Multiple Choice Randomization

Ian McLeod, Ying Zhang and Hao Yu
University of Western Ontario

Key Words: Academic Integrity and Cheating; Multiple-choice examinations; Multiple-choice items and answer-options arrangement; Student evaluation; Teaching large classes

Abstract

Multiple-choice randomized (MCR) examinations in which the order of the items or questions as well as the order of the possible responses is randomized independently for every student are discussed. This type of design greatly reduces the possibility of cheating and has no serious drawbacks. We briefly describe how these exams can be conveniently produced and marked. We report on an experiment we conducted to examine the possible effect of such MCR randomization on student performance and conclude that no adverse effect was detected even in a quite large sample.

1. Introduction and Summary

The multiple-choice test is widely used in all school subjects and at all educational levels for measuring a variety of teaching objectives. Many jurisdictions now require standardized testing in order to graduate from high school and these tests frequently have a multiple-choice component. Multiple-choice examinations are used to evaluate student progress in many undergraduate courses in Statistics as well.

To overcome cheating, instructors often prepare several versions of these exams. In spite of this, students still may have the opportunity to cheat if they are able to observe a student nearby with the same exam. With the advent of economical digital photocopiers, it is now very easy to produce multiple-choice examinations in which the order of the questions as well as the order of the answers is scrambled. We used the Perl scripting language and developed scripts for performing the randomization as well as the marking of these exams. MCR could also be implemented on various other platforms such as VBA with Microsoft® Word or with *Mathematica* notebooks.

The first step is to produce the MCR exams. Inserting some simple markups in the document file that contains the examination questions and running a script we developed, produces as many MCR exams as required. Our Perl scripts were used with source files in RTF or LaTeX format. In the future we plan to use HTML as well. Each exam has the questions and its possible responses randomized. A three digit Exam Code uniquely identifies each exam and is associated with a keyfile that indicates the exact randomization that was used for that particular exam. After the exams are produced they may be put on a CD and taken to a digital photocopier to be printed, collated and stapled.

The second step is to mark these exams. During the examination, the students are required to indicate the Exam Code on the scantron answer sheet. For small classes, a Perl script can be used to produce a listing of the correct responses for each Exam Code and the exams can be marked manually. For larger classes, an optical reader is used to read in the student scantron sheets and produce a grade report and this was the method that we used for all of our exams. Students are encouraged to keep a copy of their responses, so they can later verify the correctness of the marking. We never had any problem either with the optical scanner or our marking scripts.

The resulting Grade Report indicates for each Exam Code the student's score and in addition, it is helpful to show for each exam what the correct answer is and which answer the student selected. An example Grade Report is available from our Statistics Laboratory Homepage.

For our own further analysis of the examination questions, we also produce a Response Analysis. For this purpose, we select the original ordering of questions as the exam corresponding to Exam Code 000. With respect to this ordering of questions, our script also produces various other statistical summaries. For each question how many students selected each possible answer as well as the proportion of students correctly answering this question are tabulated. We have also found it very helpful to compute for each question the correlation coefficient between an indicator variable defined as 1 if the student answered correctly and 0 otherwise and the exam score for that student. A low correlation suggests a poor question. A good discriminating question has a low or moderate proportion of students answering correctly but a high correlation. A non-parametric correlation coefficient such as Kendall's tau or Spearman's rank correlation could be used but our preference was simply to use the Pearson correlation coefficient. Although it is not optimal in this situation, it is quite adequate for our purpose. For an example Response Analysis, see our Statistics Laboratory Homepage.

One concern with MCR examinations is whether or not this type of exam may possibly adversely affect student performance. To examine this question, we first give a brief literature review on multiple-choice examination question and answer arrangement. Then the results of an experimental investigation with our own students taking one of our MCR examinations are then reported.

2. Brief Literature Review

A high quality multiple-choice test needs careful planning, constructing, and editing of the test's items or questions as well as the answer-options or multiple-choice responses for each item. After preparing the multiple-choice items and answer-options, the next step is to decide on the precise arrangement and ordering of the items and answer-options.

Previous empirical studies discussed by Gerow (1980, p.93) on the sequencing of questions have all failed to indicate any difference between random ordering of questions and questions organized by the order it was taught. Gerow (1980) presented further empirical evidence that arranging the items in order of difficulty also has no effect provided that there is enough time for students to complete the test. A further study by Allison (1984) confirmed that even for sixth grade students there was no effect on performance by ordering the items according to difficulty provided that there was enough time to complete the test.

Tuck (1978) found that when students in an introductory psychology class were asked for their preference in item arrangements, 64% preferred a random organization.

These studies all support the use of the MCR examination design but none of the previous studies apply directly to the MCR case where both the order of the items as well as the order of the possible responses is randomized.

3. Our Experimental Investigation

Our null hypothesis is that student grades are not affected by the MCR procedure.

One specific alternative hypothesis of interest is that ordering the exam questions in the same order as taught could result in higher scores than just random ordering. If this was in fact the case, one could question whether the type of learning that has occurred is really what is wanted. Our opinion, shared by educational psychologists we talked to, is that most likely one would not want to reward such a type of learning anyway and so, if there were a difference in grades, this in itself would be a good reason for selecting an MCR design. Also Hopkins (1998, p. 234) suggests that it is necessary to avoid arranging items in the order in which they were presented in the textbook in order to achieve the logical validity test.

Another specific hypothesis of interest is whether the ordering of the answer-options could result in an improved score. The specific ordering we have in mind here is either a logical or numerical ordering of the possible answers. If speed is really the determining factor in the examination then this ordering might be expected to improve the student scores. Once again though the pedagogical value of such an exam is open to question. At our university, students with disabilities may be allowed up to about 50% more time. This fact really means that we should probably not put too much emphasis on speed of processing the examination material but rather more on the depth of understanding. The examination that we experimented with was designed so that most students would be able to complete it in the time allotted.

The exam chosen for our experiment covered four chapters of the textbook. There were eight questions from the first chapter, nine from the second, three from the third, and seven from the fourth. In total there were 27 questions and four answer-options for each question. The item difficulties on this exam were approximately the same and are independent of the position of the items within the exam.

To test the first alternative hypothesis we put the questions from each chapter in a corresponding section of the exam and randomized them within their section. And to test the second alternative hypothesis we carefully chose the questions and the possible answers so that there was either a logical or numerical ordering of these answer-option choices.

A two-factor experimental design with a covariate was used. Each factor had two levels. The first factor, denoted by I, was the question ordering. This factor was either to use a randomized order of questions or a partially randomized order in which the questions were placed in one of four sections of the exam that corresponded to the chapter of textbook and then all questions within each section were randomized. The second factor, denoted by O, was the ordering of the answer-options. The two levels corresponded to using a randomized order for the answer-options or else using a fixed order in which the answer-options were presented in a logical or numerical order. Thus there are four treatment combinations in our experiment:

(a) questions randomized and answer-options randomized
(b) questions partially randomized and answer-options randomized
(c) questions randomized and answer-options ordered
(d) questions partially randomized and answer-options ordered.

Notice that with this design all students will receive a unique exam – at least with very high probability. Neither the instructor nor student would be able to tell exactly which treatment combination was used for a particular exam without some careful examination. This was the second mid-term examination in a two-term course and so the first examination, which was completely random with respect to item and answer-option arrangement, was used as a covariate to reduce experimental error. If our null hypothesis was rejected we were prepared to make a statistical adjustment to the students' grades.

We ensured that each student received one of the four examination types at random. This was done by generating 500 exams of each of the four types and then randomly selecting without replacement four samples of size 125 from {1, 2, …, 500}. The exams with codes corresponding to the number selected in each sample were used to obtain 125 examinations for each of the four treatment combinations. These selected exams were then printed.

Our exam was administered in a double-blind fashion to 442 students with neither the student nor instructor knowing which type of exam was used for a particular student. The means and standard deviations for the students writing each type of exam are shown in Table 3.1 below.

Table 3.1. Summary of Means and Standard Deviations for Each Treatment Combination

	Number of Students	**mean**	**sd**
(a) I & O random	108	52.45	11.88
(b) I partial & O random	114	53.75	13.06

(c) I random & O fixed	107	54.72	12.65
(d) I partial & O fixed	113	53.40	11.32

It is obvious from Table 3.1 that it is very unlikely that there is any difference in scores between the treatment combinations. Note that in Table 3.1, the standard error of the mean is approximately the SD indicated in column 4 divided by 10.

Table 3.2 shows the means for each factor level. In the case of item randomization the observed mean is slightly less than when ordered. In the case of answer-option randomization, the observed mean is slightly higher than for when the items are ordered.

Table 3.2. Summary of Means and Standard Deviations for Each Factor Level

	Number of Students	**mean**	**sd**
I random	222	53.12	12.49
I partially random	220	53.70	12.00
O random	215	53.58	12.29
O ordered	227	53.24	12.21

Table 3.3 presents the analysis of variance that confirms that there is no statistically significant difference among the treatment effects. The covariate is, as expected, highly significant due to the fact the performance on this exam was highly correlated with their performance on the first exam.

Table 3.3. ANOVA of Our Experiment

Source	**DF**	**Sum of Squares**	**Mean Square**	***F*-Value**	**Pr > *F***
First mid-term	1	18840.1211	18840.1211	176.536	3.6×10^{-40}
I (item)	1	44.65097	44.65097	0.4184	0.518
O (answer-option)	1	0.44916	0.44916	0.0042	0.948
error	440	46637.2100	106.7213		

4. Concluding Remarks

The MCR examination design was used in our department for nine examinations with 1947 individual examinations being written and marked. The students were pleased with MCR examinations since it obviously increased the integrity of the examination process.

In one of these examinations we investigated the effect of randomization of the questions and possible answers on student performance and found that, as might be expected from previous empirical studies, there was no evidence for any effect.

Our Statistics Laboratory is available for producing and marking MCR examinations. If interested please contact our StatLab Manager whose contact information is given on the StatLab homepage

ACKNOWLEDGEMENTS

We would like to especially thank Professor Tom Wonnacott for his numerous insightful comments, suggestions and encouraging support for this project. The authors also wish to thank Professors Mike Atkinson, John Braun, David Bellhouse, Angela Jonkhans, Maree Libal, Harry Murray, Evelyn Vingilis for helpful comments and/or discussions. The StatLab at the University of Western Ontario developed the software for this project. Special thanks to Dr. Valery Didinchuk for assistance with Perl. We also acknowledge helpful comments from three JSE referees.

References

Allison, D. E. (1984), Test anxiety, stress, and intelligence-test performance. *Measurement and Evaluation in Guidance*, 16, 211-217

Brenner, M. H. (1964), Test difficulty, reliability and discrimination as functions of item difficulty order. *Journal of Applied Psychology*, 48, 98-100.

Gerow, J. R. (1980), Performance on achievement tests as a function of the order of item difficulty. Teaching of Psychology, 7, 93-96.

Hopkins, Kenneth D. (1998), *Educational and Psychological Measurement and Evaluation.* Allyn and Bacon.

Tuck, J.P. (1978), Examinee's control of item difficulty sequence. *Psychological Reports*, 42, 1109-1110.

Web References

Comprehensive Perl Archive Network (CPAN), http://www.perl.com/CPAN/README.html

HTML Homepage, http://www.w3.org/MarkUp/

LaTeX Project Homepage, http://www.latex-project.org/

Mathematica, Wolfram Research, http://www.wolfram.com/

Microsoft® Visual Basic for Applications, http://msdn.microsoft.com/vba/

Perl Homepage, http://www.perl.com/

Rich Text Format, Version 1.5 Specifications, http://www.biblioscape.com/rtf15_spec.htm

Scantron Homepage, http://www.scantron.com/

Statistics Laboratory, University of Western Ontario, http://www.stats.uwo.ca/statlab**.**

A. Ian McLeod, Professor and Chair,
Department of Statistical and Actuarial Sciences,
University of Western Ontario,
London, Ontario, N6A 5B7
Canada

Ying Zhang, Ph.D. Candidate and StatLab Manager
Department of Statistical and Actuarial Sciences,
University of Western Ontario,
London, Ontario, N6A 5B7
Canada

Hao Yu, Associate Professor
Department of Statistical and Actuarial Sciences,
University of Western Ontario,
London, Ontario, N6A 5B7
Canada