Sangnier, Marc

Working Paper

A (partial) replication study of Hjort and Poulsen (2019)

I4R Discussion Paper Series, No. 30

Provided in Cooperation with:
The Institute for Replication (I4R)

Suggested Citation: Sangnier, Marc (2023) : A (partial) replication study of Hjort and Poulsen (2019), I4R Discussion Paper Series, No. 30, Institute for Replication (I4R), s.l.

This Version is available at:
https://hdl.handle.net/10419/270965

INSTITUTE for REPLICATION

No. 30
I4R DISCUSSION PAPER SERIES

A (partial) replication study of Hjort and Poulsen (2019)

Marc Sangnier

May 2023

INSTITUTE for REPLICATION

I4R DISCUSSION PAPER SERIES

I4R DP No. 30

A (partial) replication study of Hjort and Poulsen (2019)

Marc Sangnier[1]

[1]*Aix-Marseille University, Marseille/France*

MAY 2023

E-Mail: joerg.peters@rwi-essen.de
RWI – Leibniz Institute for Economic Research
Hohenzollernstraße 1-3
45128 Essen/Germany
www.i4replication.org

ISSN: 2752-1931

A (partial) replication study of Hjort and Poulsen (2019)[1]

Marc SANGNIER[2]

October 2022

Abstract

Hjort and Poulsen (2019) examine how fast Internet affects employment in Africa. Their difference-in-differences estimates exploit differences in the time at which locations were connected to the network of fast Internet cables. The authors find that fast Internet increases employment rates and that this effect is driven by high-skilled occupations. Authors show that, if anything, employment inequality falls when fast Internet becomes available. This study uses replication materials made available with the original article. It first attempts to reproduce results of the original paper from available replication materials. Most results are reproducible, but some are not. Second, this study presents a sensitivity analysis that tests how reported results vary depending on whether a specific country (or region) is excluded from the sample. The paper's results are found to be differently sensitive to the composition of the sample of observations. This analysis also helps to uncover that some specifications that use a large number of fixed effects might actually be too demanding for reasonable identification to be achieved from the data.

[1] This study has benefited from the financial support of Open Philanthropy. The project leading to this publication has received funding from the French government under the "France 2030" investment plan managed by the French National Research Agency (reference :ANR-17-EURE-0020) and from Excellence Initiative of Aix-Marseille University - A*MIDEX.
[2] Aix-Marseille University (Aix-Marseille School of Economics). Aix Marseille Univ, CNRS, AMSE, Marseille, France. Email: marc.sangnier@univ-amu.fr.

1 Introduction

Hjort and Poulsen (2019), henceforth HP, examine how fast Internet affects employment in Africa. Their difference-in-differences estimates exploit differences in the time at which locations were connected to the network of fast Internet cables.

HP uses several series of data: information about Internet cables and network connections in Africa, Demographic and Health Survey (DHS), the Afrobarometer, the South Africa Quarterly Labor Force Survey (SA-QLFS), data from the South Africa Companies and Intellectual Property Commission, the Ethiopia Large and Medium Scale Manufacturing Industries Survey, light density at night and the World Bank Enterprise Survey. The paper's results are obtained from individual (or firm) observations that are compared across or within countries depending on whether their location is connected to the network of fast Internet.

The paper's main results are (see HP, page 1034): (i) that "the probability that an individual is employed increases by [3.1 to 13.2 percent] when fast Internet becomes available"; (ii) that "the probability of being employed in a position belonging to a skilled occupation increases substantially, but the probability of holding an unskilled job is statistically unaffected when fast Internet becomes available", so that "fast Internet appears to shift employment shares towards higher-productivity occupations"; (iii) that "employment inequality if anything falls when fast Internet arrives" as "the percentage point increase in the probability of having a job is […] of comparable magnitude for those who only completed primary school and those with secondary or tertiary education".

The present study investigates whether HP results can be reproduced and tests their sensitivity to the exclusion of countries (or regions) from the sample. Both the reproducibility and the sensitivity analysis use original replication materials made available with the original paper. Due to sensitive location information and to access restriction rules, these materials only include final data prepared by the authors for data to which access is not restricted. As a consequence, this study leaves aside results that

could not be made available by the authors in original replication materials and does not cover the full process of constructing final data from raw data.[3]

In terms of reproducibility, most potentially reproducible results reported by HP could be reproduced once minor coding errors were corrected. Only the estimates presented in Table 3, panel B, column (2), Table 5, panel B, columns (1)-(6), and one of the estimates from Table 8, column (7) could not be reproduced from the paper's data and code. Inability to reproduce results of Table 5 questions the paper's second main result about fast Internet shifting employment toward higher-productivity occupations.

The sensitivity analysis shows that some of HP results are sensitive to the exclusion of some countries from the sample of observations. In a preliminary analysis, HP show that connection to fast Internet affects Internet speed and use (Table 2). These results proved sensitive to the exclusion of some countries from the sample. The sensitivity analysis also revealed that identification is possibly restricted to a small number of observations in the most demanding specifications that use a large number of fixed effects. The sensitivity analysis shows that HP results about the impact of fast Internet on employment (Table 3) is sensitive to the inclusion of some countries in the sample. So do some robustness checks (Table 4). The sensitivity analysis also suggests that some of the robustness tests might be too demanding given the distribution of observations across space and time. Results along different levels of education that are used to assess the effect of fast Internet on employment inequality (Table 6) appears moderately sensitive to the exclusion of countries or regions from the sample. Finally, the sensitivity analysis shows that the complementary results about how much fast Internet affects incomes (Table 9) are highly sensitive to the exclusion of a country from the sample.

[3] Authors of the original paper were contacted over the replication process and demonstrated their willingness to help conduct a deeper replication by sharing more detailed data with authorized people.

2 Reproducibility

This section reports about the computational reproducibility of the paper's results. Reproducibility was assessed using the data and codes made available with the original paper.[4] Codes were run using STATA MP 15.

Table 2

Results of Table 2 can be fully reproduced from the paper's data and code. Only the mean of outcome in columns (4) and (5) slightly differs (0.10 in the paper, 0.11 when running code), likely because of reporting or editing issues.

Table 3

Results of Table 3, panel A, column (1) have not been reproduced as they rely on restricted access data.

Results of Table 3, panel A, columns (2) and (3) can be fully reproduced from the paper's data and code. Only the mean of outcome in column (3) slightly differs (0.72 in the paper, 0.71 when running code), likely because of reporting or editing issues.

Results of Table 3, panel B, columns (1), (3) and (4) can be fully reproduced from the paper's data and code. Only the means of outcomes in columns (3) and (4) slightly differ (0.48 and 0.12 in the paper, 0.49 and 0.11 when running code), likely because of reporting issues.

Results of Table 3, panel B, column (2) cannot be reproduced. This regression uses a variable called "*morework*" that is missing from the data. According to HP, this variable is constructed from SA-QLFS variable "*q422more*" that is available from the data. The "*morework*" variable is meant to be a dummy equal to 1 for respondents who "wants to work more". According to the SA-QLFS 2008 questionnaire, "*q422more*" is coded as follows:

[4] Hjort, Jonas, and Poulsen, Jonas. Replication data for: The Arrival of Fast Internet and Employment in Africa. Nashville, TN: American Economic Association [publisher], 2019. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2019-10-12. https://doi.org/10.3886/E113156V1

"Last week, would you have liked to work more hours than you actually worked, provided the extra hours had been paid?"

1 = YES, in the current job
2 = YES, in taking an additional job
3 = YES, in another job with more hours
4 = NO
5 = DON'T KNOW

Tests to reconstruct the "*morework*" variable from the different values of the "*q422more*" variable were non-concluding as it was not possible to find a coding rule that would allow to obtain 0.66 as mean of outcome as reported in the paper, nor to reach the reported number of observations (457,192).

Table 4

Results of Table 4, columns (1) and (2) have not been reproduced as they rely on restricted access data.

Results of Table 4, columns (3) and (4) can be fully reproduced from the paper's data and code after the correction of minor coding errors. First, the code is using a dataset named "*afrobarometer_road_electricity.dta*", while the correct name of the dataset is actually *"afrobarometer_roads_electricity.dta*". Second, the code uses a variable named "*test_med2*" that is missing from the data. Investigation revealed that it can be created as a copy of the "*connected*" variable that is available from the data. Third, the code uses a variable named "*grid_new1*" that is missing from the data. Investigation revealed that is can be replaced by the "*grid10*" variable that is available from the data.

Results of Table 4, columns (4)-(8) can be fully reproduced from the paper's data and code. Only the standard error of the estimate reported in column (6) slightly differs (0.010 in the paper, 0.008 when running code), likely because of reporting or editing issues.

Table 5

Results of Table 5, panel A have not been reproduced as they rely on restricted access data.

Results of Table 5, panel B cannot be reproduced from the paper's data and code. A first and minor issue is that the code creates a series of variables that already exist in the data. Some lines must therefore be neutralized. Once this has been done, the code runs smoothly but estimated coefficients largely differ from those reported in the paper. The table below displays original estimates from Table 5, panel B and estimates returned when running the code.

Table 5, panel B						
	(1)	(2)	(3)	(4)	(5)	(6)
Outcome	Skilled	Unskilled	Highly skilled	Somewhat skilled	Moderately skilled	Unskilled
HP estimates	0.014 (0.006) [0.02]	-0.001 (0.005) [0.84]	0.001 (0.004) [0.80]	0.003 (0.004) [0.45]	0.010 (0.006) [0.10]	-0.001 (0.005) [0.84]
Recalculated estimates	0.012 (0.009) [0.18]	0.010 (0.007) [0.15]	-0.006 (0.006) [0.32]	0.008 (0.005) [0.11]	0.010 (0.009) [0.27]	0.010 (0.007) [0.15]
Standard errors in parentheses, p-values between brackets. P-values added to HP estimates. Recalculated estimates obtained from running the original code and using original data. The numbers of observations are identical for HP and recalculated estimates.						

Results presented in columns (1) and (2) of Table 5, panel B complement those of panel A and are used to support "a positional skill bias of fast Internet in Africa" as "the arrival of fast Internet increases the probability that an individual holds a skilled job" while "the probability of unskilled employment is statistically unaffected". However, recalculated estimates suggest that the effect barely differs between skilled and unskilled jobs.

Results presented in columns (3)-(6) of Table 5, panel B are further used to argue that most of the effect concentrates on moderately skilled jobs. Recalculated estimates suggest that the effect hardly differs between somewhat skilled, moderately skilled and unskilled jobs.

Table 6

Results of Table 6, top panel have not been reproduced as they rely on restricted access data.

Results of Table 6, middle and bottom panels can be fully reproduced from the paper's data and code.

Table 7

Results of Table 7 have not been reproduced as they rely on restricted access data.

Table 8

Results of Table 8, columns (1)-(6) can be fully reproduced from the paper's data and code.

Results of Table 8, column (7) can be partially reproduced from the paper's data and code. The table below displays original estimates from Table 8, column (7) and estimates returned when running the code.

Table 8, column (7)		
	HP estimates	Recalculated estimates
Capital	0.276 (0.018)	0.184 (0.009)
Unskilled	0.337 (0.064)	0.337 (0.069)
Skilled	0.497 (0.043)	0.497 (0.043)
SubmarineCables x connected x Unskilled	-0.176 (0.058)	-0.176 (0.058)
SubmarineCables x connected x Skilled	0.026 (0.033)	0.026 (0.033)
Standard errors in parentheses. Recalculated estimates obtained from running the original code and using original data. The numbers of observations are identical for HP and recalculated estimates.		

The standard error of the recalculated estimate of the "*unskilled*" variable slightly differs (0.064 in the paper, 0.069 when running code), likely because of reporting issues. Larger differences are found for the "*capital*" variable for which the recalculated point estimate and the standard error both differ from values reported in the paper. The point estimate is reduced by about one third and the standard error is halved. However, these differences do not question the interpretation that is made of this estimate in the paper as its magnitude remains comparable.

Table 9

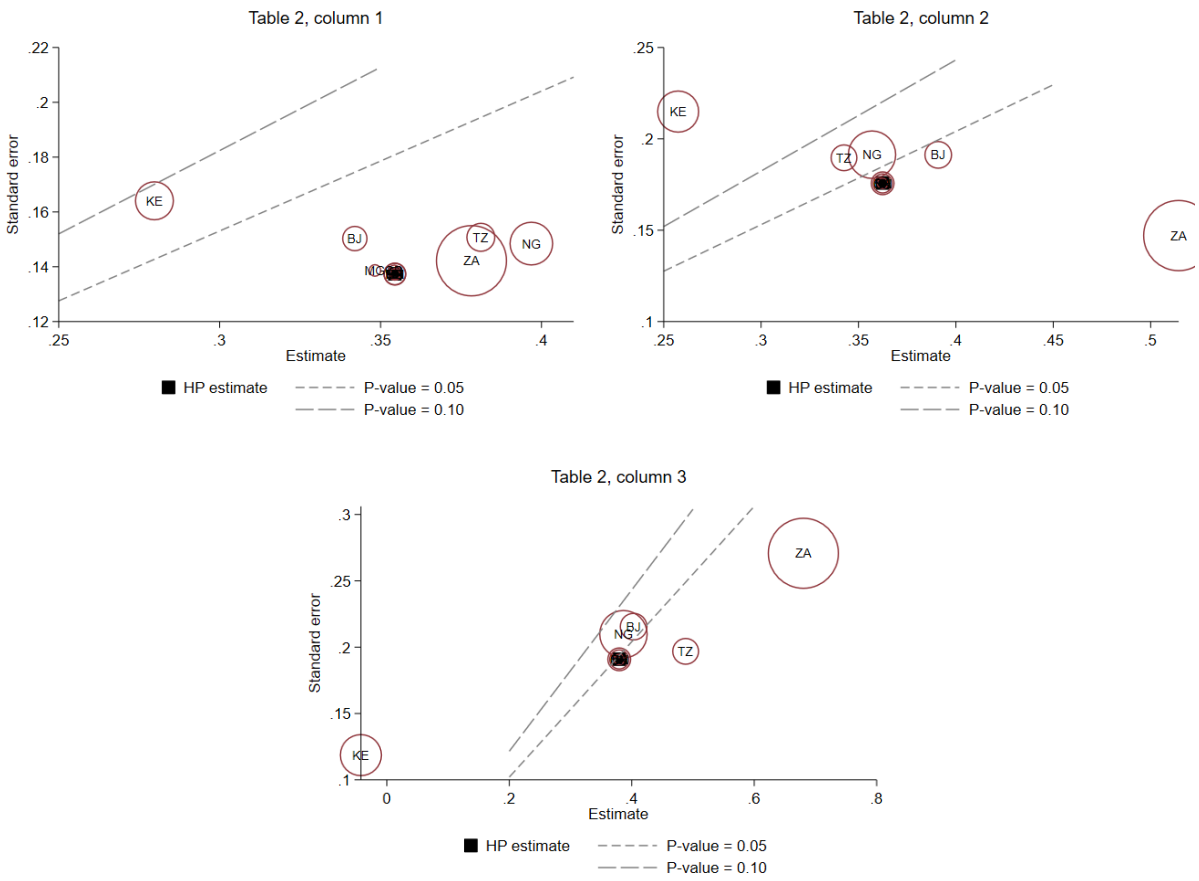
Results of Table 9 can be fully reproduced from the paper's data and code.

3 Sensitivity analysis

As the paper's results rely on a sample made of different countries, a simple sensitivity analysis was conducted by testing how reported results vary when removing each country from the sample. To this end, the code used to test reproducibility of the results was used to re-estimate coefficients of interest while removing countries one by one, keeping track of the sample share that each country represents. The logic of this sensitivity analysis is to test whether results are likely or not to depend on the sample of countries included in the analysis. This sensitivity analysis was conceived after looking at the original programs and it was not pre-registered. The sensitivity analysis was applied to results that could be successfully reproduced (see the Reproducibility section).

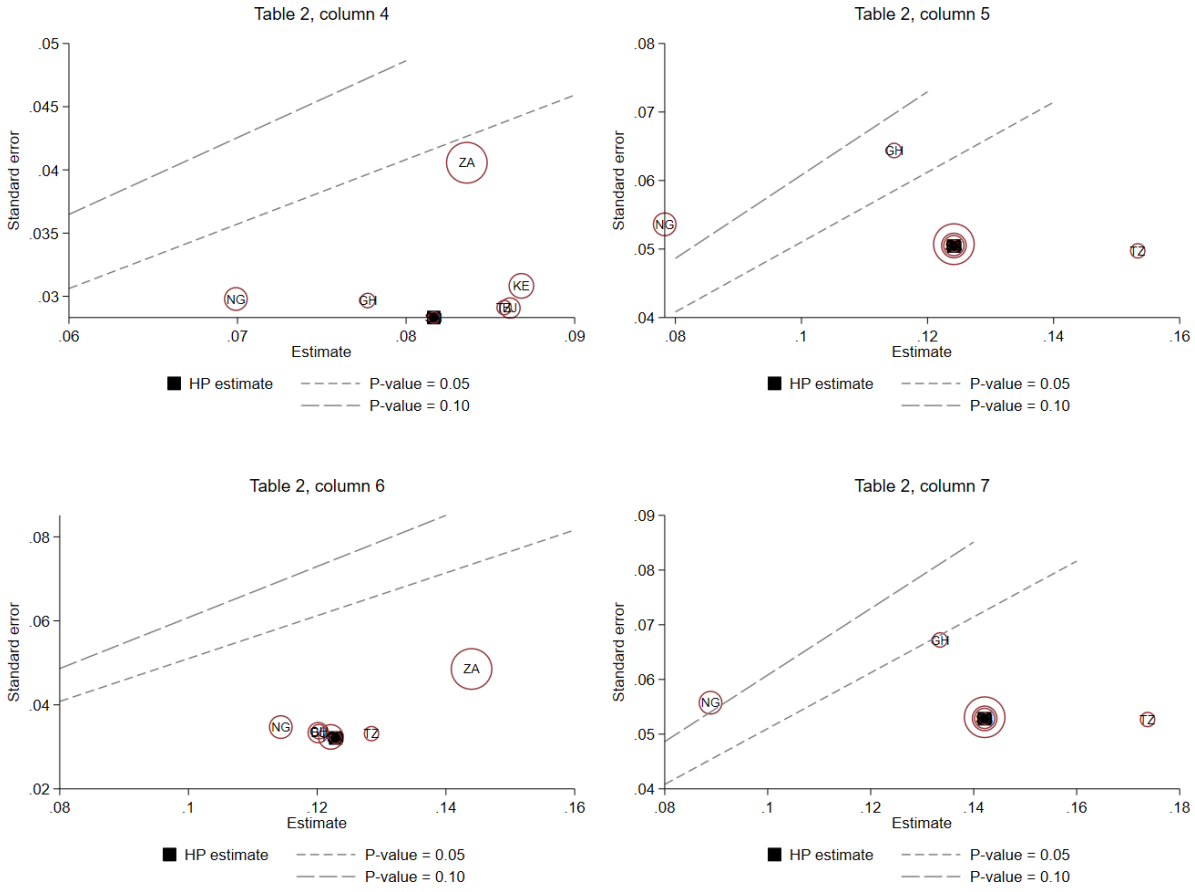
Table 2

The figures below display estimated coefficients of interest from Table 2, columns (1)-(3) when excluding countries one by one from the sample. Circles are proportional to share of each countries' observations in the original sample.



Two observations can be made from the above figures. First, a number of countries do not contribute to the estimation of the coefficient of interest as shown by the fact that removing them from the sample lead to a point estimate that is identical to the one obtained when using the full sample. This is due to the fact that there is no variation in the treatment variable for some countries. For example, there are 5 out of the 12 countries used in Table 2, column (1) for which the treatment variable does not vary. Second, estimates are quite sensitive (both in terms of magnitude and statistical significance) to the exclusion of some countries that represent a large share of the original sample. For example, excluding South Africa from the sample leads to an increase of about 40% for the estimate of Table 2, column (2). In contrast, excluding Kenya lowers the point estimate by about 30% and strongly reduces its statistical significance (the p-value moves from 0.04 to 0.23). Similar comments apply for the estimate of Table 2, column (3).

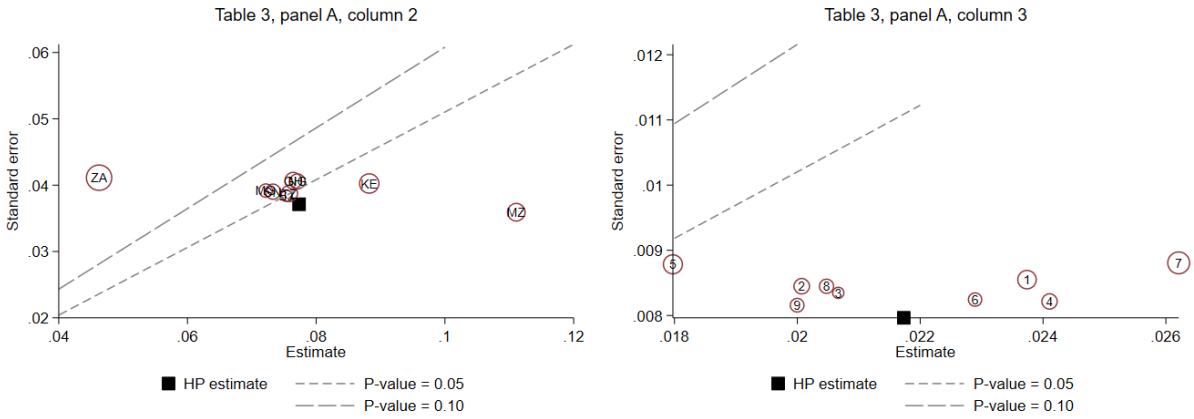
The figures below display estimated coefficients of interest from Table 2, columns (4)-(7) when excluding countries one by one from the sample. Circles are proportional to share of each countries' observations in the original sample.



Estimates obtained for columns (4) and (6) of Table 2 are reasonably sensitive to the exclusion of countries from the sample. In contrast, estimates obtained for columns (5) and (7) exhibit two common patterns. First, a number of countries do not contribute to the estimation of the coefficient of interest as show by estimates' clustering around the full sample estimate. This is likely due to the fact that these columns use demanding *connected x time fixed* effects that mechanically neutralize a large number of observations as they come in addition to *location* and *country x time* fixed effects. Second, the estimate is quite sensitive to the exclusion of 3 countries (Nigeria, Ghana and Tanzania). However, estimated coefficients remain reasonably comparable to the full sample estimate in terms of magnitude and statistical significance.

Table 3

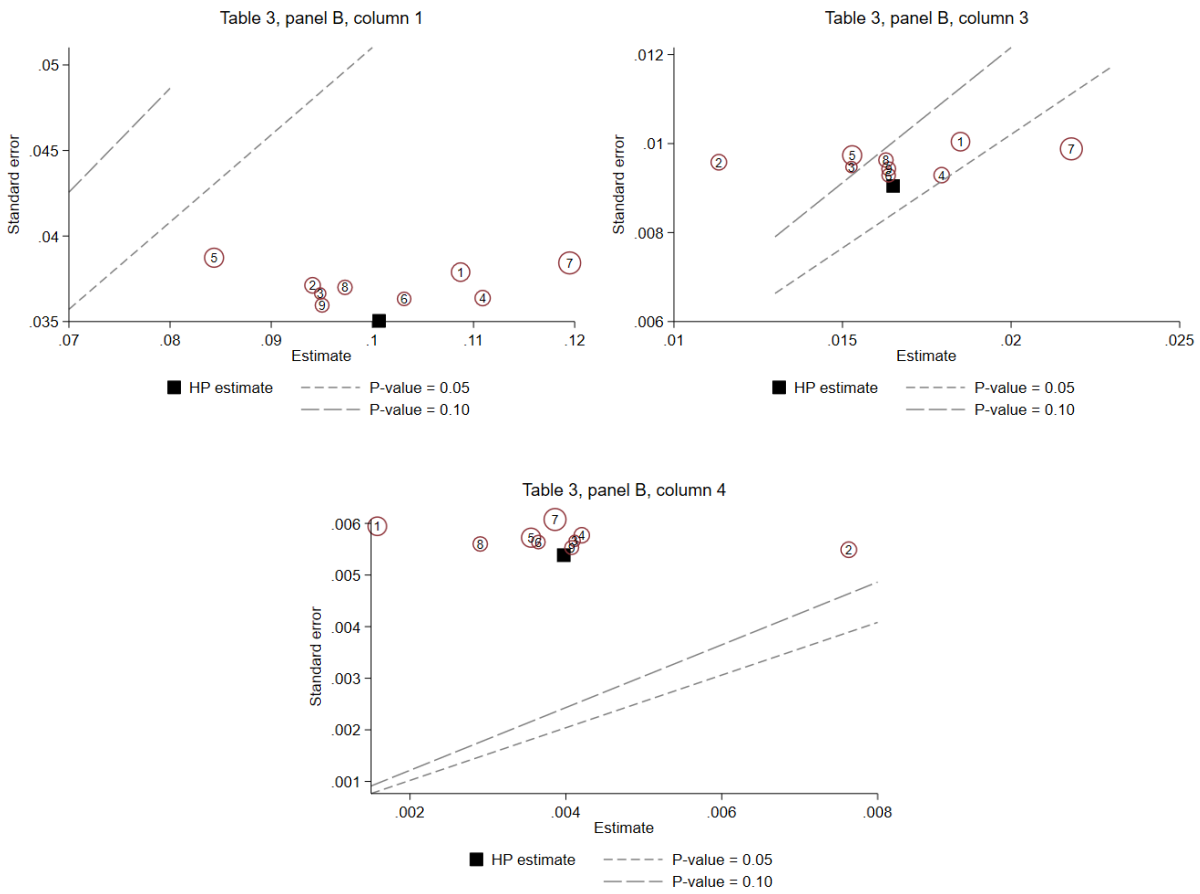
The figures below display estimated coefficients of interest from Table 2, panel A, columns (2) and (3) when excluding countries or regions one by one from the sample. Circles are proportional to share of each countries' observations in the original sample. While column (2) estimates are obtained from Afrobarometer data, column (3) estimates are obtained with SA-QLFS data. For column (3), groups of observations are thus excluded depending on the first digit of observations' enumeration area.



The estimate of column (2) shows somehow sensitive to the exclusion of some countries. For example, excluding Mozambique increases the estimated coefficient by about 40%. Excluding South Africa reduces the estimated coefficient by about 40% and strongly reduces its statistical significance (the p-value moves from 0.04 to 0.26).

In contrast, the estimate of column (3) appear less sensitive to the exclusion of some regions (changes in the coefficient remain in the [-20%; +20%] range).

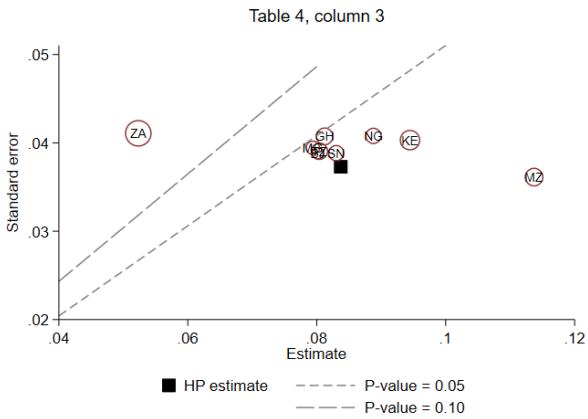
The figures below display estimated coefficients of interest from Table 3, panel B, columns (1), (3) and (4) when excluding regions one by one from the sample. Circles are proportional to share of each countries' observations in the original sample.



The estimate of column (1) proves very stable to the exclusion of some regions. In contrast, large changes in both the magnitude and the statistical significance of the estimated coefficients can be found for column (3) when excluding some regional groups of observations. Finally, while the estimated coefficient of column (4) seems to vary substantially when excluding some regions from the sample, it remains largely non-significant at conventional levels.

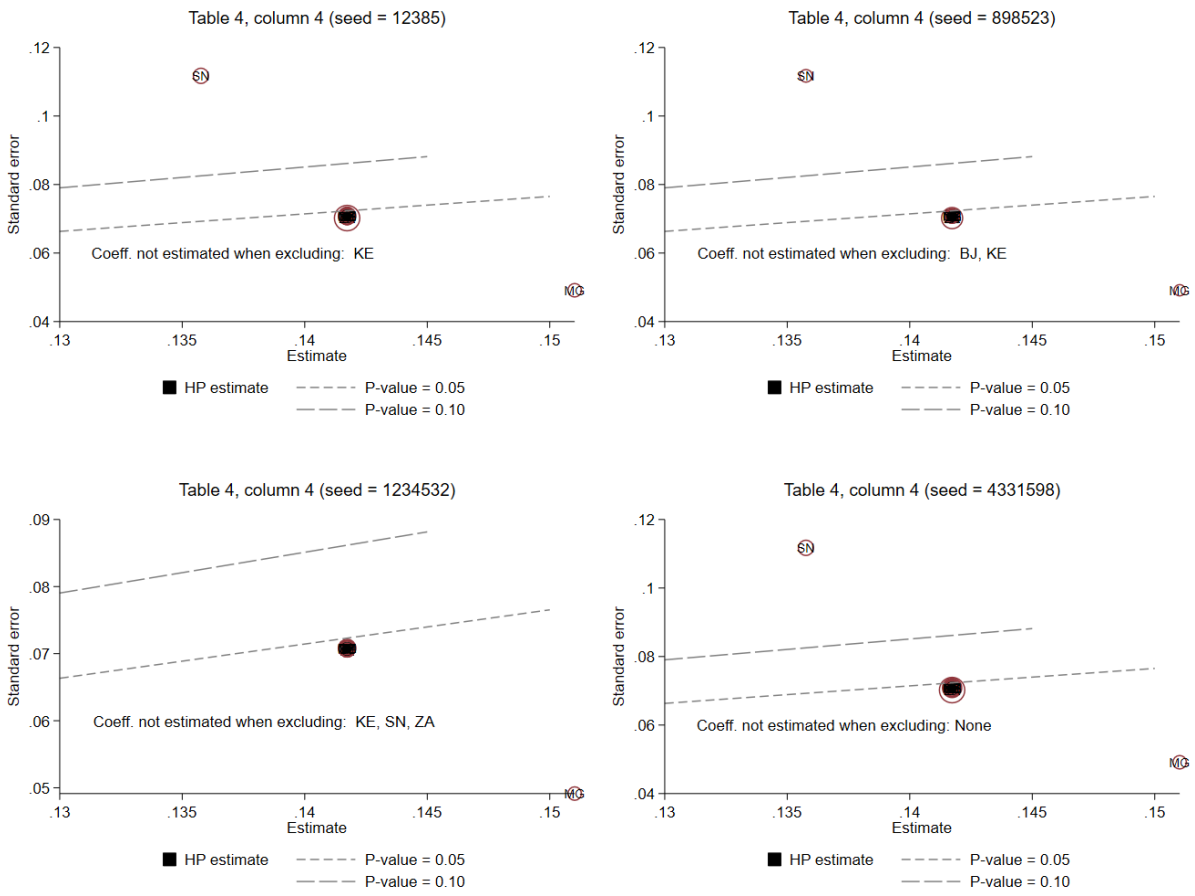
Table 4

The figure below shows how the estimate displayed in column (3) of Table 4 varies when excluding countries one by one. This estimate turns out to be somehow sensitive to the exclusion of some countries. For example, excluding Mozambique increases the estimated coefficient by about 35%. Excluding South Africa reduces the estimated coefficient by close to 40% and strongly reduces its statistical significance (the p-value moves from 0.02 to 0.20).



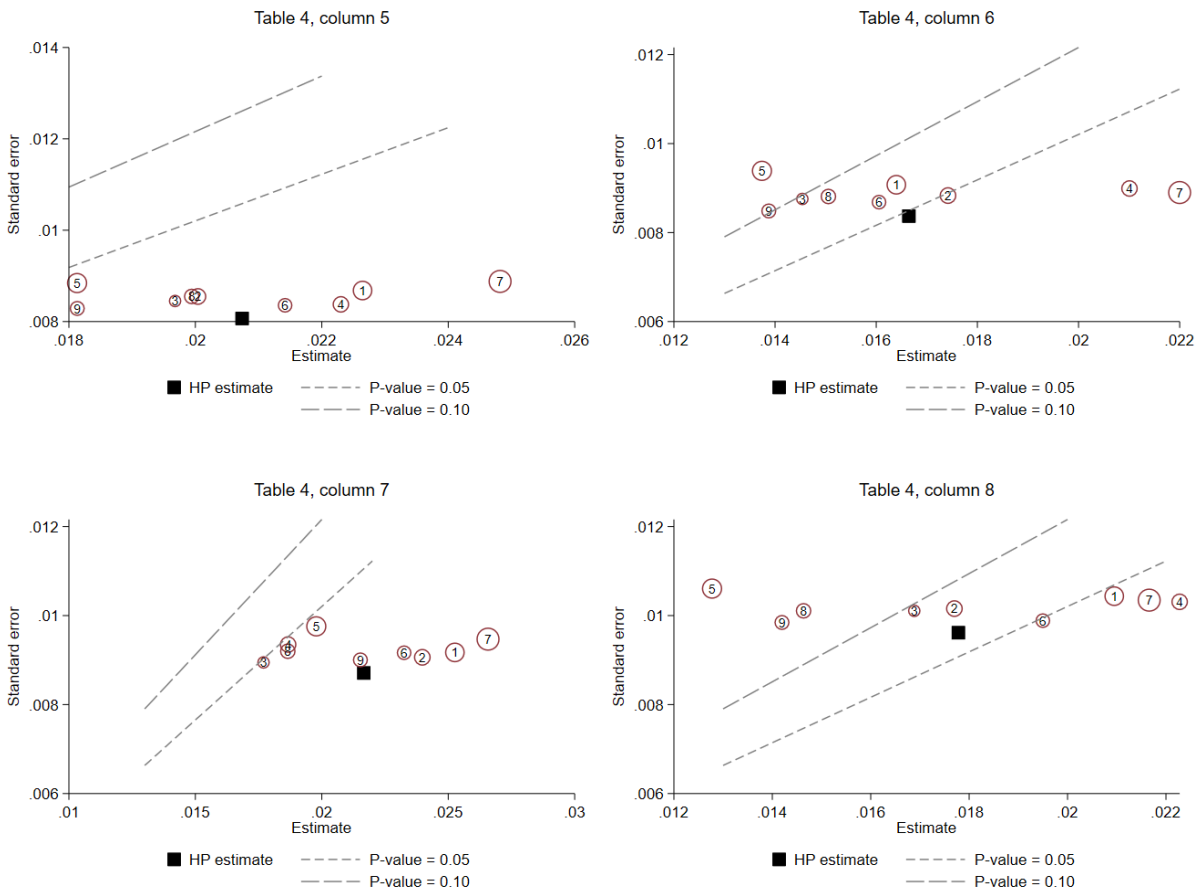
Running the original code used to compute the estimate of interest displayed in Table 4, column (4) revealed that the coefficient of interest cannot be estimated when excluding some countries. The specification used in column (4) actually supplement *grid-cell x connected* and *country x time* fixed effects with *connected x time fixed* effects. Such a large set of fixed effects is likely to neutralize information conveyed by many observations and to reduce a lot the quantity of information used for identification. It also likely creates collinearity issued that lead Stata to randomly drop some fixed effects.

Figures below show how the estimate of Table 4, column (4) vary when excluding countries one by one, depending on four possible seeds as STATA relies on randomness to decide which fixed effects to drop in case of collinearity.



These figures show that most countries do not contribute to the estimation of the coefficient of interest as point estimates are identical when excluding some countries) and that the model cannot be estimated when excluding some countries. These two observations suggest that the demanding sets of fixed effects used in this estimation lead to important collinearity issues that question what is actually captured by the coefficient reported in column (4) of Table 4 and how it should be interpreted.

Figures below shows how the estimates displayed in column (5)-(8) of Table 4 vary when excluding level-1 regions one by one.



Estimates of columns (5) and (7) appear weakly sensitive to the exclusion of some groups of observations. They remain stable in terms of magnitude and statistical significance.

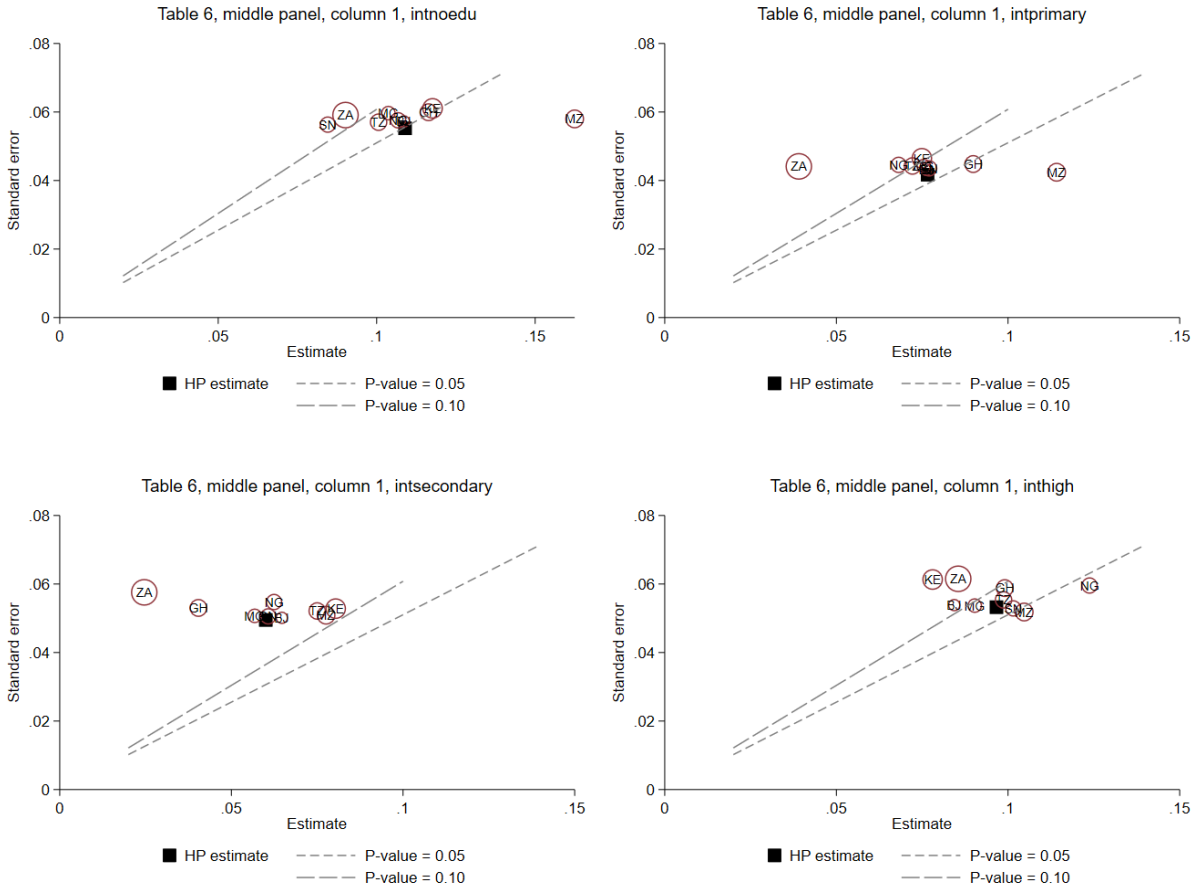
In contrast, estimates of columns (6) and (8) seem much more sensitive to the exclusion of some regional groups. The estimate of interest can be increased by about 25% when some regions are excluded. It can also be decreased by about 25% and turn not statistically significant at conventional levels when others regions are excluded from the sample. For example, the p-value of column (8) estimate moves from 0.06 to 0.15 or 0.23 when excluding any of the regions numbered with 5, 8 or 9.

Table 5

Sensitivity of Table 5, panel B was not investigated as results could not be reproduced (see comments about Table 5, panel B in the Reproducibility section).

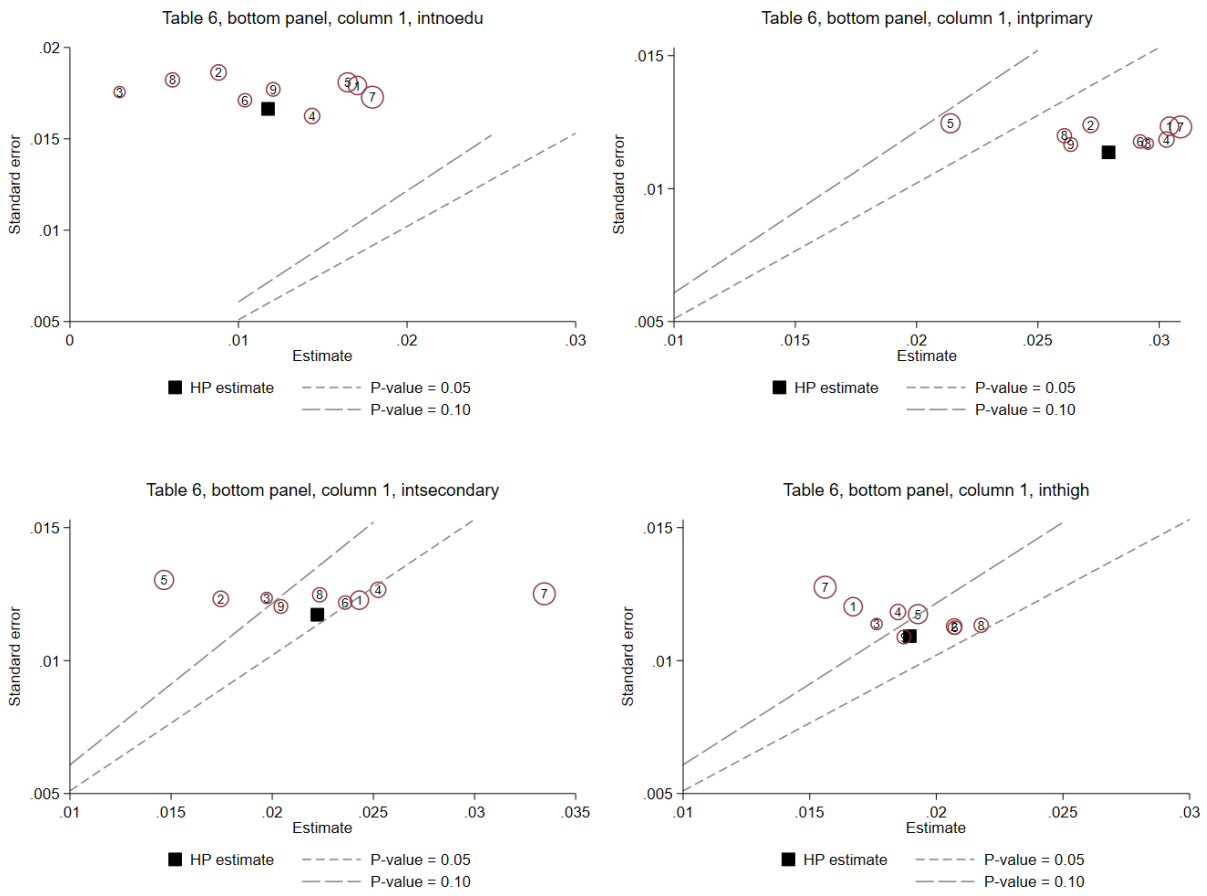
Table 6

Estimates reported in Table 6, middle panel are interaction terms between the treatment variable and respondents' level of education. Figures below shows how these estimates vary when excluding countries one by one.



The interaction terms with primary and secondary education appear sensitive in terms of magnitude and statistical significance to the exclusion of some countries such as South Africa, Ghana or Mozambique. In contrast, interaction terms with no primary or higher education prove only marginally sensitive to changes in the sample of countries.

Estimates reported in Table 6, bottom panel, column (1) are interaction terms between the treatment variable and respondents' level of education. Figures below shows how these estimates vary when excluding regions one by one.



Interaction terms with the highest levels of education appear sensitive to the exclusion of some groups of observations, mostly in terms of statistical significance. Note that all interactions terms are equally affected by the exclusion of respondents of the same regions. This observation might be related to the spatial sorting of jobs and social categories.

Figures below display how the results reported in columns (2) and (3) of Table 6, bottom panel vary when excluding regions one by one. While estimates of interaction terms show comparably sensitive as those of column (1) to the exclusion of regions, excluding specific regions does not substantially alter the comparison between results obtained for unskilled [column (2)] and skilled [column (3)] workers.

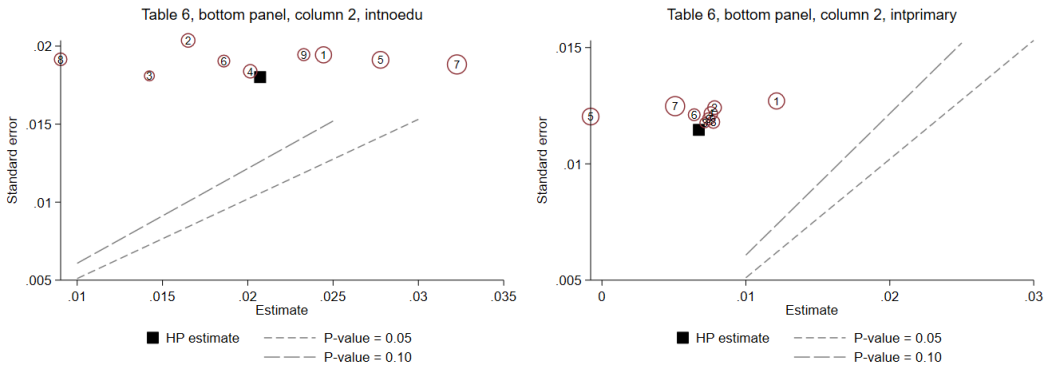


Table 6, bottom panel, column 2, intnoedu
Standard error
Estimate
HP estimate
P-value = 0.05
P-value = 0.10

Table 6, bottom panel, column 2, intprimary
Standard error
Estimate
HP estimate
P-value = 0.05
P-value = 0.10

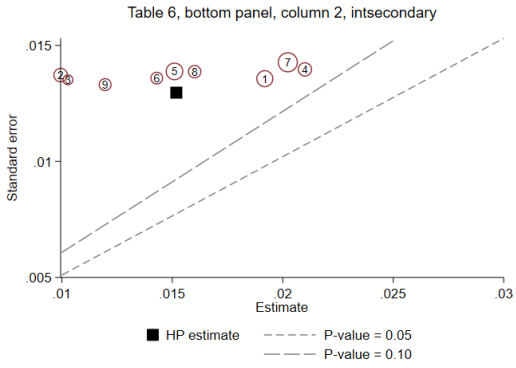


Table 6, bottom panel, column 2, intsecondary
Standard error
Estimate
HP estimate
P-value = 0.05
P-value = 0.10

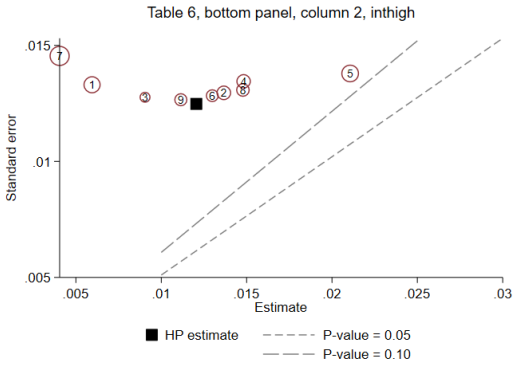


Table 6, bottom panel, column 2, inthigh
Standard error
Estimate
HP estimate
P-value = 0.05
P-value = 0.10

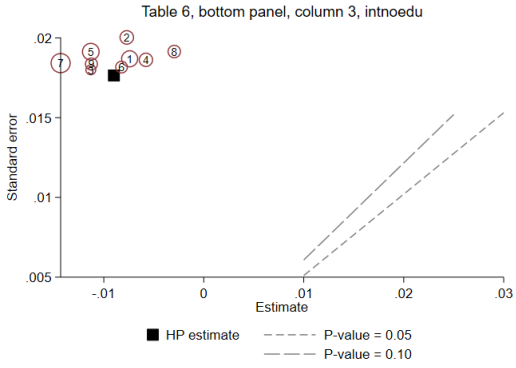


Table 6, bottom panel, column 3, intnoedu
Standard error
Estimate
HP estimate
P-value = 0.05
P-value = 0.10

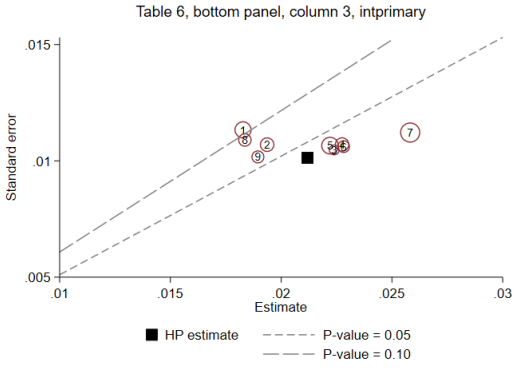


Table 6, bottom panel, column 3, intprimary
Standard error
Estimate
HP estimate
P-value = 0.05
P-value = 0.10

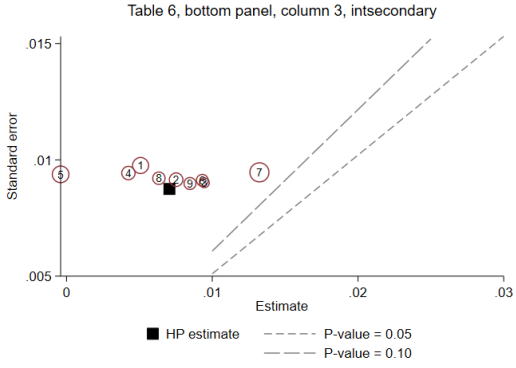


Table 6, bottom panel, column 3, intsecondary
Standard error
Estimate
HP estimate
P-value = 0.05
P-value = 0.10

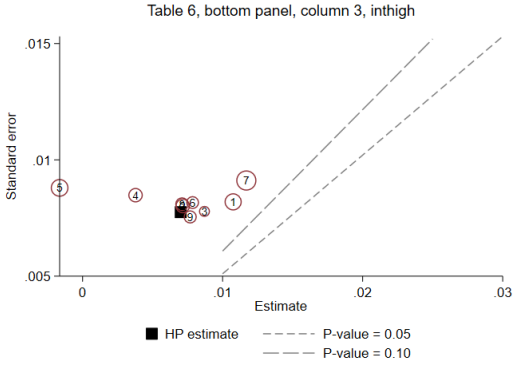
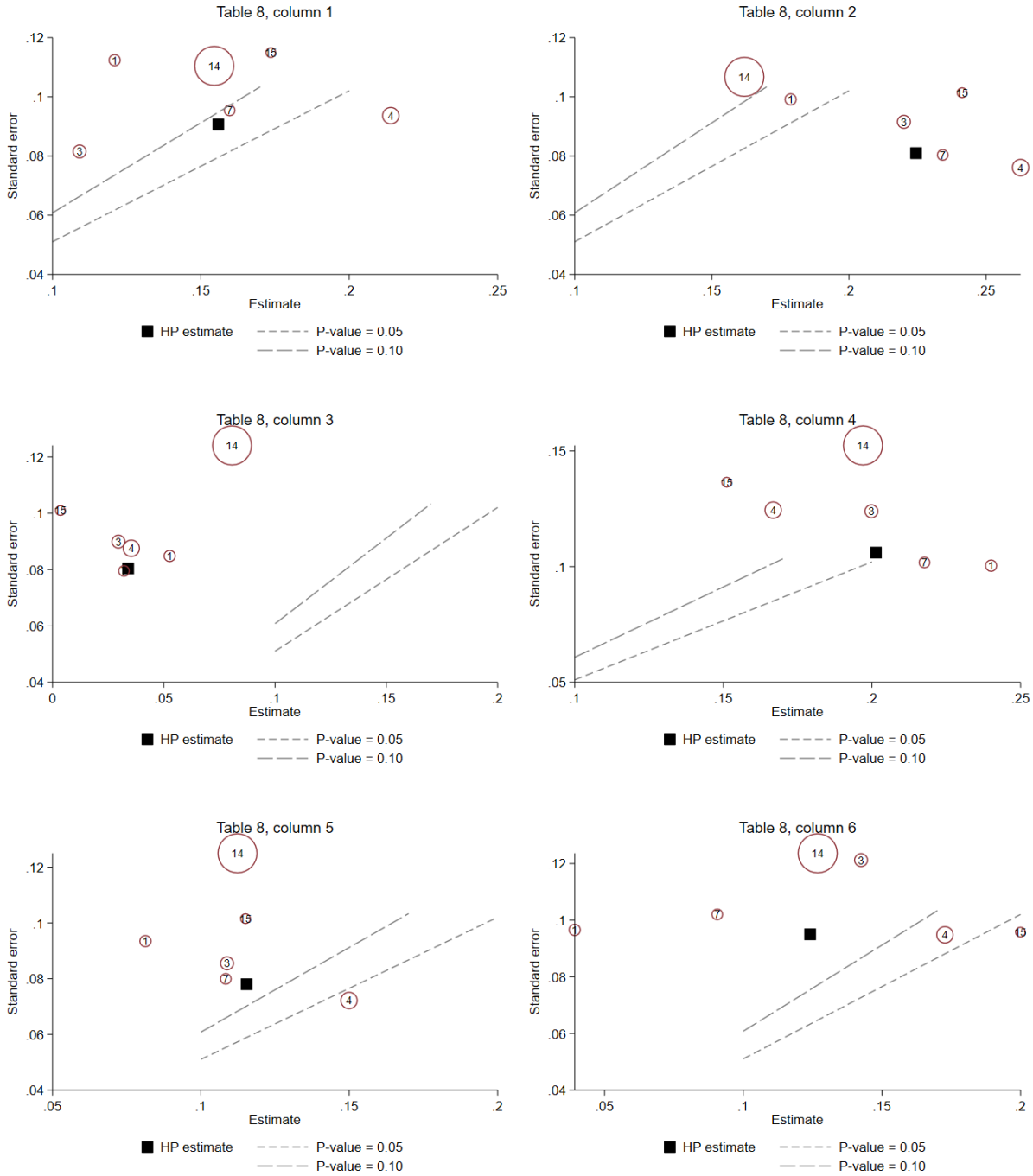


Table 6, bottom panel, column 3, inthigh
Standard error
Estimate
HP estimate
P-value = 0.05
P-value = 0.10

Table 8

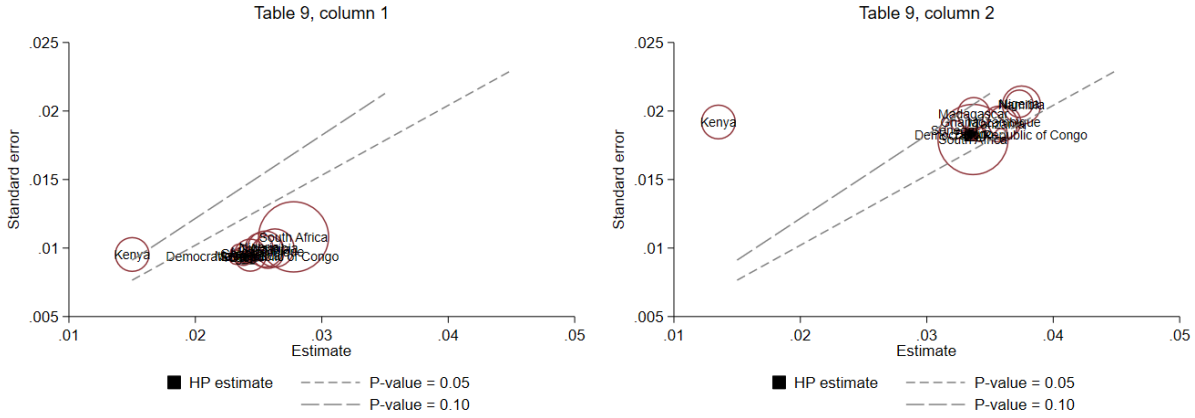
Table 8 uses data from the 2006-2013 census of Ethiopian firms. Sensitivity was tested to the successive exclusion of the 6 regions that are identified in the data. Figures below show how estimates reported in columns (1)-(6) vary when excluding regions one by one.



Sensitivity of results reported in columns (7)-(9) of Table 8 was not tested.

Table 9

Figures below display how estimates of Table 9 vary when excluding countries one by one.



Reported estimates appear to be highly sensitive to the exclusion of Kenya from the sample.

4 Conclusion

This (partial) replication study shows that most potentially reproducible results reported by HP can be reproduced from the paper's data and code. However, some could not be reproduced. In particular, results about the differential effect of fast Internet on employment depending on skills could not be reproduced from the available South Africa Quarterly Labor Force Survey data.

The sensitivity analysis shows that the paper's results are differently sensitive to the composition of the sample of observations. It also shows that some specifications might actually be too demanding for reasonable identification to be achieved from the data given the distribution of observations and treatment across time and space.

All in all, further empirical work would be needed to re-construct final data from potentially enlarged raw data so as to further explore the robustness of the paper's results.

References

Hjort, Jonas, and Jonas Poulsen. 2019. "The Arrival of Fast Internet and Employment in Africa." American Economic Review, 109 (3): 1032-79. DOI: 10.1257/aer.20161385

Hjort, Jonas, and Poulsen, Jonas. Replication data for: The Arrival of Fast Internet and Employment in Africa. Nashville, TN: American Economic Association [publisher], 2019. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2019-10-12. https://doi.org/10.3886/E113156V1.