graded, and analyzed. Due to variation in the topics covered in the course at each university, not all questions were asked of all students. Professors A and B assigned all 35 questions, while professor C only assigned 18 of the 35 questions. The students at each school could choose to opt-in to the study on each exam. A total of 57 students had the option of including their exam results in the study. The minimum number of students to answer a question in the study was 43, while an average of 50 students answered all of the exam questions.

Topics Typically Covered in a Reinforced Concrete and Topics Covered in this Study

The topics shown in Table 1 are frequently covered in an initial reinforced concrete design course. At university A and B, the courses covered 18 of the same topics. University C covered 13 of the topics, focusing on the core content of a reinforced concrete design course. All three universities covered 13 topics in common.

Table 1. Topics frequently covered in a reinforced concrete design course

Topic #	Topics Frequently in a Reinforced Concrete Course	Topics Taught at Each University		
		A	B	C
1	Review of structural analysis / mechanics of materials / deformable bodies	X	X	X
2	Material properties of concrete and steel rebar	X	X	
3	Load and resistance factor design process and limit states	X	X	
4	Loads and load paths	X	X	
5	Live load reduction	X	X	
6	Flexural un-cracked and cracked transformed-section analysis	X	X	X
7	Ultimate flexural capacity using equivalent rectangular Whitney stress block	X	X	X
8	Flexural failure types	X	X	X
9	Deflections	X	X	X
10	Reinforcement detailing	X	X	X
11	Flexural beam design with known and unknown dimensions	X	X	X
12	Flexural analysis of doubly-reinforced beams	X	X	X
13	Flexural analysis of non-rectangular beams	X	X	X
14	One-way slab analysis	X	X	
15	One-way slab design		X	
16	Shear analysis	X	X	X
17	Shear design	X	X	X
18	Column analysis	X	X	X
19	Column design	X	X	X
20	Footing analysis		X	
21	Footing design			
	Count:	18	20	13

Topics Covered in the Exam Questions

Out of the 21 topics typically covered in a reinforced concrete design course, there were 14 topics that were included in exam questions in this study as shown in Table 2. The topics not studied included two that are typically thoroughly covered in a structural analysis course (#3 and #5) and four (#14, #15, #20, and #21) that are only briefly explored in a basic reinforced concrete design course. These four topics are frequently cut or abbreviated due to time constraints, and they are good candidates for coverage in an advanced reinforced concrete design course. The other topic not covered (#11) was explored extensively through the design project at University A and B.

Two different types of questions were asked on the exams: short answer and computational. The short answer questions required the students to recall knowledge without any prompts and the computational questions required the students to complete analysis or design related to a common reinforced concrete topic. Out of the 14 topics included on the exams, only one was exclusively covered with a short answer question (topic #2). Out of the remaining 13 topics, there were seven covered with both question types and six covered with only a computational question. The value of short answer questions ranged from 2 to 8 points and the value of the computational questions ranged from 5 to 18 points.

Table 2. Number of participants and exam question weights, topics, and types

Question #	Question Weight	# Participants (57 max)	Topic #	Question Type	
				Short Answer	Computational
1	15	56	1		X
2	10	56	6		X
3	18	56	6		X
4	12	56	6		X
5	10	54	9		X
6	10	48	4		X
7	15	48	4		X
8	3	48	2	X	
9	8	48	2	X	
10	3	48	2	X	
11	2	48	2	X	
12	4	48	6	X	
13	10	54	12		X
14	10	54	12		X
15	10	50	7, 8		X
16	10	50	7, 8		X
17	10	50	7, 8		X
18	15	43	10		X
19	10	43	10		X
20	4	44	8	X	
21	4	44	8	X	
22	4	44	10	X	
23	4	44	7	X	
24	4	46	12	X	
25	10	54	13		X
26	10	54	7, 13		X
27	10	54	16		X
28	15	54	17		X
29	5	54	16		X
30	5	53	18		X
31	10	53	18		X
32	15	53	18		X
33	10	45	19		X
34	2	45	16	X	
35	4	45	4	X	
Average:	8.6	50	Count:	13	22
Min:	2	43			
Max:	18	56			

Data Analysis and Statistics

Basic Data Analysis

Simple statistics, such as the mean and ratio of two mean values, were used to obtain an overall idea of how the scoring from each grader related to the entire data set. The total number of points assigned by the professors to all 35 of the problems prior to grading the questions summed to a value of 301. The highest weighted question was 18 points, the lowest weighted

question was 2 points, and the distribution of question weights is shown in Table 3. An overall average of the scores assigned by each professor was calculated by summing the averages of each individual problem.

Table 3. Distribution of question weights

Question Weight (points)	2	3	4	5	8	10	12	15	18
Number of Questions	2	2	7	2	1	14	1	5	1

Tukey Method

The goal of the additional statistical analysis was to answer the following question: Are there significant differences among the grades generated by each professor for each problem? Because there were three graders for each problem (i.e., more than two levels of the independent variable), a multi-comparison procedure was used to determine if there were any statistical differences in the scores for each problem. This was assessed using Tukey's method [22] to compare individual means in the analysis of variance. Tukey's method was selected because of its ability to investigate all possible pairwise comparisons with equal sample sizes (e.g., grader A vs. B, grader B vs. C, and grader A vs. C).

Initially, an overall *p*-value was used to determine if there were any significant differences among any of the mean grades, considering the entire dataset. If there were significant differences, a comparison of sets of two means at a time was conducted to determine specifically where the significant differences were located (using Tukey's method). A confidence level of 95% (significance level, $\alpha = 0.05$) was used to determine significant differences in the grades. During the pairwise investigations, a *p*-value less than $\alpha = 0.05$ indicated that there were significant differences between the two graders in that comparison. The Tukey method results indicated which grader was significantly different when two of the three pairs had *p*-values less than $\alpha = 0.05$.

Results

Basic Data Analysis

Out of 301 total points available, the sum of the average score for all 35 questions ranged from 225 (75%) to 247 (82%) points awarded. Total scores from Grader A were 5 and 7 percentage points higher than total scores from Graders B and C, respectively. Total scores from Grader B were 2 percentage points higher than total scores from Grader C. These differences may indicate a difference in letter grades assigned by the professors over the course of a semester. Figure 1 shows the comparison of the average score for all 35 questions among the three graders.

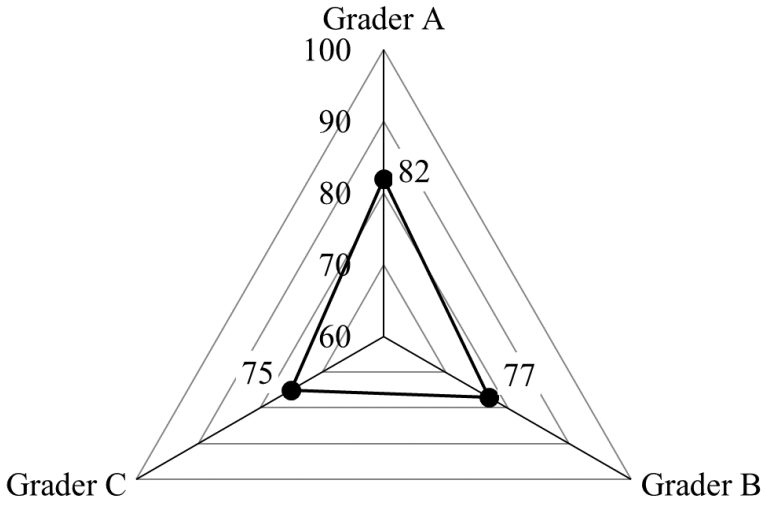


Figure 1. The overall average score for all 35 questions for each grader as a percentage

Tukey Method

Overall, the *p*-value for this data set indicated that there were highly significant differences within the overall results ($p < 0.0001$). Within the entire data set, *p*-values from the Tukey method showed significant differences between graders A and B ($p < 0.05$), no significant differences between graders B and C ($p > 0.05$), and highly significant differences between graders A and C ($p < 0.0001$). Given the overall results, a more in-depth analysis was warranted.

Of the 35 total questions, 12 had *p*-values that were less than 0.05, which indicated significant differences within the grades from that particular question. In other words, 34% of the questions had statistically significant differences in grades that may be associated with some type of grading bias, while 66% of the questions had grades with no statistically significant differences among the three professors. Within the Tukey method results, six (i.e., half) of the problems that had statistically different grades (*p*-value < 0.05) had point totals of five or less (on a 100 point exam), which meant they were short answer or simple problems. For each of the statistically different questions, when two of the three pairs of graders had *p*-values less than 0.05, the remaining pair indicated which professor was statistically different. For example, comparing three graders required the comparison of three pairs: grader A vs. B, grader B vs. C, and grader A vs. C. If the only pair that was not statistically different was the pair containing grader A vs. B, grader C is a member of both of the other pairs (B vs. C and A vs. C) and is the statistically different grader. Table 4 shows the *p*-values for each of the grader pairs for questions where the Tukey method indicated significant difference in the grades.

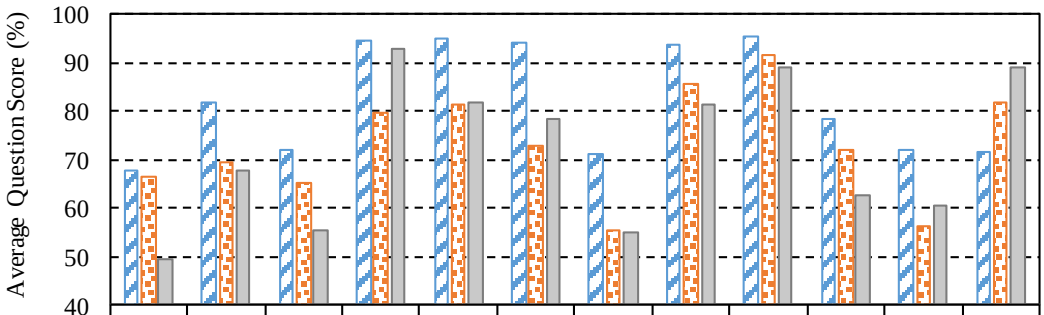
Table 4. Questions with statistically different grades (*p*-values < 0.05)

Question #	Question *p*-value	*p*-values for Pairs of Graders			Question Weight (points)	Statistically Different Grader
		A vs. B	B vs. C	A vs. C		
3	0.0011	0.9643	0.0057	0.0025	18	C
5	0.0002	0.0023	0.9214	0.0006	10	A
7	0.0131	0.4434	0.1886	0.0096	15	N/A
8	0.0004	0.001	0.0033	0.9349	3	B
12	0.0066	0.0139	0.9934	0.019	4	A
22	0.0048	0.0046	0.6748	***0.0501***	4	A
24	0.0337	***0.0675***	0.9967	***0.0562***	4	A
26	0.0047	***0.0948***	0.4578	0.0036	10	A
27	0.0481	0.2758	0.6265	0.0393	10	N/A
29	0.0343	0.5388	0.2715	0.0266	5	N/A
32	0.0002	0.0002	0.4837	0.0106	15	A
35	0.0126	0.1946	0.4226	0.0091	4	N/A
Count:	12	5	2	10		

shading indicates *p*-values for pairs of graders < 0.05

bold italics indicates *p*-values near 0.05 used to identify the statistically different grader

A comparison of how the average scores for each question that was statistically different was warranted to investigate how the graders compared to each other. The Tukey method results indicated that, typically, grader A had the highest average question score compared to graders B and C. Grader C typically had the lowest average question score and grader B was frequently in the middle. Four of the 12 questions did not have *p*-values that indicated one grader was statistically different from the other two. Figure 2 shows how the statistically different grader related to the average question score from each grader. In 11 of the 12 questions, Grader A gave the highest score. In eight of the 12 questions, Grader B gave higher grades than Grader C. There was a clear grading pattern throughout.



	3	5	7	8	12	22	24	26	27	29	32	35
Grader A	68	82	72	94	95	94	71	94	96	79	72	72
Grader B	67	69	65	80	81	73	55	86	91	72	56	82
Grader C	50	68	55	93	82	78	55	81	89	63	61	89
Question Weight	18	10	15	3	4	4	4	10	10	5	15	4

= statistically different grader (if applicable)

Question Number

Figure 2. Comparison of average question scores to the statistically different grader

How Professors Graded Students Enrolled in their Course Compared to the Group

The literature review indicated mixed results in regards to how professors graded their own students versus students taught by a different instructor. In different studies, professors both did and did not grade their students more favorably [13], [14], [15]. In this study, the researchers were interested in how the grading of their own students compared to that of the total group of students. Typically, grader A assigned the highest grades regardless of the student group, grader C assigned the least amount of points, and grader B was in the middle. This trend aligns with the Tukey method results. The data in Figure 3 did not indicate any specific trends related to grading students taught by the instructor more favorably.

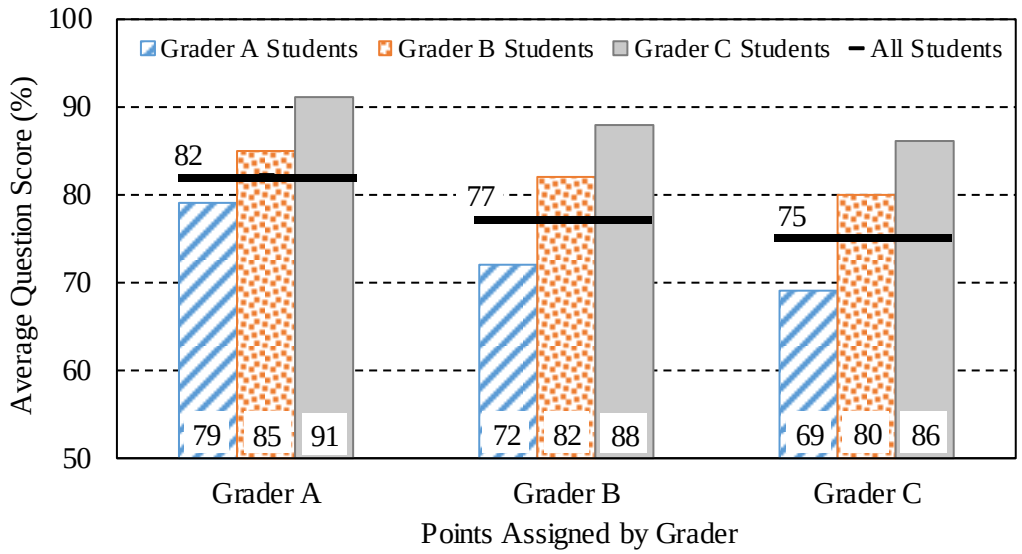


Figure 3. How professors graded students enrolled in their course compared to the group

Summary and Conclusions

The following conclusions were drawn from this study, which investigated potential bias present during grading of reinforced concrete exams:

1. Out of the 35 problems analyzed, there were statistically different results in only 12 problems (34%). The majority of the problems did not have statistically different grades between three professors.
2. Of the 12 questions that had statistically different results, half were short answer and half were computational problems. The type of problem was not an indicator of statistical difference.
3. In the 12 questions with statistically different grades, grader A gave the highest average score on 11 of the problems. Grader C gave the lowest score on eight of the 12 problems. Grader B frequently assigned grade values between graders A and C.
4. While the majority of problems showed no statistical difference in grades, the overall score computed by summing the average results of all 35 problems indicated the following differences: Grader A gave an average of 82%, grader B gave an average of 77% (5 percentage points less than grader A), and Grader C gave an average of 75% (7 percentage points less than grader A).
5. The question scores did not reveal a bias toward the professor's own students. The grading patterns were the same regardless of the student set graded (student sets varied by university).
6. External factors were held constant to eliminate as much bias as possible when grading the exams. This included eliminating student identifiers. The results indicated that an individual grader did have bias that may have been present when grading. This bias was likely manifested in the grading rubric when valuation was placed on a particular error for each problem. While these differences in valuation did show clear patterns in the grades, the overall scores varied by 7 percentage points or less. This is less than one letter grade using a traditional A through F assessment scale.

Acknowledgements

The authors would like to thank Dr. Salli Dymond for her proficiency in statistics and savvy manipulation of SAS to gather Tukey method results.

References

[1] T. Ahoniemi, E. Lahtinen and T. Reinikainen, "Improving Pedagogical Feedback and Objective Grading," in *ACM Technical Sumposium on Computer Science Education*, 2008.

[2] A. Ieta, T. Doyle and R. Manseur , "Grading Techniques for Tuning Student and Faculty Performance," in *ASEE Annual Conference and Exposition*, 2010.

[3] R. Manteufel and A. Karimi, "Grade-Based Correlation Metric to Identify Effective Statics Instructors," in *ASEE Annual Conference and Exposition*, 2010.

[4] E. Aimiuwu, "A Case of Bias in Teaching, Grading, and Plagiarism," in *AMCIS 2012 Proceedings* , 2012.

[5] A. C. Krautmann and W. Sander, "Grades and Student Evaluations of Teachers," *Economics of Education Review*, pp. 59-63, 1999.

[6] S. Habeshaw, G. Gibbs and T. Habeshaw, "53 Problems with Large Classes: Making the Best of a Bad Job," Technical and Educational Services, Bristol, 1992.

[7] A. Karimi, "Bringing Uniformity in Topic Coverage and Grading Fairness in Multiple Sections of an Engineering Course," in *International Mechanical Engineering Congress and Exposition*, 2015.

[8] C. E. Work, "Nationwide Study of the Variability of Test Scoring by Different Instructors," *Journal of Engineering Education*, pp. 241-248, 1976.

[9] K. Becker, "Grading Programming Assignments Using Rubrics," in *Conference on Innovation and Technology in Computer Education*, Thessaloniki, 2003.

[10] M. K. Thompson, L. H. Clemmensen and B.-U. Ahn, "Effect of Rubric Rating Scale on the Evaluation of Engineering Design Projects," *International Journal of Engineering Education*, 2013.

[11] J. Malouff, "Bias in Grading," *College Teaching*, pp. 191-192, 2008.

[12] R. D. Manteufel and A. Karimi, "Proposed Renormalized Grade Point Average Accounting for Class GPA," in *ASEE Annual Conference and Exposition*, 2011.

[13] I. Dennis, "Halo Effects in Grading Student Projects," *The Journal of Applied Psychology*, pp. 1169-76, 2007.

[14] B. McKinstry, H. Cameron, R. Elton and S. Riley, "Leniency and Halo Effects in Marking Undergraduate Short Research Projects," *BMC Medical Education*, 2004.

[15] A. Nyamapfene, "Involving Supervisors in Assessing Undergraduate Student Projects: Is Double Marking Robust?," *Engineering Education*, pp. 40-47, 2012.

[16] A. N. Lyerly and G. Dixon, "Grading the Capstone Written Design Reports: A Comparison of External Judges and Faculty Scores," in *ASEE Annual Conference and Exposition*, 2016.

[17] T. Ward, N. Jackson, J. Issitt and B. Baruah, "Is There Consistency in Grade Allocation When Assessing Student Presentations?," in *International Conference on Information Technology Based Higher Education and Training*, 2016.

[18] S. E. Stemler, "A Comparison of Consensus, Consistency, and Measurement Approaches to Estimating Interrater Reliability," *Practical Assessment, Research & Evaluation (PARE) Journal*, 2004.

[19] American Concrete Institute, Building Code Requirements for Structural Concrete and Commentary, Farmington Hills: American Concrete Institute, 2014.

[20] C. Carroll, "Competition Based Learning in the Classroom," in *120th ASEE Annual Conference & Exposition*, Atlanta, 2013.

[21] American Concrete Institute, "Student Competitions," [Online]. Available: https://www.concrete.org/students/studentcompetitions.aspx. [Accessed January 2018].

[22] J. W. Tukey, "Comparing Individual Means in the Analysis of Variance," *Biometrics*, vol. 5, no. 2, pp. 99-114, 1949.