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

Bruno Ferman and Lucas Finamor

(EESP-FGV)

Motivation


Tatyana Deryugina
@TDeryugina
Me: these results look so strange, better set aside a few hours to figure out what's going on.
...5 minutes later...
f&$k, I didn't specify the regression correctly!
2:37 AM · Oct 7, 2021 · Twitter Web App


Susan Athey replied
Morning Brew @MorningB... · 8h
Write a horror story for your industry using just four words
3,830
7,987
1,421
Susan Athey @Susan_Athey · 1m
Found bug, results changed
2


Rachael Meager
@economeager
daily reminder that the best time to make a histogram of your key outcome variable is the day you get the data, and the second best time is when you see a regression result and think "....huh....".
10:57 AM · Jul 14, 2021 · Twitter Web App

Motivation

- Economics rely heavily on coding
- ... coding errors are common

Motivation

- Economics rely heavily on coding
- ... coding errors are common
- ... even in published work

  Brodeur et al. (2024) finds coding errors in 1/4 of papers published in leading journals

Motivation

- Economics rely heavily on coding
- ... coding errors are common
- ... even in published work

  Brodeur et al. (2024) finds coding errors in 1/4 of papers published in leading journals

- **Are coding errors more likely to be detected when they produce unexpected results?**

Motivation

Are coding errors more likely to be detected when they produce unexpected results?

- If **no**, they can still affect inference procedure
  - Non-standard errors literature
    Simmons et al. (2011), Menkveld et al. (2024), Silberzahn et al. (2018), Landy et al. (2020), Huntington-Klein et al. (2021, 2025)
  - Excess degrees of freedom in data collection and analysis

Motivation

Are coding errors more likely to be detected when they produce unexpected results?

- If **no**, they can still affect inference procedure
  - Non-standard errors literature
    Simmons et al. (2011), Menkveld et al. (2024), Silberzahn et al. (2018), Landy et al. (2020), Huntington-Klein et al. (2021, 2025)
  - Excess degrees of freedom in data collection and analysis
- If **yes**, they additionally bias our results in the direction of researcher's priors, expectations and/or judgments

This paper

- We embed an RCT in a data task used to screen research assistants at the World Bank

- Experimentally vary whether coding errors lead to expected/unexpected results

This paper

- We embed an RCT in a data task used to screen research assistants at the World Bank

- Experimentally vary whether coding errors lead to expected/unexpected results

- **Yes!** Debugging probability depend on the nature of results

$\Rightarrow$ Individuals are 20% more likely to find an error if it leads to an unexpected result

Literature

- Reproducibility and Replicability of research

  Ankel-Peters et al. (2023), many ... ,Wood et al. (2018)

- P-hacking, research questionable practices and publication bias

  Adda et al. (2020), many ... ,Vivalt (2019)

- Excessive degrees of freedom in data collection and analysis

  Huntington-Klein et al. (2021, 2025), Landy et al. (2020), Menkveld et al. (2024), Silberzahn et al. (2018), Simmons et al. (2011)

- Non-standard errors

  Huntington-Klein et al. (2021, 2025), Menkveld et al. (2024), Silberzahn et al. (2018)

Contribution

- Experimentally show that coding errors are more likely to be detected when they lead to unexpected results
- New source of variability that can not only increase the variance but also bias results
- Researcher act that is not necessarily intentional

Experiment

Context

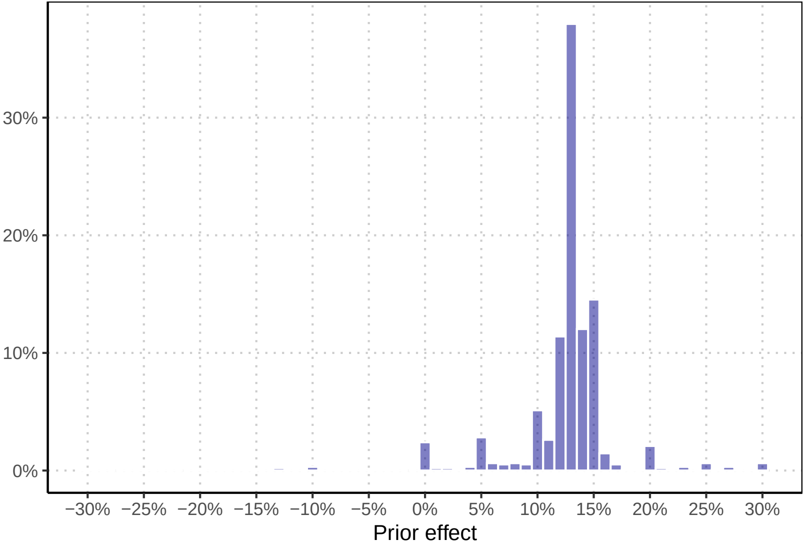
- World Bank recruitment process (2024–2025)
- Coding task for RA/Fellowship applicants
- 1,036 participants, randomized into treatment/control

Data Task

- Showed 5 results from RCTs implementing teaching at the right level, with results 0.08–0.16 sd.

- Elicited priors of what to expect of a similar RCT inspired by the literature

Priors



Data Task

- Presented a hypothetical RCT implemented following the literature

Data Task

- Presented a hypothetical RCT implemented following the literature
- Two datasets (implementation in different states) + data dictionary
- 4 coding questions

  Q1 Descriptive statistics
  Q2 Balance of age of students (regression)
  Q3 Treatment effect of the RCT in state 1 (regression)
  Q4 Treatment effect of the RCT in state 2 (regression)

  - Q1-Q2 screening for coding abilities
  - Q3-Q4 experiment

Experiment

- Outcome variable had 99 recording missing values

Experiment

- Outcome variable had 99 recording missing values

- If not taking that into account:
  - 50% test-takers (control) would estimate a **positive** significant effect
  - 50% test-takers (treated) would estimate a **negative** significant effect

- For both groups the true effect was zero

	Treated	Control
Panel A. Keeping missing values (99)		
Question 3	-0.1602^{***}	0.1488^{***}
Panel B. Removing missing values (99)		
Question 3	-0.0042	0.0068

Experimental Design

- As soon as they answer results for Q3 we store their answers

Experimental Design

- As soon as they answer results for Q3 we store their answers
- Q4 flip the results

	Treated	Control
Panel A. Keeping missing values (99)		
Question 3	-0.1602^{***}	0.1488^{***}
Question 4	0.1488^{***}	-0.1602^{***}
Panel B. Removing missing values (99)		
Question 3	-0.0042	0.0068
Question 4	0.0068	-0.0042

Why?

- Why flipping the results in Q4?

- Being ex-post fair. Both groups experience the same results and only final answers are considered for recruitment process

- It also allows us to explore the behavior of the control group in Q4

Identification and Empirical Strategy

Latent types

1. Always-spot (AS): those who always spot the error, irrespectively of the result;

2. Never-spot (NS): those who never spot the error, irrespectively of the result;

Latent types

1. Always-spot (AS): those who always spot the error, irrespectively of the result;
2. Never-spot (NS): those who never spot the error, irrespectively of the result;
3. Complier I (CI): those who spot the coding error if it leads to unexpected results;
4. Complier II (CII): those who spot the coding error if they find conflicting results between the two answers

Identification

Control

		Q4	
		Spotted	Not Spotted
Q3	Spotted	AS	-
	Not Spotted	CI, CII	NS

Treated

		Q4	
		Spotted	Not Spotted
Q3	Spotted	AS, CI	-
	Not Spotted	CII	NS

Identification

Control

Q3 \ Q4	Spotted	Not Spotted
Spotted	AS	-
Not Spotted	CI, CII	NS

Treated

Q3 \ Q4	Spotted	Not Spotted
Spotted	AS, CI	-
Not Spotted	CII	NS

Identification

Control

Q3 \ Q4	Spotted	Not Spotted
Spotted	AS	-
Not Spotted	CI, CII	NS

Treated

Q3 \ Q4	Spotted	Not Spotted
Spotted	AS, CI	-
Not Spotted	CII	NS

Empirical Strategy

- Using Q3

$$Y_i^{Q3} = \alpha_1 + \beta_1 T_i + \gamma_1 Z_i + \epsilon_i$$

- Using Q4

$$Y_i^{Q4} = \alpha_2 + \beta_2(1 - T_i) + \gamma_2 Z_i + \varepsilon_i$$

- As β_1 and β_2 identify the same parameter, we can optimally combine the two estimators:
  - Optimally choosing weights (given their variances)
  - GMM

Results

Variable	Overall Mean	Control Mean	Treated Mean	Diff	P-value Diff	N Obs
Female	0.377	0.352	0.402	0.051	0.095	1025
	(0.485)	(0.478)	(0.491)	[0.030]		
Master	0.937	0.927	0.947	0.019	0.203	1031
	(0.243)	(0.260)	(0.225)	[0.015]		
Econometrics	0.864	0.859	0.868	0.009	0.663	1034
	(0.343)	(0.348)	(0.338)	[0.021]		
Stata	0.508	0.524	0.491	-0.033	0.295	1036
	(0.500)	(0.500)	(0.500)	[0.031]		
R	0.295	0.287	0.305	0.018	0.526	1036
	(0.456)	(0.453)	(0.461)	[0.028]		
Python	0.159	0.150	0.169	0.019	0.403	1036
	(0.366)	(0.357)	(0.375)	[0.023]		

Variable	Overall Mean	Control Mean	Treated Mean	Diff	P-value Diff	N Obs
Score Q1-Q2	4.446	4.438	4.454	0.016	0.891	991
	(1.828)	(1.850)	(1.807)	[0.116]		
Coding Score	0.055	0.000	0.110	0.110	0.109	886
	(1.025)	(0.998)	(1.049)	[0.069]		
Impact Evaluation Score	0.032	0.000	0.064	0.064	0.328	949
	(1.009)	(1.000)	(1.018)	[0.066]		
Prior effect	0.126	0.127	0.125	-0.002	0.466	947
	(0.046)	(0.046)	(0.047)	[0.003]		
P-value joint test					0.463	

Main Results

Estimator	$\hat{\beta}_1$ (1)	$\hat{\beta}_2$ (2)	$\hat{\beta}_{\text{combined}}$ (3)	$\hat{\beta}_{\text{GMM}}$ (4)
Panel A - Controls for wave				
Treat	0.0103	0.0146	0.0141	0.0141
(s.e.)	(0.0199)	(0.0077)	(0.0073)	(0.0073)
[p-value]	[0.6053]	[0.0596]	[0.0514]	[0.0546]
{p-value2}	{0.3026}	{0.0298}	{0.0278}	{0.0273}
N Obs	788	788	788	788
Intercept	0.078	0.005	-	-

Main Results

Estimator	$\hat{\beta}_1$ (1)	$\hat{\beta}_2$ (2)	$\hat{\beta}_{\text{combined}}$ (3)	$\hat{\beta}_{\text{GMM}}$ (4)
Panel A - Controls for wave				
Treat	0.0103	0.0146	0.0141	0.0141
(s.e.)	(0.0199)	(0.0077)	(0.0073)	(0.0073)
[p-value]	[0.6053]	[0.0596]	[0.0514]	[0.0546]
{p-value2}	{0.3026}	{0.0298}	{0.0278}	{0.0273}
N Obs	788	788	788	788
Intercept	$\boxed{0.078}$	0.005	-	-

Main Results

Estimator	$\hat{\beta}_1$ (1)	$\hat{\beta}_2$ (2)	$\hat{\beta}_{combined}$ (3)	$\hat{\beta}_{GMM}$ (4)
Panel A - Controls for wave				
Treat	$\boxed{0.0103}$	0.0146	0.0141	0.0141
(s.e.)	(0.0199)	(0.0077)	(0.0073)	(0.0073)
[p-value]	[0.6053]	[0.0596]	[0.0514]	[0.0546]
{p-value2}	{0.3026}	{0.0298}	{0.0278}	{0.0273}
N Obs	788	788	788	788
Intercept	0.078	0.005	-	-

Main Results

Estimator	$\hat{\beta}_1$ (1)	$\hat{\beta}_2$ (2)	$\hat{\beta}_{\text{combined}}$ (3)	$\hat{\beta}_{\text{GMM}}$ (4)
Panel A - Controls for wave				
Treat	0.0103	0.0146	0.0141	0.0141
(s.e.)	(0.0199)	(0.0077)	(0.0073)	(0.0073)
[p-value]	[0.6053]	[0.0596]	[0.0514]	[0.0546]
{p-value2}	{0.3026}	{0.0298}	{0.0278}	{0.0273}
N Obs	788	788	788	788
Intercept	0.078	0.005	-	-

Main Results

Estimator	$\hat{\beta}_1$ (1)	$\hat{\beta}_2$ (2)	$\hat{\beta}_{\text{combined}}$ (3)	$\hat{\beta}_{\text{GMM}}$ (4)
Panel B - Adding demographics controls				
Treat	0.0100	0.0157	0.0150	0.0149
(s.e.)	(0.0197)	(0.0080)	(0.0075)	(0.0077)
[p-value]	[0.6126]	[0.0508]	[0.0400]	[0.0525]
{p-value2}	{0.3063}	{0.0254}	{0.0213}	{0.0262}
N Obs	788	788	788	788
Intercept	0.078	0.005	-	-

Main Results

Estimator	$\hat{\beta}_1$ (1)	$\hat{\beta}_2$ (2)	$\hat{\beta}_{\text{combined}}$ (3)	$\hat{\beta}_{\text{GMM}}$ (4)
Panel C - Adding data test variables				
Treat	0.0111	0.0145	0.0141	0.0141
(s.e.)	(0.0198)	(0.0079)	(0.0075)	(0.0077)
[p-value]	[0.5738]	[0.0674]	[0.0564]	[0.0664]
{p-value2}	{0.2869}	{0.0337}	{0.0308}	{0.0332}
N Obs	788	788	788	788
Intercept	0.078	0.005	-	-

Main Results

Estimator	$\hat{\beta}_1$ (1)	$\hat{\beta}_2$ (2)	$\hat{\beta}_{\text{combined}}$ (3)	$\hat{\beta}_{\text{GMM}}$ (4)
Panel D - Adding screening test variables				
Treat	0.0079	0.0130	0.0124	0.0124
(s.e.)	(0.0198)	(0.0076)	(0.0072)	(0.0074)
[p-value]	[0.6878]	[0.0864]	[0.0792]	[0.0949]
{p-value2}	{0.3439}	{0.0432}	{0.0380}	{0.0475}
N Obs	788	788	788	788
Intercept	0.078	0.005	-	-

Additional Exercises

Sample	Qualified	All	Non-negative Prior	Correct Prior
	(1)	(2)	(3)	(4)
Effect	0.0141	0.0097	0.0141	0.0154
(s.e.)	(0.0073)	(0.0065)	(0.0073)	(0.0082)
[p-value]	[0.0514]	[0.1237]	[0.0498]	[0.0541]
{p-value2}	{0.0278}	{0.0615}	{0.0289}	{0.0299}
N Obs	788	944	785	697
Spotted First	0.078	0.068	0.078	0.085
Spotted Second	0.005	0.007	0.005	0.006

Heterogeneity

	Effect	Std Error	P-value	N Obs	Spotted First	Diff P-value
Benchmarks						
Entire Sample	0.0097	(0.0065)	[0.1237]	944	0.068	-
1. Clustered SE?						
No	-0.0079	(0.0109)	[0.3510]	155	0.014	0.0981
Yes	0.0138	(0.0073)	[0.0552]	789	0.077	
2. Score in Q1–Q2						
Below Median	0.0060	(0.0088)	[0.4852]	452	0.039	0.5384
Above Median	0.0137	(0.0089)	[0.0987]	489	0.096	
3. Coding Score						
Below Median	-0.0021	(0.0100)	[0.8312]	405	0.062	0.3205
Above Median	0.0112	(0.0089)	[0.2089]	432	0.097	
4. Impact Evaluation Score						
Below Median	0.0043	(0.0107)	[0.6709]	423	0.087	0.6365
Above Median	0.0107	(0.0083)	[0.1982]	454	0.063	

Heterogeneity

	Effect	Std Error	P-value	N Obs	Spotted First	Diff P-value
Benchmarks						
Entire Sample	0.0097	(0.0065)	[0.1237]	944	0.068	-
1. Clustered SE?						
No	-0.0079	(0.0109)	[0.3510]	155	0.014	0.0981
Yes	0.0138	(0.0073)	[0.0552]	789	0.077	
2. Score in Q1–Q2						
Below Median	0.0060	(0.0088)	[0.4852]	452	0.039	0.5384
Above Median	0.0137	(0.0089)	[0.0987]	489	0.096	
3. Coding Score						
Below Median	-0.0021	(0.0100)	[0.8312]	405	0.062	0.3205
Above Median	0.0112	(0.0089)	[0.2089]	432	0.097	
4. Impact Evaluation Score						
Below Median	0.0043	(0.0107)	[0.6709]	423	0.087	0.6365
Above Median	0.0107	(0.0083)	[0.1982]	454	0.063	

Heterogeneity

	Effect	Std Error	P-value	N Obs	Spotted First	Diff P-value
Benchmarks						
Entire Sample	0.0097	(0.0065)	[0.1237]	944	0.068	-
1. Clustered SE?						
No	-0.0079	(0.0109)	[0.3510]	155	0.014	0.0981
Yes	0.0138	(0.0073)	[0.0552]	789	0.077	
2. Score in Q1–Q2						
Below Median	0.0060	(0.0088)	[0.4852]	452	0.039	0.5384
Above Median	0.0137	(0.0089)	[0.0987]	489	0.096	
3. Coding Score						
Below Median	-0.0021	(0.0100)	[0.8312]	405	0.062	0.3205
Above Median	0.0112	(0.0089)	[0.2089]	432	0.097	
4. Impact Evaluation Score						
Below Median	0.0043	(0.0107)	[0.6709]	423	0.087	0.6365
Above Median	0.0107	(0.0083)	[0.1982]	454	0.063	

Heterogeneity

	Effect	Std Error	P-value	N Obs	Spotted First	Diff P-value
Benchmarks						
Entire Sample	0.0097	(0.0065)	[0.1237]	944	0.068	-
1. Clustered SE?						
No	-0.0079	(0.0109)	[0.3510]	155	0.014	0.0981
Yes	0.0138	(0.0073)	[0.0552]	789	0.077	
2. Score in Q1–Q2						
Below Median	0.0060	(0.0088)	[0.4852]	452	0.039	0.5384
Above Median	0.0137	(0.0089)	[0.0987]	489	0.096	
3. Coding Score						
Below Median	-0.0021	(0.0100)	[0.8312]	405	0.062	0.3205
Above Median	0.0112	(0.0089)	[0.2089]	432	0.097	
4. Impact Evaluation Score						
Below Median	0.0043	(0.0107)	[0.6709]	423	0.087	0.6365
Above Median	0.0107	(0.0083)	[0.1982]	454	0.063	

Heterogeneity

	Effect	Std Error	P-value	N Obs	Spotted First	Diff P-value
Benchmarks						
Entire Sample	0.0097	(0.0065)	[0.1237]	944	0.068	-
5. Master's degree						
Yes	0.0109	(0.0069)	[0.1094]	888	0.069	-
6. Econometrics course						
Yes	0.0111	(0.0074)	[0.1290]	830	0.076	-
7. Gender						
Men	0.0083	(0.0071)	[0.1267]	578	0.074	0.6713
Women	0.0144	(0.0125)	[0.2453]	356	0.058	

Dicussion and Conclusion

Conclusion

- Coding errors are prevalent
- Error detection is not a neutral process: it depends on the nature of the results they generate
- Individuals are 20% more likely to identify coding errors if they lead to unexpected results
- More training, experience, and skills do not make individuals less vulnerable to the bias
- This probably extends to *favorable/not-favorable* results (and not only expected/unexpected results)
  - Significant results being more likely to be published
  - Placebo results where you do not expect an effect

Appendix

Presentation

- Introduction
- Experiment
- Identification and Empirical Strategy
- Results
- Conclusion

Non-response

Variable	Control Mean	Treated Mean	Diff	P-value Diff	N Obs
Female	0.008	0.014	0.006 [0.006]	0.336	1036
Master	0.006	0.004	-0.002 [0.004]	0.682	1036
Econometrics	0.004	0.000	-0.004 [0.003]	0.157	1036
Coding Score	0.161	0.128	-0.034 [0.022]	0.124	1036
Impact Evaluation Score	0.095	0.073	-0.022 [0.017]	0.197	1036
Prior effect	0.085	0.086	0.001 [0.017]	0.952	1036

By wave

Sample	All (1)	Wave 2024 (2)	Wave 2025
Effect	0.0141	0.0172	0.0093
(s.e.)	(0.0073)	(0.0077)	(0.0135)
[p-value]	[0.0514]	[0.0273]	[0.4652]
{p-value2}	{0.0278}	{0.0224}	{0.2285}
N Obs	788	450	338

Heterogeneities - qualified sample

	Effect	Std Error	P-value	N Obs	Spotted First	Diff P-value
Benchmarks						
Entire Sample	0.0097	(0.0065)	[0.1237]	944	0.068	-
Qualified Sample	0.0141	(0.0073)	[0.0514]	788	0.078	-
1. Clustered SE?						
Yes	0.0145	(0.0077)	[0.052]	747	0.081	
2. Score in Q1–Q2						
Below Median	0.0131	(0.0093)	[0.1763]	390	0.043	0.9441
Above Median	0.0141	(0.0108)	[0.1461]	398	0.113	
3. Coding Score						
Below Median	0.0035	(0.0100)	[0.7003]	362	0.074	0.4690
Above Median	0.014	(0.0105)	[0.2005]	362	0.102	

Heterogeneities - qualified sample

	Effect	Std Error	P-value	N Obs	Spotted First	Diff P-value
4. Impact Evaluation Score						
Below Median	0.0154	(0.0121)	[0.1637]	365	0.087	1.3871
Above Median	0.0078	(0.0089)	[0.4067]	371	0.063	
5. Master's degree						
Yes	0.0154	(0.0077)	[0.0454]	750	0.078	-
6. Econometrics course						
Yes	0.0156	(0.0081)	[0.0476]	714	0.084	-
7. Gender						
Men	0.0101	(0.0085)	[0.2128]	477	0.089	0.4519
Women	0.0221	(0.0135)	[0.1055]	304	0.061	

Latenty types expanded

Complier II (CII): those who spot the coding error if they find conflicting results between the two answers

- CII-A: spot the error if they find conflicting results between the two answers irrespectively of their signs
- CII-B: spot the error if they find conflicting results between the two answers, if the first is the positive and the second is the negative one. But not the other way around.
- CII-C: spot the error if they find conflicting results between the two answers, if the first is the negative and the second is the positive one. But not the other way around.

Latenty types expanded

Control

		Q4: Spotted	Q4: Not Spotted
Q3	Spotted	AS	-
	Not Spotted	CI, CII-A, CII-B	NS, CII-C

Treated

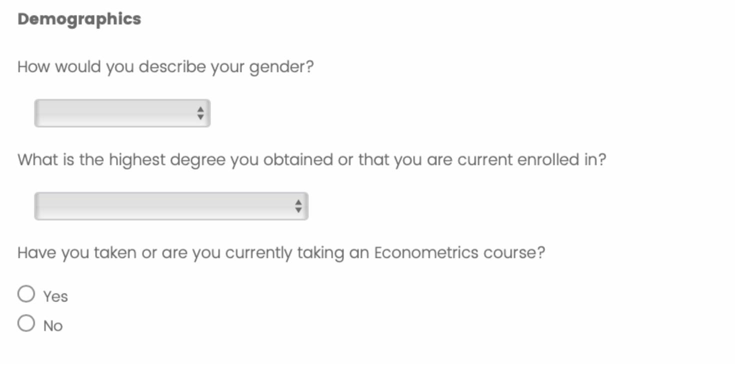
		Q4: Spotted	Q4: Not Spotted
Q3	Spotted	AS, CI	-
	Not Spotted	CII-A, CII-C	NS, CII-B

Latenty types expanded

$$\begin{aligned}\Delta_2 &= (\text{CI} + \text{CII-A} + \text{CII-B}) - (\text{CII-A} - \text{CII-C})) \\ &= \text{CI} + (\text{CII-B} - \text{CII-C}).\end{aligned}$$

- CI proportion if these two types have the same proportion (or do not exist)
- If CII-B $<$ CII-C, that is, it is more likely to spot the error after seeing negative-positive, than positive-negative results, then we would underestimate our target parameter
- If CII-B $>$ CII-C, we would overestimate the proportion of CI types. Note however, that this difference is also manifested by individuals with differential debugging probabilities based on the results they face, exactly what we want to test with this RCT.

Figure 1



Demographics

How would you describe your gender?

What is the highest degree you obtained or that you are current enrolled in?

Have you taken or are you currently taking an Econometrics course?

- Yes
- No

Figure 2

Initial information

The data task can be completed using any software/ programming language. Please, make sure you are using a computer where you can open the software of your choice and analyzed small datasets.

On the next page, we will provide a description of the problem and the data files for download.
Please, download the files to your computer.

Note that a summary of the problem and the link to the data will also be displayed on all subsequent pages.

Even though you cannot go back to a previous page, at the end, you will be able to review and edit all your answers. Only the final answers will be graded.

For the test you will be required to open two small csv files (around 700kb each), manipulate the data and perform simple econometric analysis. In which language do you intend to execute there analysis?

- ○ Stata
- ○ R
- ○ Python
- ○ [ ] Other

Figure 3

Problem Description

Learning levels in primary school are a challenge in several developing countries. Students with low proficiency in math and language create challenges for teachers and peers.
To address this pressing problem, several initiatives have tried to tailor content to student's appropriate levels. This can take several formats, such as remedial education for selected students, tutoring, adaptation of material, dividing students into groups with similar proficiency, and many other activities.

Researchers evaluated these interventions in several developing countries using randomized control protocols. The results of selected studies are presented below:

For the purpose of this test, you are not required, and there is no benefit of knowing or reading any of these studies.

Figure 4

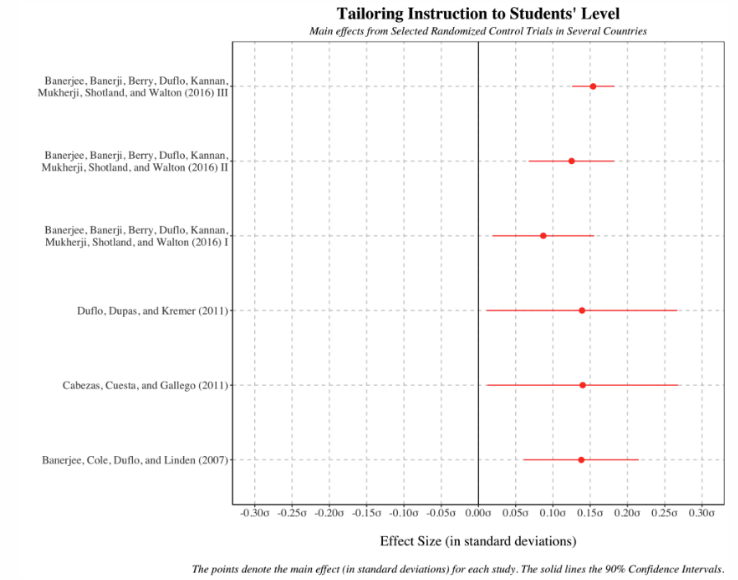

Tailoring Instruction to Students' Level
Main effects from Selected Randomized Control Trials in Several Countries
Banerjee, Banerji, Berry, Duflo, Kannan, Mukherji, Shotland, and Walton (2016) III
Banerjee, Banerji, Berry, Duflo, Kannan, Mukherji, Shotland, and Walton (2016) II
Banerjee, Banerji, Berry, Duflo, Kannan, Mukherji, Shotland, and Walton (2016) I
Duflo, Dupas, and Kremer (2011)
Cabezas, Cuesta, and Gallego (2011)
Banerjee, Cole, Duflo, and Linden (2007)
-0.30σ -0.25σ -0.20σ -0.15σ -0.10σ -0.05σ 0.00σ 0.05σ 0.10σ 0.15σ 0.20σ 0.25σ 0.30σ
Effect Size (in standard deviations)
The points denote the main effect (in standard deviations) for each study. The solid lines the 90% Confidence Intervals.

Figure 5

This box will be displayed on all screens

You are required to analyze data from a fictitious randomized control trial (RCT) that, inspired by the literature reviewed above, tailored instructions to students in 5th grade in a developing country.

The experiment took place in two different states within the same country: State 1 and State 2, which are very much alike.

In each state, 480 schools were invited and accepted to participate in the program. Each school was randomly allocated to treatment or control.

In treated schools, 5th-grade teachers received material and training to implement the tailored program.

In control schools, instruction for 5th-grade students remained the same.

You will access data that includes student-level data on language scores administered 12 months after the program started.

Figure 6

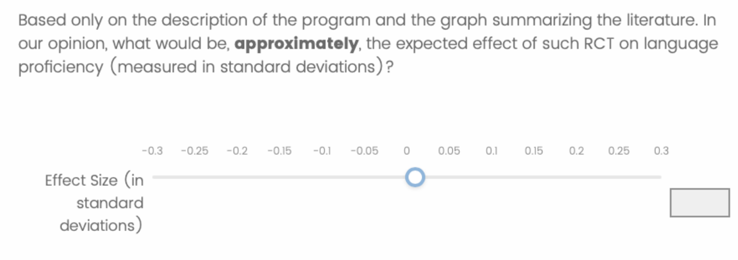

Based only on the description of the program and the graph summarizing the literature. In our opinion, what would be, **approximately**, the expected effect of such RCT on language proficiency (measured in standard deviations)?
-0.3 -0.25 -0.2 -0.15 -0.1 -0.05 0 0.05 0.1 0.15 0.2 0.25 0.3
Effect Size (in standard deviations)

Figure 7

Data

You have access to the data (csv files) for state 1 (`data_state_1.csv`) and for state 2 (`data_state_2.csv`), and the codebook (txt file). Download them here: Data

You should use 3 decimal digits, rounding the numbers either up or down. Please use a dot as the decimal separator.

- The number 1.8765980 should be inputted as 1.876 or 1.877.
- The number 98.31% should be inputted as 0.983 or 0.984.

Input all answers here to be graded. Even though you cannot go back to a previous page, at the end, you will be able to review and EDIT all your answers. Only the final answers are scored.

Figure 8

Question 1 – For this question, <u>consider only the data for State 1</u>, that is, the file `data_state_1.csv`

Remember:

- The number 1.8765980 should be inputted as 1.876 or 1.877.
- The number 98.31% should be inputted as 0.983 or 0.984.

A) What is the proportion of female students?

B) Among male students in treated schools, what is the average age?

C) Find the school with the largest number of students aged 12 years old. What is the number of 12 year-old students in this school?

Figure 9

Question 2 - For this question, <u>consider only the data for State 1</u>, that is, the file data_state_1.csv

Remember:

- The number 1.8765980 should be inputted as 1.876 or 1.877.
- The number 98.31% should be inputted as 0.983 or 0.984.

We are interested in assessing the balance of the experiment, that is, whether treated and control studets have similar age. Estimate the balance according to these instructions:

- Run a regression of age (variable AGE) on a constant and the treatment indicator (variable TREATED).
- <u>You should cluster your standard errors at the school level</u>

Now, answer the following questions regarding the coefficient for the treatment indicator (variable TREATED).

A) What is the estimated coefficient?

B) What is the standard error?

C) What is the p-value?

Figure 10

Question 3

We are interested in estimating the effect of the program on language proficiency, comparing students in treated and control schools.

We will start with data from **${e://Field/q1_text}.** Estimate the effects according to these instructions:

- Run a regression of language score (variable `LANGUAGE_SCORE_STD`) on a constant and the treatment indicator (variable `TREATED`).
- The outcome variable `LANGUAGE_SCORE_STD` is already standardized, ranging from -6 to 6
- You should cluster your standard errors at the school level

Now, answer the following questions regarding the coefficient for the treatment indicator (variable `TREATED`).

A) What is the estimated coefficient?

B) What is the standard error?

C) What is the p-value?

Figure 11

Question 4

We are interested in estimating the effect of the program on language proficiency, comparing students in treated and control schools.

Now, we will compute the same results, but using data from **${e://Field/q2_text}.** Estimate the effects according to these instructions:

- Run a regression of language score (variable `LANGUAGE_SCORE_STD`) on a constant and the treatment indicator (variable `TREATED`).
- The outcome variable `LANGUAGE_SCORE_STD` is already standardized, ranging from -6 to 6
- You should cluster your standard errors at the school level

Now, answer the following questions regarding the coefficient for the treatment indicator (variable `TREATED`).

A) What is the estimated coefficient?

B) What is the standard error?

C) What is the p-value?

Figure 12

Review your answers. You are free to edit them. Only these final numbers will be considered.