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

Bruno Ferman* Lucas Finamor†

First version: 02/02/2022
This version: 01/30/2024

Pre-Analysis Plan

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

*Sao Paulo School of Economics - FGV, bruno.ferman@fgv.br.
†Institute for Fiscal Studies, lucas.finamor@ifs.org.uk.

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. This new version updates the first version of the pre-analysis plan (PAP) written and uploaded to the AEA RCT Registry on 02/02/2022. Section 4 details the changes in this document.

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, a large international organization conducting research in Economics. Hereinafter, we will be using the term "Partner Institute" to refer to this organization. In this recruitment process, candidates are asked coding questions to assess their coding abilities. As part of this experiment, we include a direct data task as an additional module of their recruitment process.

In the data task, we first present the main results of 6 different real randomized evaluations of interventions tailoring content to students' appropriate levels on test performance.[1] The estimated effects are in the interval 0.08–0.16 standard deviations, all of them statistically significant at the 5% level.

Next, we provide candidates with a data set from a hypothetical randomized evaluation on the effects of an intervention that tailored content for students' appropriate level for language instruction, following the presented literature. The implementation took place in two similar states, 1 and 2. In one of the questions, we ask candidates to estimate the effect of this program for a specific state. 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 with a dictionary for the data set that has this information about how missing values are labeled.

The data sets 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 language test scores in case they included 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), Cabezas et al. (2011), Duflo et al. (2011) and Banerjee et al. (2016).

2.2 Sample

The sample for this experiment comes from candidates who apply for a fellowship program at the Partner Institute. After filling out the application, which already includes multiple choice questions on coding, candidates are invited to complete the data task. While completing it is not a requirement, they are encouraged to complete it. The data task is embedded in an online survey created using the software Qualtrics. Each candidate has their own personalized link and identifier to log in.

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 only use data for the experiment from candidates who consented to share their data.

Given the uncertainty regarding the total number of candidates, the proportion of candidates that will take the test, and consent to share their information, we have uncertainty regarding the final number of observations. Our target is to attain a sample of around 800 candidates after applying our filters. We expect to achieve this number in this unique recruitment process. For the purpose of the evaluation, we will consider as part of the experiment: (a) those who took the data task and consented to share their data, and (b) those who were able to correctly run a regression in one of the screening questions. The last restriction is necessary as the coding error only applies if individuals know how to run an OLS regression.

Data collection is expected to start on February 1st, 2024.

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 state (e.g., state 1). 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 state (in our example, state 2). 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 state — randomization only changes the order they appear.

We also piloted the experiment in four different recruitment processes, with 247 candidates who agreed to share 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. That implies the task taker identified the 99 as a code for missing.

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.

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

Where $\tilde{X}_i$ are all demeaned covariates: gender, whether the candidate took an econometrics course, whether the candidate has a master's degree or above, the initial score in the screening questions, the score in the multiple choice coding questions (for the specific language that the test taker chose to execute the analysis), and the score on screening questions on knowledge of Econometrics.

The main estimation will be using OLS with robust standard errors.

3.3 Hypothesis

We hypothesize that individuals who 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 who answered that the expected result was negative. This number is well below the estimates presented in the data task.
2. Heterogeneity analysis on the following dimensions:
   (a) Gender
   (b) Master's degree or above qualification
   (c) Coding skills, as measured by whether the individual correctly clustered their standard errors in the screening question
   (d) Coding skills, as measured by all data screening questions and coding multiple-choice questions. For this, we will implement an above/below software-specific median cut.
   (e) Econometrics/Microeconometrics knowledge, as measured by multiple-choice screening questions. For this, we will implement an above/below median cut.

3. 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.

4 Changes to the first version

The RCT has been running with another Partner Institute since February 2022. However, contrary to the pilot with this original partner, the number of openings was lower than originally expected, and each opening had very few candidates. Therefore it was clear we would not reach the desired sample size. In this context, we approached a Second Partner Institute, a large international organization, which also hires several economists to conduct applied work. We developed a very similar, but different data task to be applied in their recruitment process. While the design is very similar to the original data task, the fictitious example and some of the questions were modified. In particular, we increase the number of coding questions before the coding error questions, to have a better measure of the coding abilities and knowledge of task takers. This is necessary now because the recruitment process of the Second Partner Institute is broader and may have task takers with less coding experience and knowledge of Economics and Econometrics. The following list layouts the main changes to the original plan:

1. The new data task now includes one more screening question, asking task takers to run a regression where there are no missing values, with clustered standard errors. The purpose of this question is, together with the other screening questions, to generate information on the baseline coding abilities, knowledge of running regression, and knowledge of clustered standard errors. This is important as the relevant question for the RCT is based on a regression with clustered standard errors.

2. On top of the questions designed by the researchers, the original recruitment process also asks multiple-choice questions on coding and knowledge of Econometrics that will also be used as control variables for our analysis as well as in the heterogeneity analysis.

3. The fictitious setting was changed from computer aid learning platforms to teaching at the right level techniques. The list of results from real RCT was also updated to reflect this change.

REFERENCES

Banerjee, A., Banerji, R., Berry, J., Duflo, E., Kannan, H., Mukherji, S., Shotland, M. and Walton, M. (2016). Mainstreaming an effective intervention: Evidence from randomized evaluations of "teaching at the right level" in india, *Technical report*, National Bureau of Economic Research.

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.

Cabezas, V., Cuesta, J. I. and Gallego, F. (2011). Effects of short-term tutoring on cognitive and non-cognitive skills: Evidence from a randomized evaluation in chile, *J-PAL Working Paper* .

Duflo, E., Dupas, P. and Kremer, M. (2011). Peer effects, teacher incentives, and the impact of tracking: Evidence from a randomized evaluation in kenya, *American economic review* **101**(5): 1739–1774.