Grading Bias in the Grading of Writing Assignments in the Principles of Macroeconomics: An Exploratory Study

Geok H. Simpson[1] and Roger L. Adkins[2,3]

ABSTRACT

This case study explores potential grading bias when grading writing assignments. Given seven grading criteria, the mean difference between the scoring of two graders is statistically significant for three criterions: identification of concepts, organization, and writing style and mechanics. The differences in identification of concepts may be due to asymmetric information available to the graders while the differences in the other two presentation type criteria are likely due to the graders' subjective preferences. Differences in these three criterions account for over 81 percent of the differences in the writing assignments' overall scores when all criterions are weighted equally.

Introduction

Economics instructors employ assessment tools in and beyond the classroom to determine student comprehension and performance. Assessments are also useful to obtain feedback from students to identify learning problems. Some of these assessments may be in the form of writing assignments – problem sets, short essays, or research papers. Writing assignments are usually given when freedom of response from students is desired. As a learning instrument, writing assignments can help students develop the skills of economic reasoning and analysis because they require the students to explain what they are thinking (Hansen 1998). As a testing and grading device, writing assignments are appropriate to evaluate for higher-ordered economic thinking skills (Hansen, 1986 and 2001). Although student writing has long been recognized as being important for the learning of economics and the evaluation of mastery of economics, it is not as widely utilized as some other forms of assessments (Siegfried et al. 1996). The usual explanations for this are diseconomies in scoring, relatively limited coverage of the material, and greater potential for grading bias (Walstad 2001).

Even though the three explanations for the lack of use of writing assignments in economics classes appear to be self-evident, the third reason –grading bias– warrants further attention. Grade determination is a very important activity for students and instructors alike, and the prospect of bias occurring in the process is somewhat disconcerting. Instructors generally recognize that the subjective nature of writing assignments implies that some uncertainty in measurement, i.e. grading bias, may be present. However, little work has been done in this area to advance our understanding of this phenomenon. Such research should be helpful if it enables instructors to actively reduce/correct the problem given instructional goals. Increasing instructor awareness about the costs and benefits of giving writing assignments should also lead to a more objective approach to overall course assessment (Walstad 2001). Importantly, if student learning is a function of accuracy of grading by the instructor (Becker 1982), reducing the potential for bias in the grading of written work may ultimately promote student learning through writing. Hansen (1998) proposes that evaluations of the writing assignments can be used to reinforce the focus on developing student knowledge of the subject matter and the complementary skills to demonstrate this knowledge.

Seeing that the potential for grading bias with writing assignments has not been examined in the literature, we seek to begin the preliminary discussion with this case study by investigating two fundamental and related questions. Specifically, we attempt to detect for the presence of scoring bias in the

[1] Assistant Professor of Economics, The University of Texas – Pan American, Edinburg, TX 78541, gsimpson@utpa.edu.
[2] Professor of Economics, Marshall University, Huntington, WV 25755, adkinsr@marshall.edu.
[3] The authors thank two anonymous referees for helpful comments.

grading of written assignments and simultaneously identify where that bias exist. The next section describes the experimental design and presents the grading criteria. Following that, we detail the methods of analysis and provide the empirical results. The final sections summarize and conclude, and identify issues for future research.

Experimental Design

This case study examining grading bias in the grading of written assignments has a sample of 80 student essays. Each essay was independently graded by two economics professors. The data pertinent to our analysis is the *differences* in the scores given by the two graders for a particular essay. This information is used to detect for the presence of grading bias and to examine the patterns of the grading bias, if any.

The essays are the work of 41 students at a medium-size comprehensive state university; all but two of the students who participated in the study turned in two essays each. Student nonparticipation, student withdrawal, and missing assignment affected our sample size because they reduce the number of available essays. However, given the purpose of our study and the nature of the data, we are unable to control for this bias using the usual sample selection bias tests and concede that this inability is a limitation of the study. All the students were enrolled in the *same* section of a Principles of Macroeconomics class; Principles of Macroeconomics is taught as the second course in a two-semester sequence at this institution. We utilized only one section of the course to control for the instructional aspects of the study (i.e. all the students received the exact same lectures, assignments, and exams, and were subject to the same syllabus). Using a Principles class for this study is advantageous because the relatively large number of students in the class provided a reasonable sample size for an exploratory study. Furthermore, because the quality of written work in Principles classes tends to vary greatly, this increases the likelihood that our sample would include an assortment of "weak" to "excellent" essays.

The essays were a required part of the class. They were designed based on Hansen's (1986, 1998, and 2001) suggestions for successful economics writing assignments to develop proficiencies in accessing existing knowledge, displaying command of existing knowledge, interpreting existing knowledge, interpreting economic data, and applying existing knowledge; the assignments did not require the students to demonstrate Hansen's proficiency in manipulating economic data or creating new knowledge. Each writing assignment was linked to a collection of articles selected from the popular press. This is appropriate given that materials from the popular press are the second most widely assigned resources in introductory economics classes behind textbooks (Becker and Watts 1996). The class instructor chose the readings to control for the uniformity of topics across the students' essays and to ensure that they encompass the economic concepts covered in class. For each batch of articles, the students were instructed to submit a 2-3 page essay that summarizes and analyzes the articles using the economic concepts learned in class.

As previously mentioned, each essay was independently graded by two economics professors. Both teach Principles of Macroeconomics regularly at their respective institutions. One grader was the class instructor and the other a faculty member at another university. In other words, the first grader interacted with the students and was responsible for the class syllabus, lectures, and coursework (the reading and writing assignments, as well as the exams). The second grader, on the other hand, did not have any contact with any of the students in the study and was not responsible for any aspect of the class. The latter's only tie to the class is through this study, that is grading the students' essays to provide a point of comparison for the scores given by the class instructor; the scores assigned by this grader has no impact the students' actual course grades.

Both graders graded the essays using the same set of seven predetermined grading criteria presented in Table 1; many of the criterions integrate Hansen's proficiencies listed above. Most economics instructors will agree that written assignments should be graded on both their substance and their presentation. Generally, the substance of the writing refers to how well the work addresses the question(s) posed. This includes the quantity and the quality of the information conveyed as well as the adequacy and the richness of the analysis. The presentation, on the other hand, refers to the organization and the flow of the thoughts and points, and the writing style and mechanics. For exposition purposes, the substance criteria are grouped under "summary" and "analysis," and the presentation criteria under "presentation." Both graders used the same scoring scale of 0 to 10 for each of the seven grading criteria. The class instructor did not provide a detailed rubric of grading criteria to the students since most economics instructors do not usually do this. The students were only informed that separate scores would be given for the summary and the analysis as well as the presentation.

Table 1 – Grading Criteria

	Summary
Criterion 1	Identification of main and secondary points.
Criterion 2	Understanding and accuracy of points.
	Analysis
Criterion 3	Identification of concepts.
Criterion 4	Understanding and accuracy of the concepts.
Criterion 5	Adequacy and richness of analysis; and demonstration of higher level thinking skills.
	Presentation
Criterion 6	Organization of paper (e.g. thoughtful organization; logical flow; and length).
Criterion 7	Writing style and mechanics (e.g. appropriate and effective style and tone; paragraph development; sentence structure; word use; grammar; punctuation; and spelling).

The essays were not blindly graded by either grader. As such, the scores given by the class instructor may be influenced by the instructor's assessment of the students' work and the instructor's personal bias (favorable or unfavorable) towards one or more of the students. Since the second grader has no personal knowledge of the students, we assume the scores assigned by this grader reflect only the grader's assessment of the written work and preclude any personal bias. For the purposes of this study, we define grading bias as the difference in the scores awarded by the two graders. That is, the grading bias measures the difference in the individual graders' subjective evaluations of the written assignments as well as the personal bias of the class instructor.

Methods of Analysis

Given its definition, grading bias is presumed to be present when the difference in the scores assigned by the two graders is statistically significantly. The simple *paired difference t-test* is used to detect this. The hypothesis is set up to determine whether the mean difference between the score awarded by the class instructor (CI) and score awarded by the other grader (OG) is equal to zero. We use the two-tailed test since there is no reason to assume *a priori* that the score given by one grader should be dissimilar from the score given by the other grader.

$$H_0 : \mu_{CI} - \mu_{OG} = 0$$

This paired difference t-test is employed seven times for each one of the seven grading criteria. Comparison of these results should reveal where the grading bias, if any, exists.

Since students are usually more concerned with the overall grade of an assignment than the breakdown of that grade, we will also examine grading bias from this perspective. That is, we also perform the paired difference t-test on the *Total Score* given by the two graders. We obtain $Total\ Score_{CI}$ by summing the scores assigned by the class instructor for each of the seven grading criterions. $Total\ Score_{OG}$ is calculated in a similar manner using scores given by the other grader. It should be noted that the same weight is applied to the grading criteria for illustrative purposes only. In practice, an instructor might utilize unequal weights to the various criterions according to their relative importance to the individual instructor. Regardless whether the weights are equal or not, because a *Total Score* is an aggregated score, it may augment the presence of small and insignificant biases if the biases are in the same direction or hide the presence of significant grading biases if the biases are in opposite directions.

We use *ordinary least squares* (OLS) analysis to determine the degree of variation in the difference in the overall scores that can be attributed to the grading criterions which mean differences are found to be statistically significantly different from zero.

$$(\Delta\ Total\ Score)_i = \alpha + \sum \beta_j\ (\Delta\ Criterion\ j)_i + \varepsilon_i$$

where for the i^{th} essay, $(\Delta\ Total\ Score)_i$ is the difference in the *Total Score* given by the CI and the OG; $(\Delta\ Criterion\ j)_i$ is the difference in the scores given by the CI and the OG for grading criterion j; and ε_i is the stochastic error term.

Empirical Results

Table 2 – Summary Statistics of the Scores Awarded by the Two Graders

	Class Instructor's Mean	Other Grader's Mean	Mean Difference
Criterion 1	7.98	7.83	0.15
Criterion 2	8.00	7.81	0.19
Criterion 3	7.60	7.21	0.39
Criterion 4	7.38	7.15	0.23
Criterion 5	6.49	6.58	– 0.09
Criterion 6	8.18	7.43	0.75
Criterion 7	8.03	7.56	0.46
Total Score	53.64	51.57	2.08

Note: Difference = Score awarded by class instructor – Score awarded by the other grader.

Table 2 presents the summary statistics for the scores awarded by the two individual graders and the differences in those scores. We find that the mean score assigned by the class instructor is higher than the mean score assigned by the other grader for every criterion except *Criterion 5* (the adequacy and the richness of the analysis). However, the mean difference in the scores is less than 1 (based on the scale of 0 to 10) in every instance. There was a high degree of correlation between the scores given by the graders; the correlation coefficients ranged from 0.7 to 0.8.

Table 3 – The p-values for the Two-Tailed Paired Difference T-Tests

Criterion 1	*Criterion 2*	*Criterion 3*	*Criterion 4*	*Criterion 5*	*Criterion 6*	*Criterion 7*	*Total Score*
0.306	0.221	0.012	0.212	0.613	0.000	0.004	0.011

The p-values for the paired difference t-tests are summarized in Table 3. At the 0.05 level, the mean difference between the scores awarded by the two graders is statistically significantly different from zero for *Criterion 3* (the analysis identifies and uses the appropriate economic concepts), *Criterion 6* (the organization of the essay), and *Criterion 7* (the writing style and mechanics). Only the last two are also significant at the 0.01 level. There is insufficient evidence to suggest that the two graders differ in how they evaluated the essays relative to *Criterion 1* (the summary identifies the points made in the articles), *Criterion 2* (the summary expresses the points made in the articles correctly), *Criterion 4* (the analysis applies the economic concepts correctly), and *Criterion 5* (the adequacy and the richness of the analysis). The mean for $Total\ Score_{CI}$ is statistically significantly higher than the mean for $Total\ Score_{OG}$ at the 0.05 level but not at the 0.01 level.

OLS results are shown in Table 4. The model uses only *Criterion 3, Criterion 6*, and *Criterion 7* as the regressors because their paired difference t-tests are significant at the 0.05 level. This regression is significant at the 0.01 level as demonstrated by the p-value of the F-statistic. The adjusted R^2 shows that when all the grading criteria are equally weighted, over 81 percent of the variation in the differences in the total scores assigned by the two graders is due to the differences in the abovementioned grading criterions. The sizes of the coefficient estimates suggest that the best predictors of the difference in the *Total Score* given by the two graders are the difference in the scores assigned for *Criterion 3* and *Criterion 7*.

Table 4 – OLS Results

Variable	Coefficient	p-value
Intercept	– 0.671	0.089
Criterion 3	2.692	0.000
Criterion 6	0.833	0.007
Criterion 7	2.332	0.000

Conclusions

This paper responds to the call for economics instructors and researchers in economic education to give more attention to how grades are determined (Walstad 2001). Specifically, the results of this case study provide some preliminary insights into potential grading bias when evaluating writing assignments in a Principles of Macroeconomics class. Recall that we define grading bias as the difference between the graders' evaluations of the written assignments as well as the personal bias of the class instructor. Given this definition, we identify small but significant grading biases in the evaluations of the essays with respect to identification of concepts in the student analyses, the organization of the essays, and the writing style and mechanics. Moreover, we find that the difference in the two graders' overall evaluation of an essay can mostly be attributed to the graders' individual assessments of whether that essay employs the appropriate concepts in its analysis and the essay's writing style and mechanics. Limitations of our dataset prevent analysis into the dynamics leading to these outcomes.

It is perhaps not surprising that the presence of grading bias is detected in only one of the five substance criterions. In fact, a possible explanation for the difference in how the graders scored identification of concepts may simply be the asymmetric information available to the two graders. The class instructor has perfect information on the economic concepts the students have been exposed to but the other grader does not, and consequently the latter grades the essays for most or all of the economic concepts in the assigned articles. If this is the case, this type of grading bias would not be important in most post-secondary economics class settings where the class instructor or a well instructed grader typically does most or all of the grading. We conclude that the potential for grading bias is small when a written assignment is being graded with respect to the type and the quality of its content.

On the other hand, we find the presence of grading bias in both of the presentation criterions. This is not to suggest that all or even most instructors will vary greatly on how they score the presentation aspects of the written assignments. What is clear, however, is that the scores awarded for these aspects of the writing assignments may be more vulnerable to the instructors' subjective preferences for specific organizational and writing styles. Moreover, while the intent of writing assignments is to allow students to express themselves and demonstrate economic knowledge and thinking, students do not always communicate their ideas clearly, making it difficult for the instructors who read the written work to assess how much and how well the students know and understand the economics they purportedly learned. Accordingly, we conclude that the potential for grading bias is greater when assessing the presentation aspects of the written assignments.

Given the findings above, we propose that economics instructors can actively manage the potential grading bias in the overall scores when grading writing assignments. Recall that we assumed equally weighted grading criterions when examining the grading bias in the overall score. Since the presence and the size of the grading bias vary from criterion to criterion, changing the weights of the criterions will clearly affect the size of the overall grading bias. If the potential for grading bias is a major concern, the instructor may choose to reduce the weights for the presentation criterions. The obvious drawback with this approach is it might reduce the incentives for students to utilize writing as learning tool to clarify their thoughts and improve on their critical thinking skills. An alternative solution is to decrease the weights of the early writing assignments compared to the weights of the later assignments (instead of the changing the criterion weights) when determining the course grades; this should work if students respond to how the instructor score the presentation aspects of their written work. Excluding the grades from the early writing assignments from the final grades is a viable solution only if students view optional work seriously. Clearly disagreements about the appropriate weighting scheme are likely. Each instructor must determine the best combination to reflect the instructor's instructional objectives for the assignment and the course, and consider them against the potential for grading bias.

Additional Issues

Additional issues were identified during our analysis but we are not able to address them due to data constraints. The first concern is separating the grading bias into its two components –the individual graders' subjective evaluations of the written assignments and the personal bias of the class instructor– and examining them individually; this would be possible if part of the sample is blindly graded by the class instructor. The second issue is whether the presence and the degree of grading bias would be affected if

students have specific knowledge about the grading criterions and their weights. In other words, would student awareness about the grading scheme affect the quality and the efforts of the student writings to the degree that the patterns of grading bias are altered? The third question is how to determine the weights of the individual grading criteria in the overall score to manage the potential for grading bias and to balance that with the instructional goals. A final area for future investigation would be identifying factors that can help explain why graders may differ in how they evaluate writing assignments, for example instructor characteristics such as demographic variables, years of teaching experience, years at current institution, etc. As with all case studies, we acknowledge that our results may lack universality and hope that others, using other small samples or larger samples of student writings and number of graders, will confirm or contest our results.

References

Becker, William E. (1982). "The educational process and student achievement given uncertainty in measurement." *American Economic Review* 72(1): 229-36.

----- and Michael Watts (1996). "Chalk and talk: A national survey on teaching undergraduate economics." *American Economic Review* 86(May): 448-53.

Hansen, W. Lee (1986). "What knowledge is most worth knowing – For economics majors?" *American Economic Review* 76(2): 149-52.

----- (1998). "Integrating the practice of writing into economics instruction." In *Teaching economics to undergraduates: Alternatives to chalk and talk*, edited by W.E. Becker and M. Watts. Cheltenham, U.K.: Edward Elgar.

----- (2001). "Expected proficiencies for undergraduate economics majors." *Journal of Economic Education* 32(3): 231-42.

Siegfried, J.J., P. Saunders, E. Stinar, and H. Zhang (1996). "How is introductory economics taught in America?" *Economic Inquiry* 34(1): 1182-92.

Walstad, William B. (2001). "Improving Assessment in University Economics." *Journal of Economic Education* 32(3): 281-94.