MATH4561H Final Project

Do students avoid edge options?

Melissa Van Bussel

Winter 2019

Contents

Introduction 3

Background 3

Definitions . . . 3

Motivation . . . 3

Previous Studies 4

The Experiment 4

Response Variable . . . 4

Experimental Controls . . . 5

Incentive . . . 5

Guessing . . . 6

Number of Options . . . 7

Type of Options . . . 7

Presentation of Items . . . 7

Number of Questions . . . 8

The Experimental Design 8

Blocks and Experimental Units . . . 8

Factors . . . 8

First Factor: Test Version . . . 8

Second Factor: Prior Knowledge . . . 9

Third Factor: Time given . . . 10

The model . . . 11

Auxiliary Information . . . 11

Verifying Assumptions of the Model 12

Normally-distributed Errors . . . 12

Independent Errors . . . 12

Constancy of Variance Across All Treatment Levels of a Factor . . . 13

Verifying Assumptions with only 1 replicate . . . 13

A Sample Randomization Plan 13

Sources of Uncontrolled Variation 14

Summary 14

References 15

Introduction

"When in doubt, choose *C*"

This well-known phrase provides advice to those who become stumped on a multiple choice question. While this advice may sound silly, a preliminary observational study completed at Trent University in 2018[1] found significant ($\alpha = 0.05$) evidence of middle options such as 'B' or 'C' being disproportionately represented in the answer keys of multiple choice final examinations.

This report introduces the concept of edge avoidance, explores its role in post-secondary institutions and its presence in everyday life, and presents a 2^3 factorial design to examine the extent to which university students avoid edges on multiple choice examinations.

Background

Definitions

An *item* refers to an individual multiple choice (MC) question. The *stem* of an item refers to the proposed question, and the *keyed option* is the correct answer to the stem. The remaining options of the item are known as *distractors*. If the options for an item are ordered, then the first and last options are known as *edge options*, while the remaining options are referred to as *middle* options.

Edge avoidance refers to the systematic tendency of test makers and test takers to avoid edge options when hiding or guessing the keyed option of an MC item, respectively.

This psychological phenomenom is seen in many other real-world situations, too, and is sometimes referred to as "edge aversion", "middle bias", or "central tendency bias". Some examples include customers purchasing items which are centrally located on restaurant menus or in store displays, or employers favouring job candidates located centrally in a photograph[2].

Motivation

In many post-secondary institutions, multiple choice assessments are used to evaluate students' understanding of course material. Often, these assessments constitute a significant portion of students' final grades in a course. Using multiple choice as an evaluation method allows instructors to implement tests which cover a large range of material, and which can be marked rapidly. With ever-increasing class sizes and proliferating popularity of electronic and automated teaching tools, multiple choice examinations are likely to become increasingly commonplace in post-secondary courses.

Due to the wide usage of this evaluation method in universities, it is useful to understand the phenomenom of edge avoidance, in order to provide both instructors and students with strategies to get the most out of multiple choice evaluations.

Previous Studies

An observational study of 163 Scantron multiple choice examinations taken at Trent University from 2013 to 2018 was conducted in the summer of 2018[1]. In this study, the phenomenom of edge avoidance was not the central focus, and was incorporated into the analysis at the end as a point of interest. This was mainly done to determine whether or not this phenomenom was worth exploring with a properly designed experiment. In this study, each of the 163 tests had statistically significant evidence ($\alpha = 0.05$) of students avoiding edges on at least one question of the examination, meaning that the true proportion of edge distractors chosen by the test takers who had answered the question incorrectly was significantly lower than the expected proportion of edge distractors chosen by the test takers who had answered the question incorrectly. In this sense, the study looked at edge aversion tendencies on isolated questions, rather than on the test as a whole.

An experiment conducted in 2003 at the Hebrew University of Jerusalem investigated edge avoidance tendencies of both test takers and test makers of isolated multiple choice items. In this study, edges were avoided at alarming ratios of up to 3 or 4 to 1.[3]

The experiment proposed in the following section aims to investigate edge aversion tendencies of test taking individuals, and to ascertain which factors play a role in determining the degree to which an individual university student will avoid edges on an examination. This differs from previous work in the literature, which explored the phenomenom only for isolated questions. Results from the aforementioned previous studies suggest that there may be significant factors which predict edge aversion tendencies.

The Experiment

Response Variable

When a student is faced with answering a multiple choice question on a university examination, they will take into consideration the content of each option, rather than the ordering of the options. The student will choose the option whose content they deem to be the most “correct”. Even when the student is not immediately aware of which option is the keyed option, they will make an educated guess based on the substance of each option. Thus, on a real university examination, it is virtually impossible to determine whether test takers are actually avoiding edges, or if the test maker has done a poor job at creating and randomizing distractors.

If it is known with certainty that a test taker is guessing when answering each item of a multiple choice examination, then it is quite straightforward to determine whether or not the student avoided edges on the exam by performing a simple hypothesis test.

Assume that a student guessed the correct answer to each item on a multiple choice examination with n items in total. Let $\hat{p}$ be the proportion of items for which the student chose an edge option. Let p be the expected proportion of items for which the student should have chosen edge options, given that they did not avoid edges. For example, if a student completed an $n = 100$ multiple choice examination wherein each question had 3 distractors in addition to a keyed option, then it would be expected under the null hypothesis that the student chose edge options 50% of the time. Assume that the student chose edge options 36% of the time, then $p = 0.50$ while $\hat{p} = 0.36$. In this case, the following hypothesis test could be conducted:

$$H_0 : p = 0.50; \qquad H_A : p < 0.50$$

where the test statistic is t_{n-1}-distributed:

$$T = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} = \frac{0.36 - 0.50}{\sqrt{\frac{0.36(0.64)}{100}}} = -2.91667 < -T_{0.05}$$

Since $T < -T_{0.05}$, there is significant evidence at $\alpha = 0.05$ significance level that the student avoided edges on the examination.

For the proposed experiment, the response variable is the t_{n-1}-distributed test statistic from the above hypothesis test.

Experimental Controls

Incentive

On a post-secondary examination, students have large incentive for answering as many questions correctly as they are able to. Answering questions on an examination incorrectly will result in lower grades, as well as a number of other potential downfalls. Performing poorly on a university examination could result in lowered grade point average (GPA), loss of scholarships, academic probation, denial of entry to future courses, rejection of applications to other post-secondary institutions, and so on. Thus, in a real examination, students are always motivated to perform their best.

For this reason, it is important to ensure that there is adequate incentive provided to students participating in the proposed study, so that they answer questions as they would on a real examination. This incentive would depend largely on budgetary constraints, but but offering monetary compensation proportional to the student's performance on the examination given during the study is a proposed way to simulate the incentive that a student might feel during a real post-secondary examination. For research ethics purposes and for experimental control, it is recommended that all student participants receive the same amount of compensation for their time, regardless of their performance on the examination during the study; however, participants should not be informed of this detail until after the completion of the study.

If this step is omitted, it is likely that student participants will select their answers without thinking, and systematic tendencies of interest will not be observable. If the participant knows ahead of time that they will receive the same financial compensation regardless of their performance, they will not be motivated to take the examination seriously.

Guessing

As discussed previously, it is not possible to explore the phenomenom of edge avoidance in situations where the student is certain of the keyed option. It is only when the student is guessing and is actively attempting to find the correct answer that this tendency will be observable. For this reason, it is necessary to ensure that students are not only incentivized to answer each question correctly, but that they are also forced to guess the correct answer.

One way to ensure that participants guess all answers on the examination is to present them with a test which has no correct answers. For example, the following question could be presented:

The Greek word $\Psi\Psi\mu$i means which of the following?

(a) Door
(b) Floor
(c) Wall
(d) Ceiling

The question above has no correct answer, as the word actually means "bread". By presenting students with questions such as the ones above, it can always be ensured that students are guessing. Not only is there no correct answer, but it is highly unlikely that the average university student would know the Greek word for "bread".

Unfortunately, using questions which appear to be as difficult as the one presented above may introduce problems. If a student participating in the study is faced with many questions which have a high perceived difficulty, the participant will become easily frustrated and flustered, which may affect the way in which they answer the items on the examination.

Instead, it is recommended to use "Who Wants to be a Millionaire" or "Jeopardy" style questions, for which the questions are perceived to be "general knowledge" questions that a student may conceivably know the answer to. For example, consider the following popular culture trivia question which has no correct answer:

In Dreamworks Animation's 2010 hit film "Shrek Forever After", which character did Harrison Ford voice?

(a) Butter Pants
(b) Rumpelstiltzskin
(c) Brogan
(d) Ugly Stepsister

This question is well-suited to the study examination for a number of reasons. To begin with, *Shrek Forever After* is a popular movie, but it is not as popular and iconic as the original *Shrek* film. Additionally, given that it came to theatres in 2010, it is likely that most students who have seen the film have not seen it in a long time, and may have forgotten many details about the movie. Since this movie featured many famous actors and actresses, most students who are faced with this question are unlikely to suspect that there is no correct answer.

While questions which are similar in style to the *Shrek Forever After* question are preferable over questions similar to the question about the Greek word for bread, they can still present some problems. It is not advisable to create an examination for the study in which all questions are lacking correct answers, as students may still feel frustrated if they never feel certain in their answers. As this could affect the way in which they respond to the multiple choice questions on the examination, it is instead recommended to use a combination of questions which have correct answers and questions which do not.

If such a combination is used, external psychological effects which could affect analysis are minimized. In this setup, the questions which have a correct answer serve as a control. The responses to these questions will not be analyzed, and merely serve as a way to ensure that participants do not feel frustrated while completing the examination, and that they stay incentivized by believing they are performing “well” on the test. Conversely, the questions which have no keyed option will be analyzed for systematic edge aversion tendencies.

For the items which have correct answers, it is advised to use a varying range of difficulty. Some of these questions should allow for a high difficulty index (“easy” questions), while some should allow for a low difficulty index (“hard” questions). This is to again minimize external psychological effects that may occur if the examination is perceived to be too difficult or too easy as a whole.

Number of Options

For the proposed experiment, each question should have four options. This is because it is quite standard on post-secondary examinations to have multiple choice questions which have approximately four options per item. Using four options also ensures that many questions may be asked of the participants. If more distractors are to be used for each item, participants will need to spend a significant amount of time reading and processing options which do not add any additional value to the study. It has also been argued that questions with a larger number of distractors can actually serve to make the question less efficient at discriminating strong students from weak students.

By ensuring that each question on the study examination has the same number of options, the hypothesis test discussed previously can be easily conducted for each participant. This is because the proportion of edge options and middle options will be the same for each question, so the expected proportion of selected edge options for guessed answers will always be 0.50.

Type of Options

For the proposed experiment, each question should be strictly multiple choice with four distinct options. Options such as “all of the above”, “none of the above” or “Both B and C” should not be included. This is because options such as these are most frequently placed as the last option on an item, which is an edge. Including questions with options such as these could greatly affect the analysis for this reason.

Presentation of Items

In order to observe systematic edge aversion tendencies, it is important to present all options one after another, in order to reduce potential variation. For example, presenting four options in a 1 x 4 presentation visibly creates options “A” and “D” as edge options with “B” and “C” as middle options, while presenting four options in a 2 x 2 presentation does not create obvious visual edges, and may even serve to create bias. Since English-speakers read from left to right, it is possible that options presented on the left may be more favourable than options presented on the right, or vice versa. To avoid this, options must be presented in a 1 x 4 presentation. For the same reason, all questions should be presented one after another, and there should only be one column of questions on each page of the examination used in the experiment.

Number of Questions

To ensure that the test statistics from the previously described hypothesis tests are approximately Normally-distributed, a sufficiently large value of n must be used, as per the Central Limit Theorem. For this reason, it is proposed that the examination in the experiment consist of eighty questions in total, where forty of the questions have a correct answer, and forty do not. This implies that $n = 40$ for the purpose of computing the test statistics for each participant, which is sufficient.

This number is also recommended for budgetary reasons. An examination any longer than this would be very expensive to run, due to the amount of time that each participant would need in order to be part of the study.

The Experimental Design

Blocks and Experimental Units

As the response variable of interest is psychometric in nature, it is necessary to block on participant. This is because there will be large variation between subjects, and this should be incorporated into the model in order to reduce the experimental error.

It is important to note that while there may be other blocking variables of interest, these will be extremely difficult to incorporate into the model directly. For example, it would be difficult to obain Research Ethics Board approval for any kind of experiment which appears to “discriminate” based on factors which may be of interest, such as gender, GPA, or major. This sort of information is best left for post-hoc analysis, and is included in the “Auxiliary Information” section of this report.

The experimental unit of this design is the ____ .

Factors

First Factor: Test Version

It is hypothesized that edge avoidance on a particular set of questions is independent of the ordering of the questions and the options within each question, assuming that the content of the examination itself is the same. In order to explore this hypothesis, a test version factor can be included in the design.

First level: Structured

For the first level of the test version factor, the ordering of the options for each item will be completely randomized, subject to the following constraint: questions which have correct answers will have evenly-distributed correct answers. In other words, out of the forty questions which have correct answers, ten of these questions will have "A" as the keyed option, ten will have "B" as the keyed option, and so on.

This ordering will be adopted because many test takers tend to make assumptions about the distribution of correct answers, and will often count the frequency of their answers when determining which option to select on questions for which they are guessing the keyed option. This ordering is also similar to real-world post-secondary examinations, where instructors will often put effort into ensuring there is an approximately even distribution of correct answers.

Furthermore, the order of the questions themselves will be completely randomized, subject to the following constraint: the examination will be split up into ten chunks of 8 questions each, wherein each chunk will consist of four questions which have correct answers (one of "A", "B", "C", and "D") and four questions which do not.

This setup is suggested for the purposes of controlling variation. It is possible that a participant may become frustrated with the examination if they answer many questions in a row which do not have correct answers. This ordering of questions prevents this from happening, and ensures that a participant will never have to answer more than eight questions in a row which do not have correct answers.

Second level: Completely randomized

In the second version of the examination, the ordering of all questions will be completely randomized. The four options per question will also have a completely randomized order. The list of possible options will be the same for each item for both versions of the examination.

With this version of the examination, there is no structure at all. This will be interesting to analyze, because many post-secondary instructors will use randomization techniques on examinations, in order to deter students from counting the frequencies of their answers or looking for patterns in the keyed options.

Second Factor: Prior Knowledge

It is hypothesized that the information given to students at the start of an examination affects the way in which the student completes the examination. Thus, prior knowledge about the structure of the examination could have an effect on whether or not students avoid edges on the examination.

First level: Completely randomized

For the first level of this factor, study participants will be told that the ordering of all questions and the arrangement of all options for each item has been completely randomized. Thus, they are not to expect any kind of ordering or structure, and any ordering or structure is complete coincidence. It is important to note that participants which are assigned to this factor will be told this information regardless of whether or not it is true – that is, they will be told that the examination has been completely randomized, regardless of whether they have test version 1 or test version 2.

It is hypothesized that telling this information to students at the start of the examination (regardless of whether or not it is true) will minimize middle bias, and will therefore decrease the participant's edge avoidance. If this hypothesis is true, it would allow scientists to provide valuable recommendations to the educational community.

Second level: No prior knowledge

For the second level of this factor, study participants will not be told anything at all about the structure of the examination. They will simply be given the examination and asked to complete it.

It is hypothesized that students who have not been given prior information about the structure of the examination will seek out patterns in their responses, whether subconsciously or not. This could have an effect on the participant's likelihood to avoid edges on the examination.

Third Factor: Time given

The amount of time given to a student to complete an examination may have a substantial effect on the mental state of the student during the course of the examination.

First level: Ninety minutes

The first level of the time factor is ninety minutes. This corresponds to an average of 1.125 minutes per question, which should be enough time for the student to read and process the question and select a response, without having much time to deliberate between the options.

This treatment level will create a sense of urgency for participants. For some students, this could induce major text anxiety. Other students may perform well under pressure, and will be highly focused when completing the examination.

It is hypothesized that a smaller amount of time to complete the examination will result in a higher proportion of middle options selected. This is because students will have less time to think about their responses, and will subconsciously avoid edges.

Second level: Unlimited time

For the second level of this factor, participants will not be given a time limit, and will instead be told to take as long as they need. The unlimited time given could cause some students to begin searching for patterns and, as a result, could mean that participants do not avoid edges as much as those with a time limit.

The model

The proposed experiment uses a completely randomized 2^3 factorial design. Thus, the experiment uses an RCBF, or Randomized Complete Block Factorial.

The general model is:

$$\begin{aligned} y_{ijkl} = \mu &+ b_i + \alpha_j + \beta_k + \gamma_l \\ &+ (\alpha\beta)_{jk} + (\alpha\gamma)_{jl} + (\beta\gamma)_{kl} \\ &+ (\alpha\beta\gamma)_{jkl} + \epsilon_{ijkl} \end{aligned}$$

where

$$i = 1, 2, 3; \quad j = 1, 2; \quad k = 1, 2; \quad l = 1, 2$$

and μ is the grand mean, b_i is the block effect, α_j is the test version effect, β_k is the prior knowledge effect, and γ_l is the amount of time given effect. The last term is the experimental error term, and the remaining terms are interaction terms.

It should be noted that an additional index, r, could be used on each term if the experiment had more than one replicate. If budgetary constraints allow, more than one replicate should be used, to ensure that there are enough degrees of freedom to adequately model the data.

Since there is currently no data available for this experiment, it is not possible to determine which terms would be significant and should be included in the model. This is simply the starting point for the model.

Auxiliary Information

In addition to including several factors in the experimental design itself, it is also useful to collect auxiliary information at the end of the examination for enhancing post-hoc analysis. Upon completion of the examination, participants would be given a voluntary survey where they may answer any questions they are comfortable answering. This should include questions such as:

- What is your overall GPA?
- Do you experience test anxiety?
- What is your major?
- How many multiple choice final examinations have you written at Trent University?

Participants should also be given the option to provide any additional comments or feedback.

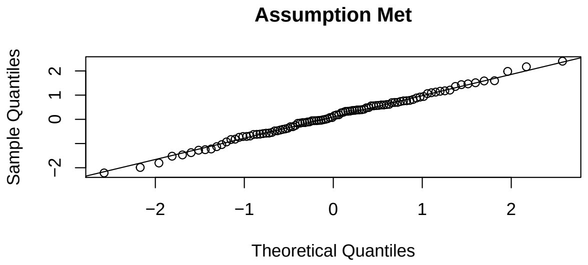
Responses to the optional survey could be used to stratify results, and to guide future research in this area.

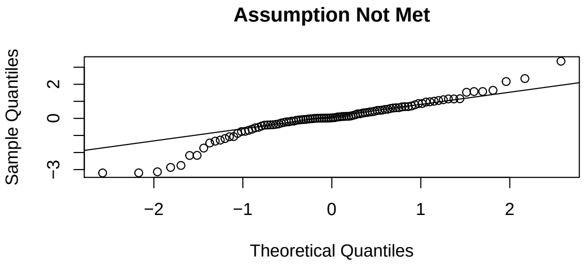
Verifying Assumptions of the Model

For the following sections, data will be “made-up” in order to demonstrate how the assumptions of the model would be verified, if there were available data.

Normally-distributed Errors

The first assumption of the model is that the experimental errors are normally distributed. There are a number of ways to verify this assumption. The first way is to plot a Normal Q-Q plot of the errors, and confirm that the points fall along the straight line. If they do not, this assumption might not be met.


Assumption Met

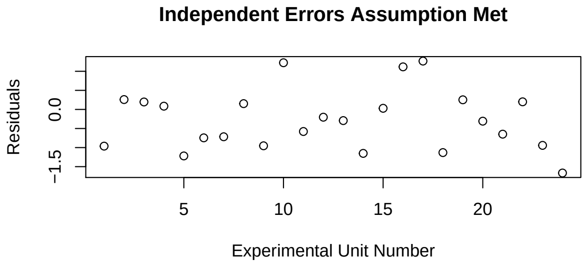

Assumption Not Met

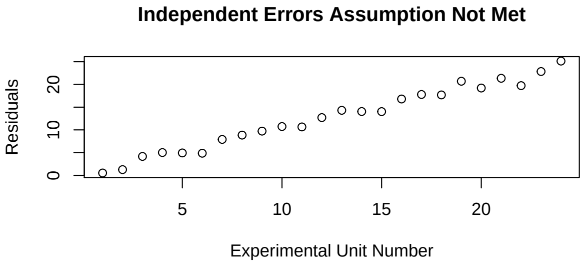
This assumption can also be verified by performing a Shapiro-Wilk test by using the `shapiro.test()` function in `R`. The null hypothesis of this test is that the errors are Normally-distributed. Thus, small p-values from the output of the `shapiro.test()` function indicate that the assumption is not met.

Independent Errors

The second assumption of the model is that the experimental errors are independent. This indicates that there was proper randomization when assigning treatments to the experimental units.

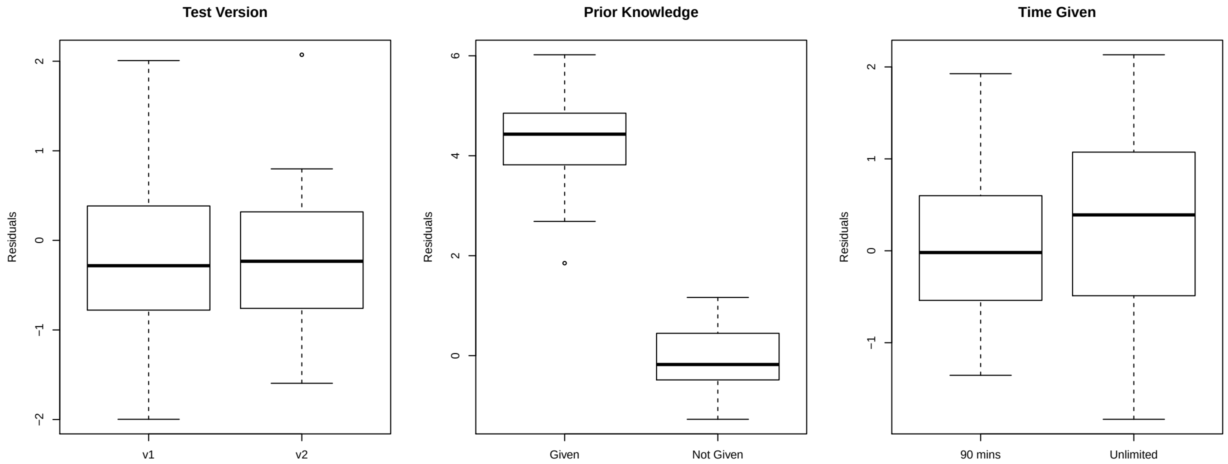
To verify this assumption, a simple scatterplot of the residuals versus the experimental unit number can be created. Any trends in the scatterplot indicate that the randomization may not have adequately balanced units across the treatment levels of the factors. This could include increasing trends, decreasind trends, or periodic trends.


Independent Errors Assumption Met


Independent Errors Assumption Not Met

Constancy of Variance Across All Treatment Levels of a Factor

The third assumption of the model is that the variance of the experimental errors is constant across all treatment levels of a factor. To verify this assumption, a boxplot of the residuals by treatment level can be created for each of the factors.



In the above example, the assumption would have been met for the test version factor and the time given factor, but not the prior knowledge factor.

Verifying Assumptions with only 1 replicate

It has already been stated that it is preferable to include multiple replicates for this experiment; however, this may not be possible due to financial constraints.

If there is only one replicate per cell, it can be extremely difficult to verify the assumptions of the model. This is because there are so few observations that a single outlier can cause the assumptions to be broken. If this occurs, Daniel's Method (1960)[4] can be used. The `Gaptest` function from the `daewr` package in `R` can detect and correct atypical values that would otherwise prevent analysis of the data. If this method is used, it is important to proceed with caution. The experiment may need to be re-run.

A Sample Randomization Plan

While it is unknown how many replicates of the design there would be, the following `R` snippet demonstrates how a randomization plan could be created for this design. Assume $r = 3$ replicates.

```r
version <- c("structured", "random")
prior <- c("given", "none")
time <- c("90", "unlimited")
combo <- expand.grid(version, prior, time)
```

```r
replicate <- rep(1:3, each = 8)
order <- sample(1:24, 24)
edge <- data.frame("Test Version" = combo$Var1,
                   "Prior Knowledge Given" = combo$Var2, "Time Given" = combo$Var3,
                   "Replicate" = replicate, "Order" = order)
```

Test.Version	Prior.Knowledge.Given	Time.Given	Replicate	Order
structured	given	90	1	3
random	given	90	1	19
structured	none	90	1	5
random	none	90	1	17
structured	given	unlimited	1	1
random	given	unlimited	1	16
structured	none	unlimited	1	10
random	none	unlimited	1	22
structured	given	90	2	24
random	given	90	2	12
structured	none	90	2	20
random	none	90	2	15
structured	given	unlimited	2	8
random	given	unlimited	2	18
structured	none	unlimited	2	13
random	none	unlimited	2	6
structured	given	90	3	23
random	given	90	3	7
structured	none	90	3	2
random	none	90	3	4
structured	given	unlimited	3	11
random	given	unlimited	3	21
structured	none	unlimited	3	9
random	none	unlimited	3	14

Sources of Uncontrolled Variation

- presumably we would have all people with factor 2 levl 1 and a time limit writing in same room same time to save costs, but where people sit in the room and the fact that theres other people in the room could stress people out and affect things

Summary

With regards to multiple choice questions, edge avoidance refers to the psychological phenomenom of test takers avoiding edge options (such as “A” or “D”) in favour of middle options (such as “B” or “C”). This phenomenom can have a significant impact on post-secondary students’ grades, finances, and even mental health. Understanding which factors affect edge avoidance will allow researchers to provide valuable recommendations to professors and students alike about how to administer and how to take multiple choice examinations.

In order to study edge avoidance, a Randomized Complete Block Factorial (RCBF) design was developed, consisting of 3 factors with 2 levels each. The first factor was the test version (structured or random), the second factor was prior information given to students (prior information given or no prior information given), and time given (ninety minutes or unlimited time).

References

1. Fitze, K., & Van Bussel, M. (2018) Making your r' PRIME: Exploring the Quality of Multiple Choice Testing at Trent. Physics Department Summer Seminar Series, Trent University.

2. Jarrett, C. (2012). People Prefer the Middle Option. The British Psychological Society Research Digest, 30 April 2012.

3. Attali, Y., & Bar-Hillel, M. (2003). Guess where: The position of correct answers in multiple-choice test items as a psychometric variable. Journal of Educational Measurement, 40(2), 109-128.

4. Lawson, R. (2014). Design and Analysis of Experiments with R. CRC Press.

5. Battista, D., & Kurzawa, L. (2011). Examination of the Quality of Multiple-Choice Items on Classroom Tests. Canadian Journal for the Scholarship of Teaching and Learning, (2):2.