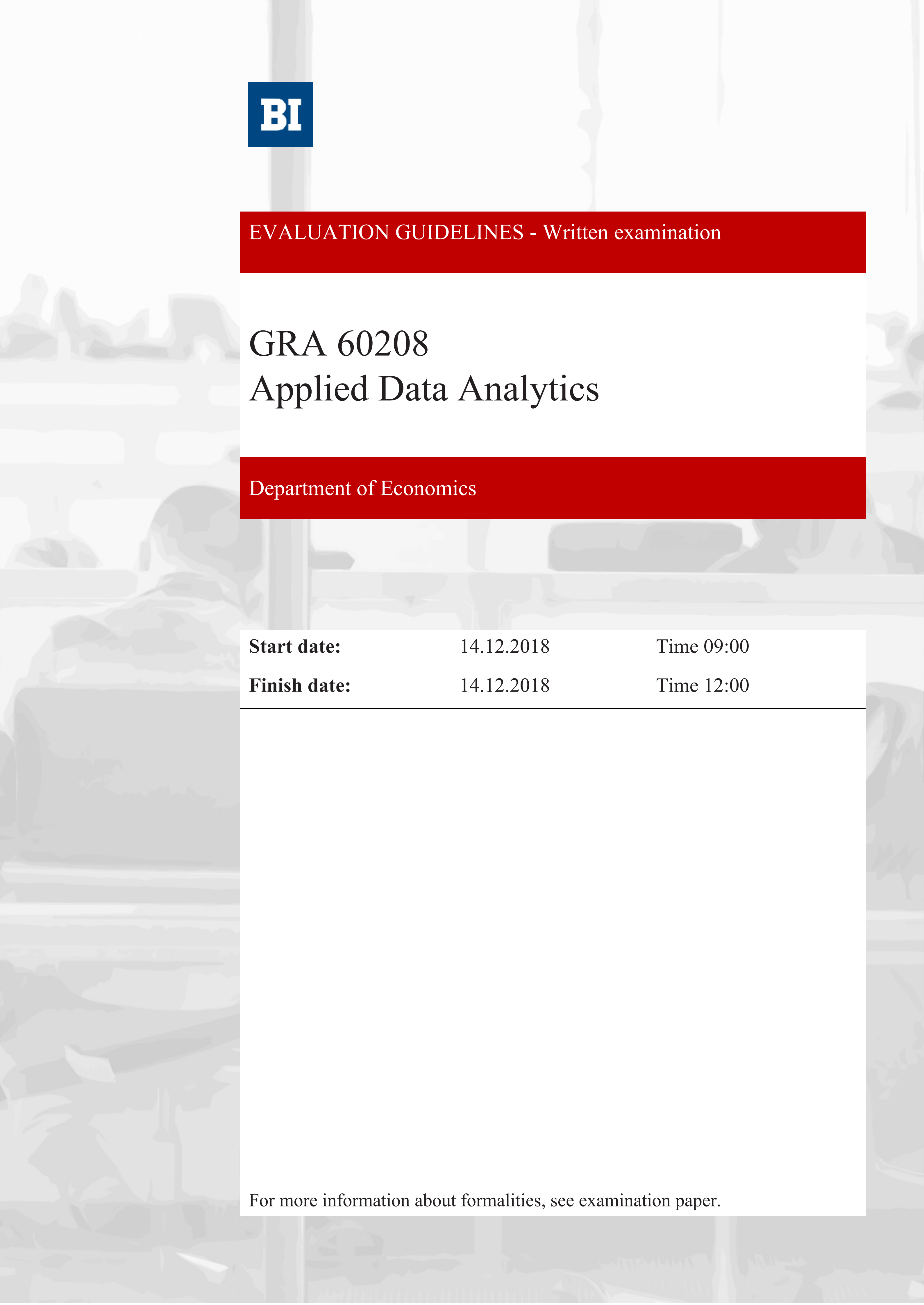

EVALUATION GUIDELINES - Written examination

# GRA 60208
# Applied Data Analytics

Department of Economics

| | | |
|---|---|---|
| **Start date:** | 14.12.2018 | Time 09:00 |
| **Finish date:** | 14.12.2018 | Time 12:00 |

For more information about formalities, see examination paper.

# SCHOOL EXAM GRA6020 – AUTUMN 2018

The first part of the exam consists of five multiple choice questions. There is just one correct answer for each question. This first part counts for 30% of the total points on the exam. For each multiple-choice question, you either get a full score, 100 points, or no points, and you will not be penalised with negative points for wrong answers. The second part (Part 2 and Part 3) of the exam is a standard written exam and counts for 70% of the total points on the exam. Within each part, each task/sub-task carries equal marks, and each task/sub-task is scored from 0 to 100, for example, Exercise 7 has the same weight as every sub-task in Exercise 8. The grade limits are as follows: A $\geq$ 90, B $\geq$ 75, C $\geq$ 55, D $\geq$40, E $\geq$ 25.

**PART 1: Multiple choice questions (Counts 30%).**

**Exercise 1:** C

**Exercise 2:** B

**Exercise 3:** D

**Exercise 4:** C

**Exercise 5:** B

**PART 2: Mathematics (Counts 20%).** In general, calculation errors, inconsistent or incomplete arguments will reduce the total score for each subtask; consequential errors (følgefeil) are not penalised very hard.

**Exercise 6:**

**a)** For the data in the table in the exercise, we have that

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{4}\sum_{i=1}^{4} x_i = \frac{1}{4}(-1.10 + 0.50 + 0.70 - 0.1) = 0$$

and similarly for the $y_i$ sequence

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{4}\sum_{i=1}^{4} y_i = \frac{1}{4}(-2.40 + 1.10 + 1.00 + 0.30) = 0$$

**b)** Since we know from **a)** that $\bar{x} = 0$ and $\bar{y} = 0$, we have that $\widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1\bar{x} = 0$, and for $\widehat{\beta}_1$ we have that

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n}(x_i - 0)(y_i - 0)}{\sum_{i=1}^{n}(x_i - 0)^2} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}.$$

**c)** From **b)** we have that $\widehat{\beta}_0 = 0$. In order to compute $\widehat{\beta}_1$ we plug the numbers from the table into the formula above, this gives

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2} = \frac{2.64 + 0.54 + 0.70 - 0.03}{1.21 + 0.25 + 0.49 + 0.01} = \frac{3.85}{1.96} \approx 1.96.$$

Since 1.96 is, for all practical purposes (at least when we draw), close to 2.00, the final part of this task is to draw a line with slope approximately equal to 2 and intercept of 0.

## PART 3: Data analysis (Counts 50%).

**Exercise 7:** The data presented in Table 3 and the corresponding description/information is incomplete and it is not possible to conclude; Serena Williams might be right, and she may not be, to paraphrase one of her countrymen.

Suggested reasons for the differing number of fines issued to women and men are listed below. In the answer, the conclusion should be clearly stated. Too many students listed descriptive facts, but failed to reach a conclusion. In this exercise several reasons for why the data do neither corroborate nor refute Serana Williams' claim are accepted. Also, stating that based on the available data, the burden of proof appears to lie with Serena Williams will give some points.

The main problem is that we do not know what actually happened. The table only provides us with the total number of fines from the collection of (potentially) biased judges. This can perhaps be solved, since the matches are most likely filmed, and in principal it could be analysed by an independent panel of judges to (hopefully) obtain proper unbiased data which can then be compared to the numbers in Table 3.

Below are additional possible explanations for the numbers given in Table 3.

- Men play best-of-five while women play best-of-three. Therefore, men play more sets than women, which means that male tennis players in general have more time to get fines.
- By playing more sets the match is usually longer, and perhaps playing longer matches (more sets) makes players more tired and aggressive, which again could result in more fines.
- We do not know if men are generally more (or less) aggressive than women, perhaps men (or women) are involved in reprehensible behaviour more often than women (or men).
- The distribution of fines over time is not shown in the aggregated table, and perhaps most of the fines issued to women were issued in the most recent years; which may support the claim of Serana Williams.
- We do not know if there are a few individuals that are causing the majority of fines. In principle, a handful of male players could single-handedly drive up the number of fines.

There are probably many more such arguments, and all convincing and coherently presented arguments are rewarded. In general, however, few students commented on the fact that we do not know what actually happened. Most argued that we could not conclude based on of the more specific reasons listed above.

**Exercise 8:** Below are some general comments that apply to most subtasks in Exercise 8:

- It is important to avoid just listing facts.
- It is important to write complete arguments.
- Listing the relevant facts, without linking facts and arguments together and providing a proper conclusion will reduce the score.
- Presentation is important. Arguments and insights should be easily accessible to the reader
- The results and conclusions made in 8a, 8b and 8c should be reflected in the non-technical summary in 8d.

**a)** The answer to this exercise must be based on the summary of the baseline model in the appendix to the exam. It is hard/impossible to answer this properly by only looking at tables and plots. This task is a lot easier to answer if one first compute the numbers corresponding to the four possible combinations of image and headline as shown below:

| | Headline I | Headline II |
|---|---|---|
| Image A | 91.6 | 65.5 |
| Image B | 102.3 | 40.5 |

The right answer, when trusting the model, is that Image B and Headline II is the least and Image B and Headline I is the most effective option. In addition, the students should comment on practical importance, which is quite high (requires computation of the actual numbers) and the statical importance, which is also high since all estimated coefficients have relatively small $p$-values.

**b)** First of all, note that a good answer for this subtask does not have to be as long as the following explanation.

This is the first task where the students must relate their answer to assumptions about the population and the particular sample of data. In order to provide an answer, we must assume something about the underlying population. It makes sense to assume that the population (users with registered accounts) are more or less balanced with respect to gender. If nothing is assumed about the underlying population, it is impossible to say anything about possible gender imbalance in the sample. There are hints about this in the assignment, where it is stated that this is one of the largest online newspapers in the country, and that the newspaper is dedicated to a gender neutral frontpage.

Furthermore, it is important to note self-selection occurs, meaning that only those who were sufficiently interested in the theme or topic interact with, or clicked on, the article. Therefore, the observed time-to-exit times are conditioned on, or given that, individuals (in the population) were sort of interested in the first place (sufficient to click on the article). Note that this also assumes that the images and headlines used are not completely misleading.

In order to obtain a full score, students are expected to comment on assumptions regarding the population, and that for the particular sample there is an imbalance in genders; as shown in the different tables. In short, more men clicked on the article. Looking at the box plot of time-to-exit and gender shows that men and women in the sample (those who were sufficiently interested in the theme or topic) are more or less equally engaged in the article; i.e. use more or less the same amount of time. Actually, it is a bit hard to use the box plot to conclude with anything, since we do not know the baseline. Meaning that we do not know if men in general use more or less time than women when reading an article. This is needed to argue that men or women spend more or less time, compared to normal behaviour, when reading this article. Therefore, under the assumption of a balanced population, the data indicate that this is an article with a theme that appeals more to men than women.

Note that regardless of the conclusion and argument used, there should be a link between this answer and the answer to subtasks 8c and 8d, since in 8c we see that the effect of gender is seen to be small and insignificant with respect to time-to-exit, and in subtask 8d students are asked to comment on gender neutrality (which should be based on the combined insights from 8b and 8c).

**c)** The following discussion is not meant to be an answer to the subtask, this is more of a summary of what is expected to be included (the most important one comes first). In order to obtain a good score on this subtask, one should **not** write a long list of descriptive facts. The student is expected to provide a coherent text that compares the so-called full model with the baseline model, with particular attention to the main hypothesis (the effect of image and headline on time-to-exit).

First of all, from the estimated coefficients we see that the conclusion found in 8a remains more or less unchanged, since the estimated coefficients for the common explanatory variabels are more or less the same. However, it is now less clear that the type of image used is important (larger *p*-value). The reason for the change in the *p*-value for the estimated coefficient for `ImageB` (which is the difference between the effect on time-to-exit between image A and B) is most likely related to the fact that image and gender is correlated (multicollinearity). However, it still looks like the type of image used is important for capturing the attention of a reader; which can be seen in the various tables. The *p*-values for the other common coefficients are still very small. Of the control variables, both `Age` and `Gender` do not contribute much (low statistical and practical importance). There is an effect of the type of device, however, this does not appear to affect the effect of the use of image and headline, hence the type of device does not appear to be a confounder for the relation we are studying. The residual plot indicate heteroscedasticity, we should therefore be careful when we interpret the *p*-values, but we can still trust the estimated coefficients. We see that the values of $\mathrm{R}^2$ and $\mathrm{R}^2_{\mathrm{adj}}$ increased, but this fact is not very important since the main goal is not to explain all the variation in time-to-exit, but rather to understand how the use of image and headline attracts readers (or affects engagement).

**d)** It is very important that the non-technical summary: (i) is clear, concise and easy to read, (ii) that conclusions and statements are based on the analysis in 8a–8c, and (iii) does not include technical language. The summary should be general, for example, instead of writing "... the difference between `Headline I` and `Headline II` is... " write "... the difference between using a descriptive and sensational headline is... "; this makes it much easier to read and understand, and is how one ought to present statistical findings to a statistically challenged boss.

According to the solution alluded to in the discussion above, the main conclusion, which should be included in the summary, is that the use of a sensational "you won't believe what happened next" type of headline (clickbait) may attract users that are not really that interested in the article, and that these tend to quickly leave the article, lowering the average engagement or time-to-exit of users, resulting in a poor user experience. Furthermore, based on the conclusion from 8b, there is reason to believe that this is an article that appeals more to men than women.

If students decide to comment on the the effect of the type of device, this should be framed in a way that makes sense to the whole summary, and similarly age, gender and the uncertainty regarding the quality of the model and/or sample.

**e)** Here, there are several different answers that give good scores. Again, listing descriptive facts is not sufficient to obtain a high score. It is very important that the text is clear, concise and easy to read. Note that stating that something is wrong, missing (data or variable) or that something may be a confounder and cause omitted-variable-bias, without explaining why it is wrong or why a certain variable might be a confounder, is generally not given any points.

1) There are several hints in the assignment. There are several aspects of the data and the full model that could be improved. All suggested improvements of the model or data should be well argued, it is not sufficient to say that "we need to control for $X$", the student is expected to provide a convincing argument for *why* we ought to control for $X$.
2) Among the few students who chose to answer this exercise, most were able to comment on the two approaches and also discuss alternative and creative ways of measuring engagement.
3) In this exercise, it is more important to discuss differences in articles than differences among users (unless a particular property of a user is a confounder). The reason is that different users will most likely have similar behaviour on similar articles, e.g. both a slow and fast reading user will spend more time reading long and complex articles compared articles that are to short and simple.
4) Few students chose this exercise, the main problem (which some students were able to comment on) is that "optimising" one article may depend on its surroundings. For

example, an eye-catching image does not stand out if all images on the webpage are of a similar eye-catching character.