There must be an error here!
Experimental evidence on coding errors' biases

Bruno Ferman* Lucas Finamor†

First draft: 02/02/2022

Pre-Analysis Plan

Link to Pre-Registration at AEA Registry — AEARCTR-0008312

*Sao Paulo School of Economics - FGV, bruno.ferman@fgv.com.
†Yale University lucas.finamor@yale.edu.

1 Introduction

This document pre-specifies the analyses for our experiment on how the probability of detecting coding errors depends on the nature of the findings.

2 Design Overview

2.1 Setting and experimental design

The study will take place in a recruitment process for research assistants of a Partner Institute, which promotes and supports research in Economics. In this recruitment process, candidates are asked to perform a data task to evaluate their coding abilities.

In the data task, we first present the main results of 5 different real randomized evaluations of computer aid learning platforms on math proficiency.[1] The estimated effects are in the interval 0.12-0.37 standard deviations, all of them statistically significant at the 5% level.

Next, we provide candidates a data set from a hypothetical randomized evaluation on the effects of a computer aid learning platform on math proficiency, following the presented literature. The implementation took place in two regions, A and B. In one of the questions, we ask candidates to estimate the effect of this program for a specific region. In this question, candidates are subject to a common coding error: not verifying that the outcome variable has a value of 99 in case the information on test scores is missing for a student. We also provide candidates a dictionary for the data set that has this information about how missing values are labeled.

The data sets that we provide to candidates experimentally vary the results candidates would find in case they do *not* take into account this coding for missing values. For the treatment group, candidates would (mistakenly) find a significant negative effect of the program on math test scores in case they include in the regression the 99 values. For the control group, they would (mistakenly) find a significant positive effect in the range of the effects found in the literature, in case they include in the regression the 99 values. The estimated effect is close to zero in both groups if the missing values are appropriately taken into account.

[1] The papers are Banerjee et al. (2007), Lai et al. (2013), Mo et al. (2014), Lai et al. (2015), and Muralidharan et al. (2019)

2.2 Sample

The sample for this experiment comes from candidates that apply for one of the research assistant vacancies at the Partner Institute. The partner institute decides which candidates will be asked to do the data task. The data task is embedded in an online survey created using the software Qualtrics. The partner institute will send personalized links to candidates with this task by email.

On the first page of the task, candidates are asked whether they consent to share de-identified information from their data task for research purposes. We will then receive de-identified data from candidates that consented.

Given the uncertainty regarding the number of candidates in each recruitment process and on the proportion of candidates that will consent to share their information, we have uncertainty regarding the final number of observations. Our target is to combine information from several recruitment processes until we attain a sample of around 800 candidates. We expect that this number will be attained with 8 to 16 recruitment processes.

In the de-identified data, we have a unique anonymous identifier for the email of the candidates. Therefore, if a candidate applies to two different positions, we will be able to identify her. We will only consider the first observation of each candidate, discarding the others.

2.3 Fairness concerns

Given the fact that the experiment takes place in a real recruitment process, it is not sufficient that it is *ex-ante* fair for all candidates — it needs to place candidates in analogous situations, despite the randomization. Our design aims to achieve this *ex-post* fairness as well.

In a first question (Q1), candidates are asked to estimate the treatment effects for a given region (e.g., region A). If they do not exclude the 99 (missing) observations, those in the treatment group will obtain a negative effect and those in the control group a positive effect. The first answer to this question is recorded to be used in the experiment. In the following question (Q2), candidates are asked to compute the same treatment effects for the other region (in our example, region B). Now, the candidates in the treated group will obtain a positive effect, and those in the control group will get the negative effect if they include the 99 observations.

After completing the questions, candidates have the opportunity to review all the questions and are free to change their initial answers. They are told since the first instructions that this would be a possibility. Only the final answers are used for the data task and the screening procedure. Therefore, while the experiment uses the answer to the first question, the screening uses only the final answer. When submitting the final answer, all candidates experienced the absolutely same problems, each receiving negative and positive estimates for one region — randomization only changes the order they appear.

We also piloted the experiment in four different recruitment processes, with 247 candidates that agreed on sharing their data for research purposes. The results are presented below. In the first column, we see that 8% of the control group candidates spot the first question error (control mean 0.08). Those in the treatment group were 6.7 percentage points more likely to see the error. This represents an 83.7% effect or approximately 25% of the standard deviation of this variable. The p-value is 0.107. In the second column, we assess the treatment effect on the final score for this question. As expected, we do not see a differential effect across the two groups. The point estimate is 0.014 out of a control mean of 2.117. This represents a higher score of only 0.7% or 1.5% of the standard deviation. The p-value is 0.899. We see the same in the total score of the task in the third column.

Table 1: Results from the Pilot — Fairness

	(1)	(2)	(3)
	Spotted the Error	Score Error Question	Total Score
Treatment	0.067	0.014	0.102
(s.e.)	(0.041)	(0.108)	(0.279)
[p-value]	[0.107]	[0.899]	[0.715]
N Obs	247	247	247
Control Mean	0.080	2.117	3.931
Control SD	0.272	0.903	2.218

Notes: The table shows the treatment effect estimates using equation 1, interacting the treatment variable with all demeaned controls (gender, education level, econometrics course and position fixed-effects). Robust standard errors are presented. The outcomes are respectively: whether the individual spotted the error, the final score in this question, and the total score.

3 Empirical Analysis

3.1 Primary outcome

Our primary outcome of interest is a variable indicating whether the answer to the first question was correct.

3.2 Estimation

Let Y_i be the indicator for whether the candidate i spotted the error in the first question. Candidate i is treated, $T_i = 1$, if she receives the negative estimate in the first question. We will estimate our treatment effects in a specification that interacts the treatment indicator with the demeaned control variables.

$$Y_i = \alpha + \beta T_i + \gamma \tilde{X}_i + \delta T_i \times \tilde{X}_i + \varepsilon_i \quad (1)$$

Where $\tilde{X}_i$ are all demeaned covariates: gender, whether the candidate took an econometrics course, whether the candidate has a master degree or above, and a position fixed-effect. The main estimation will be using OLS with robust standard errors.

3.3 Hypothesis

We hypothesize that individuals that obtained the negative estimates when not correcting the error (treated) will be more likely to revise the code and the data, and therefore more likely to spot the error. This is equivalent to testing whether $\beta > 0$. Since we do not have any prior on obtaining $\beta < 0$, we will use a unilateral test with the following hypothesis:

$$H_0 : \beta \leq 0$$

$$H_1 : \beta > 0$$

3.4 Heterogeneity Analysis

We will perform the following heterogeneities and additional analysis:

1. As robustness, we will remove all individuals that answered that the expected result was below 0.1 standard deviations. This number is well below the estimates presented in the data task.
2. To have a measure of proficiency, we will execute a heterogeneity analysis using an initial question that measures candidates' coding abilities in simple tasks as computing means, conditional means, and counting operations. We will split the sample into those individuals that scored more and less than 50% in this initial question.
3. Heterogeneity analysis on gender, master degree, and whether the individual took an econometrics course.
4. Since we do not have perfect information on the real-time individuals spent on the test, we will not do any heterogeneity analysis using the time of completion.

REFERENCES

Banerjee, A. V., Cole, S., Duflo, E. and Linden, L. (2007). Remedying education: Evidence from two randomized experiments in india, *The Quarterly Journal of Economics* **122**(3): 1235–1264.

Lai, F., Luo, R., Zhang, L., Huang, X. and Rozelle, S. (2015). Does computer-assisted learning improve learning outcomes? evidence from a randomized experiment in migrant schools in beijing, *Economics of Education Review* **47**: 34–48.

Lai, F., Zhang, L., Hu, X., Qu, Q., Shi, Y., Qiao, Y., Boswell, M. and Rozelle, S. (2013). Computer assisted learning as extracurricular tutor? evidence from a randomised experiment in rural boarding schools in shaanxi, *Journal of Development Effectiveness* **5**(2): 208–231.

Mo, D., Zhang, L., Luo, R., Qu, Q., Huang, W., Wang, J., Qiao, Y., Boswell, M. and Rozelle, S. (2014). Integrating computer-assisted learning into a regular curriculum: Evidence from a randomised experiment in rural schools in shaanxi, *Journal of development effectiveness* **6**(3): 300–323.

Muralidharan, K., Singh, A. and Ganimian, A. J. (2019). Disrupting education? experimental evidence on technology-aided instruction in india, *American Economic Review* **109**(4): 1426–60.