THE EFFECT OF SCRAMBLING TEST QUESTIONS ON STUDENT PERFORMANCE IN A SMALL CLASS SETTING

Della L. Sue[1]

ABSTRACT

A technique used by instructors is to prepare several versions of the same exam in which the multiple-choice questions appear in a different order in each version. This makes it difficult for a student to obtain answers from another student while keeping the level of difficulty of the exam constant across students since every version contains the same questions. If the order in which questions are arranged in an exam has an effect on a student's performance on the exam, then changing the sequence order may bias student performance. Previous statistical analyses of data collected from economics courses provide mixed results on whether scrambling the content order biases a student's test score. In this paper, I investigate the effect of scrambling test questions on student performance in principles of macroeconomics courses and principles of microeconomics courses that are characterized by small class size.

Key Words: pedagogy, testing, and student performance

JEL Classification: A22

Introduction

In an effort to reduce the benefits of cheating on an exam, a technique used by instructors is to prepare multiple versions of the same exam in which the multiple-choice questions appear in a different order in each test version. This makes it difficult for a student to obtain correct answers from another student while keeping the level of difficulty of the exam constant across all students since every version contains the same questions. Many computerized test banks offer question scrambling as a standard feature. This makes it easy for the instructor to prepare multiple versions of the same test questions and its standard availability highlights the popularity of the technique.

An assumption of this technique is that the level of difficulty of an exam is determined by the level of difficulty of the questions being asked in the exam and not the order in which the questions are asked. However, does content order affect a student's performance on an exam? In particular, if the content sequence of the questions on an exam is in the same order in which the material was covered in the course, then a student might perform better on the exam as a result of order association. This would suggest that content order matters and a randomly scrambled version of the same test could result in weaker student performance on the test.

To what extent does the order in which questions are arranged in an exam affect a student's performance? If there is an effect, then the degree of difficulty of the exam might not be constant across all students, even though the different versions of the exam contain the same questions. Although the intention of the instructor is to be fair to all

[1] Assistant Professor of Economics, Department of Economics, Accounting & Finance, School of Management, Marist College, 3399 North Road, Poughkeepsie, NY 12601-1387

students by reducing the benefits to cheating, scrambling the order of the questions might instead unfairly put some students to a grade disadvantage.

This paper empirically explores this issue with data collected from two introductory one-semester undergraduate courses, which were principles of macroeconomics and principles of microeconomics.

Literature Review

The question of student performance being affected by the content order of multiple-choice questions in an exam has been previously explored. The earliest studies were applied to social sciences other than economics, such as psychology and geography. Although the number of studies is not voluminous, I am limiting this literature review to studies that pertain to economics courses so as not to confound the issue with differences between disciplines.

Taub and Bell (1975) conducted one of the earliest studies that addressed the issue at hand. Using results from an undergraduate principles of economics class, they performed a regression analysis in which the sum of scores on previous exams and a dummy variable for an exam form in which the questions were randomly arranged were regressed on the final exam score. Their results indicated that the effect of the exam form on the exam score was statistically significant and that the students who took the form in which the questions were randomly arranged scored lower than their classmates whose exam contained the same questions arranged in content order.

However, two subsequent studies lead to opposite results. Gohmann and Spector (1989) found that scrambling questions did not adversely affect student performance in a principles of macroeconomics class. They suggest including a measure of student intelligence, such as grade point average or SAT score, as an additional factor that might influence student performance but were precluded from doing so by the unavailability of the data.

The study by Bresnock, Graves, and White (1989) similarly concluded that question order had no effect on test performance. However, using data from undergraduate principles of economics classes, their analysis also considered the effect of the distribution of the responses within a multiple-choice question and they found that altering the response pattern of answers could affect the degree of test difficulty as measured by the test scores. This suggests that if an instructor wants to offer different forms of an exam to minimize the benefits of student cheating without altering the difficulty of the exam, the instructor should scramble the test questions but not scramble the choices within each question.

Carlson and Ostrosky (1992) offer a contrasting suggestion in creating variations in an exam. Their study was based on data collected from a principles of microeconomics class. They look beyond the effect of question order on the mean (average) test score and address the possibility that question order affects the distribution of exam scores. While they found that scrambling the order of the questions could result in a lower test score, they recommend that the order of the responses within a multiple-choice question be scrambled instead of scrambling the questions as a method of reducing the benefits from cheating.

In a more recent study, Doerner and Calhoun (2009) found that the content order of the questions had a statistically significant effect on the exam score. Both sequentially ordered questions and reverse sequentially-ordered questions resulted in a higher grade,

on average, with the former ordering of questions having a larger impact. The data was collected from three courses. Two courses were introductory macroeconomics and the third course was introductory microeconomics.

Sue (2006) found that scrambling the content order of multiple-choice questions did not affect a student's grade. The conclusion was that offering different versions of an exam, in which each version contains the same questions, to reduce the gain from cheating does not bias student performance on the exam. Data used in the analysis was from a one-semester intermediate microeconomics course. An intermediate-level course was chosen because it was felt that students in this course were more uniformly interested in economics than are students in a principles course. All of the students had previously taken a principles of microeconomics course, most had taken a principles of macroeconomics course, and some students had taken an economics elective course. A few had previously taken an intermediate macroeconomics course.

Thus, in summary, empirical examinations of the effect of scrambling the content order of multiple-choice questions on a student's performance on the exam indicate mixed results. All of the courses used in these studies were reported to have been large lecture classes. There were 88 students in the study by Taub and Bell (1975), approximately 300 students in the class used by Bresnock, Graves, and White (1989), 191 students in the analysis by Gohmann and Spector (1989), and 400 students in the section used by Carlson and Ostrosky (1992). In the study by Doerner and Calhoun, all three courses had a large class format, with a maximum enrollment of 450 students each. In the analysis by Sue (2006), data was gathered from an intermediate level microeconomics course in which it was presumed that there is more uniformity in the students' interest in the subject but the class size was small (28 students at the beginning of the course) in comparison with the studies conducted with principles-level courses.

The focus of the analysis in this paper is to test the effect of scrambling the content order of multiple-choice questions on a student's performance on the exam at the principles-level of course material within a small class setting. The relevance of class size is expressed in a footnote by Bresnock, Graves, and White (1989): "The law of large numbers makes it unlikely that our results imply that one group of test takers is more able than the other. This might not be true for smaller classes, however" (page 244). One of the salient features of smaller class size is stronger interpersonal communication between the students and the instructor as well as between students. This has the potential to weaken an advantage of recalling information if the questions on an exam appear in the same content order as the material was covered in class.

Empirical Methodology

The principal inquiry is whether scrambling the content order of questions in a multiple-choice test introduces a bias in student performance on the test, consequently affecting the test grade. The relationship can be expressed as

$$\text{Grade} = f(\text{content order of questions})$$

The effect of the content order on the grade can be estimated with the following linear relationship

$$Y_{ij} = \beta_0 + \beta_1 X_{ij} + \varepsilon_{ij}$$

where Y_{ij} is the test grade for individual i on test j and X_{ij} represents the content order of questions. X_{ij} is a binary variable that indicates whether the multiple-choice questions are presented in the same order in which the material was covered in the course or whether the order of the questions was presented in a scrambled sequence. β_1 measures the effect of scrambling the order on the test grade.

If content order matters and a student benefits from order association when the questions are presented in the same order in which the material was covered in the course, then the expectation is that

$$\beta_1 < 0$$

and the bias from scrambling can be estimated by the magnitude of the coefficient.

The Data

Data was collected in two one-semester undergraduate courses. The semester in which the data was collected for the principles of macroeconomics course was the Spring 2006 semester. In the three class sections of the course that were sampled, the initial enrollment at the beginning of the semester was 24 students, 30 students, and 27 students, with a total enrollment of 81 students. In the following Fall 2006 semester, data was collected for 87 students who were initially enrolled in 3 class sections of the principles of microeconomics course, with individual section enrollments of 29 students, 30 students, and 28 students. In each class section, the enrollment was between 25 and 30 students, which is smaller than the class sizes of the previous analyses included in the literature review. All of the course sections used in this analysis were taught by the same instructor. For each course, the course material was identical across all sections.

Both courses are required of all students who are majoring in economics, business, or accounting, as well as those who minor in economics. Either course can also satisfy a distribution requirement in social sciences, as part of the liberal arts foundation of the college. As a result, although a large proportion of the students in the courses choose a major in economics, business, or accounting, there is a wide range of interests and other majors among the students. The implication is that students in this course are not uniformly interested in economics.

Most of the students take these courses early in their undergraduate studies, as either freshmen or sophomores. Unlike upperclassmen who become academically experienced, many of the students in these courses are learning to navigate through courses at the college level. Being novices, content order of course material could be an important aspect of a student's ability to understand, analyze, and recall the material, thus affecting a student's performance on an exam.

As part of the course requirements, the students were given two exams during the semester and a final exam at the end of the course. For each exam and the final, two versions of the exam were used, each containing the same multiple-choice questions: one form contained questions in the same content order in which the material was covered during the semester and the other form contained the same questions in a randomly scrambled order that was generated by the test bank software provided by the textbook publisher.

The exams were randomly distributed among the students at the beginning of the exam session. The distribution of exams by form ("Version") is given in Table 1 for each course. Panels A through C pertain to the principles of macroeconomics course and

Table 1. Distribution of Versions by Exam and Course

Principles of Macroeconomics

Panel A	Exam2			
Exam1	(blank)	Version 0	Version 1	Total
(blank)	3	2	3	8
Version 0	4	21	15	40
Version 1	1	15	17	33
Total	8	38	35	81

Panel B	Final		
Exam1	Version 0	Version 1	Total
(blank)	8		8
Version 0	23	17	40
Version 1	11	22	33
Total	42	39	81

Panel C	Final		
Exam2	Version 0	Version 1	Total
(blank)	5	3	8
Version 0	21	17	38
Version 1	16	19	35
Total	42	39	81

Principles of Microeconomics

Panel D	Exam2			
Exam1	(blank)	Version 0	Version 1	Total
(blank)	1	2	2	5
Version 0		16	25	41
Version 1	1	25	15	41
Total	2	43	42	87

Panel E	Final			
Exam1	(blank)	Version 0	Version 1	Total
(blank)	1	3	1	5
Version 0		29	12	41
Version 1	1	25	15	41
Total	2	57	28	87

Panel F	Final			
Exam2	(blank)	Version 0	Version 1	Total
(blank)	1	1		2
Version 0		30	13	43
Version 1	1	26	15	42
Total	2	57	28	87

Key: Version 0-on the exam, the multiple-choice questions were presented in the same content order in which the material was presented in the course.
Version 1-on the exam, the multiple-choice questions were presented in a randomly scrambled order.
blank-neither version of the exam was taken.

panels D through F pertain to the principles of microeconomics course. Each panel provides the distribution among students for each version between two of the three exams given in the course, as indicated, as well as the number of students who took a particular version of a specific exam. For example, in panel A, for the first exam in the principles of macroeconomics course, 40 students took a content-ordered exam ("Version 0") and 33 students took a content-scrambled exam ("Version 1"). Eight students took neither version ("blank"), generally because they missed the exam on the day it was administered and were given a makeup exam that contained different questions. For the second exam of the course, 38 students took a content-ordered version and 35 students took a content-scrambled exam. Similarly, there were eight students who were administered a different exam for Exam2. Twenty-one students took the content-ordered version of both Exam1 and Exam2, 17 students had scrambled versions for both Exam1 and Exam 2, 15 students had a scrambled exam for Exam1 but a content-ordered exam for Exam2, and 15 students experienced the opposite combination for the two in-class exams. In the remaining panels of Table 1, corresponding distributions are provided for the two other combinations of the exam pairs as well as for the other course.

A chi-square test of the joint distribution in each panel implies that there is no statistically significant relationship at a 95% level of confidence in the distribution of students who took the content-ordered version and those who took the scrambled version between each pair of exam combinations except for panel D. However, for Panel D, the chi-square test was not statistically significant at a 99% level of confidence. Thus, the exam versions were randomly distributed among the students at the beginning of each exam and there does not appear to be any systematic grouping among versions between exams.

The average score on the multiple-choice questions for each version for all three exams is presented in Table 2. Except for the second exam in the principles of macroeconomics course and the final exam in the principles of microeconomics course, the average score was lower for those who took the scrambled order version. However, a statistical test of the difference between two means for each exam implies that we cannot reject the null hypothesis that there is no difference between the average scores by version for all three exams in both courses.

Table 2. T-Test of the Difference Between Average Grade by Version of the Exam

Principles of Macroeconomics

Exam1	Version 0	Version 1
Average Grade	97.3750	93.4848
Number of observations	40	33
Hypothesized difference	0	
T Stat	1.1242	
P (T<=t) two-tail	0.2647	
T Critical two-tail	1.9939	

Principles of Microeconomics

Exam1	Version 0	Version 1
Average Grade	112.4390	110.7317
Number of observations	41	41
Hypothesized difference	0	
T Stat	0.4051	
P (T<=t) two-tail	0.6865	
T Critical two-tail	1.9901	

(continued)

Exam2	Version 0	Version 1
Average Grade	103.2895	106.2857
Number of observations	38	35
Hypothesized difference	0	
T Stat	-0.5837	
P (T<=t) two-tail	0.5612	
T Critical two-tail	1.9939	

Exam2	Version 0	Version 1
Average Grade	110.5814	106.3095
Number of observations	43	42
Hypothesized difference	0	
T Stat	1.0338	
P (T<=t) two-tail	0.3042	
T Critical two-tail	1.9890	

Final	Version 0	Version 1
Average Grade	152.1429	148.8462
Number of observations	42	39
Hypothesized difference	0	
T Stat	0.5435	
P (T<=t) two-tail	0.5883	
T Critical two-tail	1.9905	

Final	Version 0	Version 1
Average Grade	132.8947	133.0357
Number of observations	57	28
Hypothesized difference	0	
T Stat	-0.0211	
P (T<=t) two-tail	0.9832	
T Critical two-tail	1.9890	

Key: Version 0-on the exam, the multiple-choice questions were presented in the same content order in which the material was presented in the course.
Version 1-on the exam, the multiple-choice questions were presented in a randomly scrambled order
Note: Exam1 and Exam2 each contained 30 multiple-choice questions, each of which was worth 5 points. The final exam contained 40 multiple-choice questions, each of which was worth 5 points.
T-Test assumes equal variances.

To test whether the content order of multiple-choice questions affects a student's performance on an exam, regression analysis was performed. The null hypothesis is that scrambling the content order of questions in a multiple-choice test does not affect student performance on the test. Since there is no statistical difference between average scores by version for each of the three exams for both courses, the following specification can be estimated using the combined data across the three exams and both courses.

$$GRADE = b_0 + b_1\ VERSION + b_2\ DCOURSE + b_3\ DEXAM1 +$$
$$b_4\ DEXAM2 + b_5\ DCOURSEEXAM1 + b_6\ DCOURSEEXAM2 +$$
$$b_7\ DCOURSEVERSION + b_8\ DEXAM1VERSION +$$
$$b_9\ DEXAM2VERSION + b_{10}\ DCOURSEEXAM1VERSION +$$
$$b_{11}\ DCOURSEEXAM2VERSION$$

The dependent variable was the multiple-choice score ("GRADE"). A dummy variable ("VERSION"), which represented the content version of the exam that was taken by the student, was created as an explanatory variable. This variable assumed a value of 1 if the content order of the multiple-choice questions was the scrambled version and a 0 if the content order followed the sequence in which the material was presented in class.
In order to allow for structural variation between the exams and courses, dummy variables were added to the regression specification. DCOURSE had a value of 1 if the observation was from the principles of microeconomics course and a value of 0 if the observation was from the principles of macroeconomics course. Observations from the first exam and the second exam had a value of 1 for DEXAM1 and DEXAM2, respectively, and a value of 0 otherwise. If both of those binary variables had a 0 value,

the observation pertained to the final exam. The remaining binary variables are interaction terms between the course, exam, and content order version.

The results of the regression are presented in Table 3. The coefficients for VERSION and all of the interaction terms with VERSION are not statistically significant. On the other hand, the coefficients for the binary variables representing the course and exam are statistically significant. There are structural differences between the two courses and the three exams, but the null hypothesis that scrambling the content order of questions in a multiple-choice test does not affect student performance on the test is not rejected.

Table 3. OLS Regression Results

	Coefficients	T Stat
Intercept	152.1429	43.7548
VERSION	-3.2967	-0.6579
DCOURSE	-19.2481	-4.2003
DEXAM1	-54.7679	-11.0008
DEXAM2	-48.8534	-9.6831
DCOURSEEXAM1	34.3121	5.0546
DCOURSEEXAM2	26.5400	3.9058
DCOURSEVERSION	3.4377	0.4760
DEXAM1VERSION	-0.5934	-0.0814
DEXAM2VERSION	6.2929	0.8645
DCOURSEEXAM1VERSION	-1.2548	-0.1225
DCOURSEEXAM2VERSION	-10.7058	-1.0501
R Square	0.4124	
Number of Observations	479	

Educators presume that a student's performance on an exam is an indicator of how well the student comprehends the material. A student's comprehension of the material reflects, in part, the effort put forth by the student to learn the material, the student's innate learning ability, random chance, and perhaps the content order of the questions on the exam. Gohmann and Spector (1989) discussed the inclusion of grade point average and SAT scores as indicators of student intelligence. Although the unavailability of this information precluded them from including it in their analysis, they speculated that their finding that content order has little effect on performance would not be changed with the inclusion of grade point average or SAT scores.

In Sue (2006), the student's grade point average was included in the analysis as a measure of ability. The results indicated a statistically significant correlation between grade point average and exam score, with no change in the finding that scrambling the order of multiple-choice questions does not adversely affect a student's performance on an exam. In that study, the data was obtained from an intermediate microeconomics course. None of the students were freshmen so a grade point average was available for each student in the course. However, in the current analysis many of the students were first semester freshmen for which a grade point average was not yet available at the time of the course. Alternatively, an analysis based on the final exam can be conducted in which the student's performances on the first exam and second exam in the course are included in the regression as controls for the student's ability and effort.

$$GRADE = b_0 + b_1\ VERSION + b_2\ MC1 + b_3\ MC2 + b_4\ DCOURSE +$$
$$b_5\ DCOURSEVERSION + b_6\ DCOURSEMC1 + b_7\ DCOURSEMC2$$

"MC1" is the exam score on the multiple-choice section of the first exam and "MC2" is the exam score on the multiple-choice section of the second exam.[2] The assumption is that there is less variation in a student's ability and effort between exams than there is between students. Thus including a student's performance on prior exams within the same course could control for innate characteristics of the student. The dummy variable for course and the corresponding interaction terms were included to allow for structural variation.

The results are presented in Table 4. Recall that the data pertains to the final exam for both courses. The coefficients on VERSION and DCOURSEVERSION are not statistically significant, again inferring that scrambling the content order of the questions does not adversely affect student performance on the exam. The coefficients on previous exam scores are statistically significant, as is the coefficient on DCOURSE. The coefficient on DCOURSE is negative, implying that performance on the final exam in the principles of microeconomics course was weaker than in the principles of macroeconomics course, controlling for test version and performance on the first two exams. The coefficients on the interaction terms between course and performance on prior exams are not statistically significant. One explanation is that the negative effect of the course subject is cancelled by the positive effect of performance on prior exams.

Table 4. OLS Regression of GRADE on VERSION and Prior Exam Grades

	Coefficients	T Stat
Intercept	50.3550	2.9365
VERSION	-3.3355	-0.6904
MC1	0.6517	3.7617
MC2	0.3891	3.3881
DCOURSE	-53.4003	-2.2130
DCOURSEVERSION	7.5193	1.0706
DCOURSEMC1	-0.0164	-0.0719
DCOURSEMC2	0.1994	0.9825
R Square	0.4572	
Number of Observations	165	

Conclusions

This analysis addresses the question of whether the content order of multiple-choice questions affects student performance on an exam. Based on the empirical results obtained in two one-semester undergraduate courses in principles of macroeconomics and principles of microeconomics, exam scores do not appear to be affected by the order in which the questions are presented in the exam. The results do not change when controls

[2] In an alternative specification, "MC1" and "MC2" were replaced with "MCSUM", which is the sum of the exam scores on the multiple-choice sections of the first exam and the second exam. Corresponding interaction terms were also included in the regression. The results, which were consistent with the results presented here, are available from the author on request.

for student effort and ability are introduced into the analysis. This suggests that the technique of scrambling multiple-choice questions in order to reduce the benefits of student cheating during the exam can be done without risk of biasing student performance. The courses used in this study were conducted in a small class format, rather than in a large lecture style design that characterized the introductory level economics courses used in previous studies. The evidence suggests that any potential disadvantages from taking an exam in which the content order of the questions deviates from the order in which the material was taught is not apparent in a small class setting.

References

Bresnock, Anne E., Philip E. Graves, and Nancy White. 1989. "Multiple-Choice Testing: Question and Response Position." *Journal of Economic Education*, Summer.

Carlson, J. Lon, and Anthony L. Ostrosky. 1992. "Item Sequence and Student Performance on Multiple-Choice Exams: Further Evidence." *Journal of Economic Education*, Summer.

Doerner, William M., and Joseph P. Calhoun. 2009. "The Impact of the Order of Test Questions in Introductory Economics" (working paper). *Economics Educator: Courses, Cases & Teaching*, April 20, Economics Research Network (ERN), a division of Social Science Research Network (SSRN).

Gohmann, Stephan F., and Lee C. Spector. 1989. "Test Scrambling and Student Performance." *Journal of Economic Education*, Summer.

Sue, Della Lee. 2006."The Effect of Test Scrambling on Student Performance." *Proceedings of the 2006 NBEA Annual Conference*, Northeast Business and Economics Association.

Taub, Allan J., and Edward B. Bell. 1975. "A Bias in Scores on Multiple-Form Exams." *Journal of Economic Education*, Fall.