

Journal of Humanistic Mathematics

Volume 3 | Issue 1

January 2013

Changing the Order of Mathematics Test Items: Helping or Hindering Student Performance?


Kristin T. Kennedy

Bryant University, kkennedy@bryant.edu

Allison G. Butler

Bryant University, abutler@bryant.edu

Follow this and additional works at: <http://scholarship.claremont.edu/jhm>

 Part of the [Educational Assessment, Evaluation, and Research Commons](#), [Educational Psychology Commons](#), and the [Science and Mathematics Education Commons](#)

Recommended Citation

Kennedy, K. T. and Butler, A. G. "Changing the Order of Mathematics Test Items: Helping or Hindering Student Performance?," *Journal of Humanistic Mathematics*, Volume 3 Issue 1 (January 2013), pages 20-32. DOI: 10.5642/jhummath.201301.04 . Available at: <http://scholarship.claremont.edu/jhm/vol3/iss1/4>

©2013 by the authors. This work is licensed under a Creative Commons License.

JHM is an open access bi-annual journal sponsored by the Claremont Center for the Mathematical

Sciences and published by the Claremont Colleges Library | ISSN 2159-8118 | <http://scholarship.claremont.edu/jhm/>

Changing the Order of Mathematics Test Items: Helping or Hindering Student Performance?

Cover Page Footnote

Acknowledgements: The authors would like to thank Professor Alan Olinksy of Bryant University for helpful thoughts and ideas on statistical analysis.

Changing the Order of Mathematics Test Items: Helping or Hindering Student Performance?

Kristin Kennedy

Mathematics Department, Bryant University, Smithfield, RI 02917

kkennedy@bryant.edu

Allison G. Butler

Applied Psychology Department, Bryant University, Smithfield, RI 02917

abutler@bryant.edu

Abstract

This paper recounts an experiment by a mathematics professor who primarily teaches mathematics majors. The main question explored is whether the ordering of the questions makes a difference as to how students perform in a test. More specifically we focus here on the following research questions: (1) *Does arranging a math test with easy-to-hard items versus hard-to-easy items impact student performance?* and (2) *If so, does item order impact male and female mathematics majors and non-majors in unique ways?* We examine data collected over multiple semesters with several different classes. We find that for most of the mathematics students who were examined, the ordering of the questions on a test did not impact performance. However, female majors performed better on classroom exams when the test was arranged with the more difficult questions presented first. Readers who are interested in teaching mathematics, educational psychology, or gender issues in the classroom may find our results intriguing.

Evolution of a research project. A mathematics professor, I (Kristin Kennedy) often give two versions of the same exam to reduce the possibility of students copying from one another. Recently, I began to wonder if the order of the questions could affect a student's grade. I approached Allison Butler, professor of psychology, for her expertise. I of course wanted to ensure that both tests were fair, and Professor Butler was interested to see if there was any difference from a psychological point of view. This paper describes our joint inquiry into this matter.

1. Introduction

Constructing a fair mathematics exam that truly measures student comprehension of the material, from surface level information to a deeper understanding of the concepts, is no easy matter, and anyone teaching mathematics has struggled with writing a fair exam. Today, there is a second problem in some university classrooms when administering a test: overcrowding. Students often sit quite closely to each other at long tables rather than individual desks. To minimize the possibility of copying, a popular way of administering a classroom exam is to make up multiple versions of the same test.

The challenge of writing clear and fair problems for each version of the same test is not insurmountable, but achieving parallel forms reliability is often not an easy task. Preferable to having two potentially nonequivalent versions of a particular test would be to design just one test, but to alter the sequence of the questions in order to reduce the likelihood of cheating in a densely packed classroom. Intuitively, one might surmise that an exam which begins with easy questions and progresses to more difficult ones would be easier for students. Or perhaps one might say it would be preferable to students. Acknowledging the diversity of students that fill most college classrooms today, we think another important consideration is the possibility that the arrangement of mathematics test items may impact the performance of certain student groups in unique ways.

Previous work shows mixed results as to whether test item arrangement can affect grade outcome. Most of the previous research has been concerned with the issue of test anxiety and whether item arrangement can impact a grade when test anxiety is present. Also previous research mainly focuses on students who are not necessarily inclined to quantitative subjects. Our project adds to this body of research because the participants we study are primarily mathematics majors with very strong mathematics backgrounds.

We take an interdisciplinary perspective, drawing from research in social psychology and mathematics education, to examine how humanistic aspects of test-taking may intersect with item arrangement to influence performance. Our specific research questions are: (1) *Does arranging a math test with easy-to-hard items versus hard-to-easy items impact student performance?* and (2) *If so, does item order impact male and female mathematics majors and non-majors in unique ways?* We use data collected from several classes (for majors and non-majors) over a number of semesters.

2. Literature Review

Generally, the literature suggests that the order of how questions are presented on an exam does not have an effect on the grade outcome. According to an extensive historical review of the literature compiled in 1985 by Leary and Dorans [3], researchers began to look at the effects of item placement on exams as early as 1950. This early research was concerned primarily with the effect of item order on performance. As more studies were done, results were mixed, depending on whether one is trying to examine if test anxiety is a factor or not, or if “speeded” test results are different from those tests that have unlimited time restrictions. Studies also show that a random rearrangement of items on a test does not seem to affect examinee performance. This is important if a professor simply wants to scramble test questions as a hedge against students copying from one another during a crowded exam. Leary and Dorans summarize that easy-to-hard or hard-to-easy sequencing designs do not support clear-cut conclusions. Some studies found evidence of effect, and others did not.

Plake, Thompson, and Lowry [6] examined the effects of item arrangement, knowledge of arrangement, and test anxiety on grade outcomes on a mathematics test. There were three arrangements: easy-to-hard, hard-to-easy, and random question arrangement. No significant results were found regarding the change of order, but the participants in this experiment were volunteers from a general psychology class, not mathematics majors nor students overly motivated to do well on a mathematics exam. Plake and Lowry continued their work along with Ansorge and Parker [7] and found that item arrangement was significant with motivated upper division students under “speeded” conditions. In particular they found that males, who were given an easy-to-hard ordering, performed the best. However, the authors did note that the strict time limit to the test may have added to the significant outcome that an easy-to-hard ordering was helpful.

Tippets and Benson [11] also examined the effects of item arrangement and test anxiety and suggested that item arrangement on an exam can affect test anxiety levels. This study did not directly conclude that performance was enhanced with an easy-to-hard progression of test questioning, and the participants in the study were not mathematics majors. There is simply the suggestion that if test anxiety is low, then a student would have a better outcome on an exam. Plake *et al.* [7] did conclude that males taking strictly

timed exams performed better and in fact out-performed women when items were arranged from easy-to-hard. However, the population tested was comprised of 170 students from three introductory statistics classes. Two were educational psychology classes and one was an agricultural class. The mathematics background of the subjects varied substantially. Our paper focuses on the variable of gender and examines the performance of students who are primarily mathematics majors.

While there is limited research on the differential impact of item arrangement on males versus females, the social psychology literature on gender and stereotype threat suggests that males and females might respond differently when faced with challenging mathematics problems at the start of an exam. A stereotype is a generalized, often oversimplified or inaccurate belief about a group of people [4]. According to stereotype threat theory, performing in a domain in which one is negatively stereotyped creates anxiety and discomfort [10, 12]. These feelings of uncertainty are rooted in the fear that one's performance might confirm a negative stereotype. Research has shown that stereotype threat often impairs performance [12]. Specifically, studies detail how negative stereotypes about females' math abilities shape their math attitudes and undermine their performance in STEM fields [2, 8]. Even though females outnumber males by a considerable margin in terms of postsecondary education enrollment, far fewer females pursue studies and careers in math-intensive fields such as engineering and computer science [5].

Interestingly, research on stereotype threat theory has also revealed an opposite effect, whereby positive stereotypes may boost performance for certain individuals [9]. While our study does not intentionally evoke stereotype threat, the social psychological literature suggests that arranging math test items from hard-to-easy versus easy-to-hard could potentially impact the performance of male and female majors and non-majors in unique ways.

3. Method

We recorded data over a span of five semesters, starting with the spring of 2010 and ending in the fall semester of 2011; summer sessions were not included. Kennedy generated every test administered; test banks were not used. The tests were either constructed with easiest questions first and hardest questions listed last (Type A), or the other way around with the hardest questions first on the test and the easiest questions placed at the end of the

test (Type B), but each test had the exact same questions. Easy items were defined (by Kennedy) as those that involved one-step calculations without the application of the concept in a word problem. Also easy questions were those that required simple factual recall. Hard questions were defined as those that required multiple steps, especially those questions embedded in word problems and applications. We handed out tests randomly; only taking care to assure that students sitting very close to each other would have different tests to keep dishonesty to a minimum.

Four distinct mathematics classes were involved with the experiment, see Table 1. Three classes were for mathematics majors and one was for non-majors.

Table 1: General class information

Course (class size)	Male Percent	Female Percent	Major Percent
<i>Statistics II</i> (23)	57%	43%	9%
<i>Linear Algebra</i> (28)	54%	46%	100%
<i>Life Contingency</i> (80)	58%	42%	100%
<i>Theory of Interest</i> (42)	50%	50%	90%

Theory of Interest, a sophomore/junior level course in the theory of financial mathematics, and *Life Contingency*, a senior level course that is a blend of risk and interest theory, are two courses that students take if they are majoring in mathematics for actuarial science. The recorded information included the class name, the semester, the year, the test number (first test, second test, etc.), the each exam grade of the student, gender, major versus non-major, and whether the test was Type A or Type B. There were 594 records of data, each one containing the above list of data fields.

4. Analysis and Results

Whether the ordering of questions on a classroom mathematics test, easy-to-hard (Type A) or hard-to-easy (Type B), would show a significant difference to test grades was our main question of concern, and we were particularly interested in how male majors compared to female majors, even though we also had some data regarding non-majors. The full data set with grades used as the analysis variable followed a normal distribution, as did various subsets of the data. Subsets that were individually examined for normality were: 1) data for all males, 2) data for all females, 3) data for all the tests that had

questions of Type A, 4) data for all tests that had questions of Type B, 5) data for all tests for majors, and 6) data for all non-majors. All subsets were relatively normal in design.

Averages were computed for the different subsets.

Table 2: Average grades and standard deviations for various subsets of the data

	Observations	Mean	Standard deviation
<i>Males</i>	325	81.354	13.674
<i>Females</i>	269	82.048	13.764
<i>Easy-to-Hard (A)</i>	305	80.230	15.280
<i>Hard-to-Easy (B)</i>	289	83.187	11.660
<i>Majors</i>	462	83.409	12.570
<i>Non-majors</i>	132	75.580	15.680

All analyses were performed using SAS; the grade of each exam was the analysis variable. First we ran a linear regression analysis to examine if any of the independent variables in question were significant. The variable *Grades* was the analysis variable, and the independent variables in question were *Gender*, *Major* or *Non-major*, and whether the tests were *Type A* or *Type B*.

The regression equation was:

$$\text{Grades} = 81.6 + 1.27 \text{Gender} - 6.47 \text{Major/Non-major} + 2.4 \text{Type A/Type B}$$

Table 3: *P*-values for each of the coefficients in the regression equation

Variable	Constant	Gender	Major/Non-major	Type
<i>p</i> -value	0.0	0.225	0.0	0.022

Gender is the only non-significant independent variable at the 0.05 level of significance. The independent variable *Type* is significant, but the test score only changes by 2.4 points, depending on whether Type A or Type B test was administered.

Since *Gender* was non-significant, the regression was run again after removing the *Gender* variable, and the new regression equation was:

$$\text{Grades} = 81.99 - 7.79 \text{Major/Non-major} + 2.89 \text{Type A/Type B}$$

Table 4: *P*-values for the coefficients of the regression equation when *Gender* is removed

Variable	Constant	Major/Non-major	Type
<i>p</i> -value	0.0001	0.0001	0.0081

The new equation shows both variables to be significant, and they should remain. Initially then, *Gender* does not appear to be a significant variable for this data set. Also note that the coefficients to the independent variables did not change dramatically when we ran the regression analysis for the second time.

Next we ran a series of *t*-tests comparing two means at a time, combining all classes together. The series of *t*-tests were:

1. *Selecting out all the majors, is there a difference between the average grade between males and females?* There was no statistical significance with a *p*-value of 0.117.
2. *Selecting out all the non-majors, is there a difference between the average grade between males and females?* There was no statistical significance with a *p*-value of 0.122.
3. *Selecting out all male majors, is there a difference between grades of Type A or Type B?* There was no statistical significance with a *p*-value of 0.414.
4. *Selecting out all female majors, is there a difference between the average grades from Type A or Type B?* **These results were statistically significant** with a *p*-value of 0.027. Thus these averages are not equal, and a second test was performed on this subset of data.
5. Using the same group, results showed that with all female majors, the average grade from easy-to-hard questioning **was significantly lower** than the average grade from hard to easy with a *p*-value of 0.013. This was a surprising result! This group of all female majors performed better with harder questions starting off first on exams!
6. *Selecting out all male non-majors, is there a difference between the average grades from Type A to Type B?* There was no statistical significance with a *p*-value of 0.202.
7. *Selecting out all female non-majors, is there a difference between the average grades from Type A to Type B?* There was no statistical significance with a *p*-value of 0.718.

Thus our data show that there is no difference between the average grades of males and females or examining whether they are majors or non-majors, as noted by tests (1) and (2). Secondly, the results show that there is no difference between the average grades of (a) male majors, (b) male non-majors, and (c) female non-majors when examining if easy to hard questioning or

hard to easy questioning makes a difference. This is noted by tests (3), (6), and (7).

However, there was one surprising result! For the subset of all female majors, the average grade on tests of Type A was **less than or equal to** the average grade on tests of Type B, with a p -value of 0.013. That is, for this particular group of female majors, the grades were higher if the ordering of the test questions ranged from hardest first to easiest last. This was the only group that stood out with significant results.

Finally, we looked at course-specific findings, rather than just overall findings. Would a course-by-course examination give different results?

The course *Statistics II* was primarily a course for non-majors. A two-tailed t -test was used to examine if the average grades from Type A or Type B were equal for male non-majors. There was no statistical significance with a p -value of 0.55. A similar test for female non-majors resulted with a p -value of 0.73, and the average grades for both groups would be considered equal.

We were actually more interested in courses that mathematics majors take—namely *Linear Algebra*, *Life Contingency*, and *Theory of Interest*—since in that group a significant result was previously found. Table 5 below shows the p -values for each class, with separate columns for males and females. Each test was a two-tailed t -test, comparing grade averages from Type A and Type B exam to see if the grades are equal or not.

Table 5: P -values for six different t -tests comparing Type A = Type B exams

Class	Male majors	Female majors
<i>Linear Algebra</i>	0.33	0.20
<i>Life Contingency</i>	0.88	0.17
<i>Theory of Interest</i>	0.45	0.21

We see that for the male majors, there is no statistical difference in average grades from Type A to Type B test for any of the three courses. It also can be seen that the p -values are now higher for females in each class, suggesting that there is no statistical difference. However, this was a two-tailed test, and we can further examine one-tail tests for the female majors, which was of interest based on earlier results.

Table 6 displays the results of one-tail t -tests for female majors in each major course, comparing Type A to Type B tests for each class. The null hypothesis was that the average score for Type A test was greater than or

equal to that of Type B test.

Table 6: P -values for female majors of each class comparing type A \geq type B

Class	p-values	t statistic
<i>Linear Algebra</i>	0.09	-1.31
<i>Life Contingency</i>	0.086	-1.38
<i>Theory of Interest</i>	0.108	-1.26

The p -values found when just analyzing one course at a time are not highly statistically significant, as they were in the earlier overall test (p -value = 0.013). However, we see that the trend is present with the negative t statistics. Although not highly statistically significant, the sample size was much smaller for the analysis that was done for each course. The change in sample size can affect the significance. In test (5) above, with a p -value of 0.013, the sample size of female majors who had Type A was $n = 112$, and the sample size for Test B was $n = 101$. However, when each class was examined separately, the sample size in each test was dramatically smaller. For *Linear Algebra*, Type A exam had $n = 26$ and Type B had $n = 24$. For *Life Contingency*, Type A exam had $n = 57$ and Type B had $n = 45$. For *Theory of Interest*, Type A had $n = 25$ and Type B had $n = 29$. When sample sizes are smaller, a larger difference needs to be present to have a statistically significant result. But the trend can still be seen that the female majors did better on exams of Type B. That was not true of the male majors.

5. Conclusion

It is plausible for professors to think that in the construction of a math test, the ordering of the questions could be a factor in the grade outcome, and in fact it is reasonable to think that starting off with easier questions first and progressing to the more difficult questions would be a sensible construction. We were surprised with these results which show that for most of the mathematics students who were examined, the ordering of the questions on a test did not impact performance. Frankly, we were pleased to see that the ordering did not impact the grade. Although two versions of the same test have been administered for several semesters, the tests can be considered equivalent.

Interestingly, this work strongly indicates (p -value=0.027) that the subgroup of female majors showed a significant difference with the average grade

on the two types of tests, and this group scored significantly better (p -values=0.013) if the tests were constructed with the harder questions first. Further testing with future classes could be conducted to verify these results or to contradict them as an aberration inherent in this particular data set. As mentioned in the literature review, the results appear to contradict earlier findings by Plake *et al.* in [7]. They concluded that males performed better and in fact out-performed women with items arranged from easy to hard questioning. However, their participants were not mathematics majors.

After an overall analysis, the data were examined course by course. The results for the female majors performing better on Type B exams was less significant, but the results were trending toward the overall findings. With smaller sample sizes, such as the sizes used for the course by course analysis, rather than the larger samples sizes with the overall analysis, larger differences need to be present to show significance. However, the negative t -statistic values illustrated that the trend for female majors performing better on Type B exams was still prevalent.

We wondered why the female majors would show this result, yet the male majors did not. One possible explanation for this finding is that female majors experienced “stereotype boost” instead of stereotype threat when they encountered the challenging math problems at the very beginning of the exam. Our results connect with the work of Crisp, Bache, and Maitner [1] who found that females who successfully entered a gender counter-stereotypic quantitative domain (i.e., engineering majors) showed enhanced performance in a testing situation intended to evoke stereotype threat. However, psychology majors experienced reduced performance in the wake of stereotype threat. The implication is that female students who strongly identify with mathematics may rise to the challenge in an anxiety-inducing or stressful math testing situation. In fact, the female mathematics majors in the present study also showed higher average test scores when hard math problems were presented before easy problems.

While our study did not directly test stereotype threat, it is possible that encountering the difficult problems at the very beginning of the math test created a psychological state that lead female math majors to rise to the challenge in the same way that stereotype threat motivated the strong performance of female engineering majors in the Crisp *et al.* study [1].

We should perhaps note once again that all exams were created by Pro-

fessor Kennedy. She wrote all test questions by herself and decided which questions were “easy” and which were “hard”. That could be a point of bias in the experiment, since she alone decided which questions formed the Type A or Type B test. Difficulty was generally defined according to the complexity of the thinking skills that were necessary to complete the problem. However this is not atypical. Many professors write their own test questions, as opposed to using test bank questions. All in all, it is reassuring to see that the ordering of questions for most students does not have an effect on the grade outcome.

Acknowledgments: We would like to thank Professor Alan Olinsky for encouragement and consultation about our statistical results.

About the authors:

Kristin T. Kennedy, a professor of Mathematics at Bryant University in Smithfield, RI for over 30 years, currently serves as the Chair of the department. She graduated from Manhattanville College with a B.A. in Mathematics. She has completed two master’s degrees: an MST in Mathematics from Georgia Southern University and an MS in Computer Science from Brown University. Her Ph.D. is in Applied Science from the University of Rhode Island. Her research interests follow a wide spectrum from topics regarding the teaching of mathematics effectively to issues in accounting, healthcare, and a variety of business topics. Most recently her teaching interests are in the field of statistics and actuarial mathematics courses. Besides working and volunteering in the community, she enjoys traveling as often as possible, and she is an avid golfer.

Allison G. Butler is an Assistant Professor of Applied Psychology at Bryant University. She graduated from the College of William and Mary with a B.S. in Psychology and was a fifth grade teacher before earning her M.Ed. in Educational Psychology from the University of Virginia. Her Ph.D. is in Applied Developmental and Educational Psychology from Boston College. Her research and publications focus on children’s learning and cognition, contemporary educational policy, and higher educational pedagogy aimed at maximizing student learning and engagement. She currently teaches courses in educational psychology, testing and assessment, introductory psychology, and child development.

References

- [1] Crisp, R. J., Bache, L. M., and Maitner, A. T. (2009). "Dynamics of social comparison in counter-stereotypic domains: Stereotype boost, not stereotype threat, for women engineering majors." *Social Influence*, 4(3), pages 171–184.
- [2] Gunderson, E. A., Ramirez, G., Levine, S. C., and Beilock, S. L. (2012). "The role of parents and teachers in the development of gender-related math attitudes." *Sex Roles*, 66(3–4), pages 153–166.
- [3] Leary, L., and Dorans, N. (1985). "Implications for altering the context in which test items appear: A historical perspective on an immediate concern." *Review of Educational Research*, 55(3), pages 387–413.
- [4] Myers, D. G. (2013). *Psychology*, 10th Ed. New York: Worth Publishers.
- [5] National Science Foundation. (2012). *Women, minorities, and persons with disabilities in science and engineering*. Retrieved from <http://www.nsf.gov/statistics/wmpd/sex.cfm>, accessed January 6, 2013.
- [6] Plake, B.S., Thompson, P.A., and Lowry, S.R. (1981). "Effects of item arrangement, knowledge of arrangement and test anxiety on two scoring methods." *Journal of Experimental Education*, 49, pages 214–219.
- [7] Plake, B.S., Ansorge, C.J., Parker, C.S., and Lowry, S.R. (Spring, 1982). "Effects of item arrangement, knowledge of arrangement test anxiety and sex on test performance." *Journal of Educational Measurement*, 19(1), pages 49–57.
- [8] Shapiro, J. R., and Williams, A. M. (2012). "The role of stereotype threats in undermining girls' and women's performance and interest in STEM fields." *Sex Roles*, 66(3–4), pages 175–183.
- [9] Smith, J. L., and Johnson, C. S., (2006). "A stereotype boost or choking under pressure? Positive gender stereotypes and men who are low in domain identification." *Basic and Applied Social Psychology*, 28(1), pages 51–63.
- [10] Steele, C. M., and Aronson, J. (1995). "Stereotype threat and the intellectual test performance of African-Americans." *Journal of Personality and Social Psychology*, 69, pages 797–811.

- [11] Tibbets, E., and Benson, J. (1989). "The effect of item arrangement on test anxiety." *Applied Measurement in Education*, 2(4), pages 289–296.
- [12] Vick, S. B., Seery, M. D., Blascovich, J., and Weisbuch, M. (2008). "The effect of gender stereotype activation on challenge and threat motivational states." *Journal of Experimental Social Psychology*, 44, pages 624–630.