Topics in econometrics
VT2022

Max Engström
Uppsala University

Replication of Olken (2007): ”Monitoring Corruption: Evidence from a Field Experiment in Indonesia."

1. Introduction

In Olken (2007), the author conducts a randomized field experiment where he measures the effect of monitoring on corruption in Indonesian villages. Data was collected on over 300 road projects from villages in central Jawa and east Jawa in Indonesia, where many of the villages, through randomization, were subject to different types of monitoring. More specifically, the three forms of monitoring included: (i) govermental audits of the projects, (ii) formal invitations to project meetings in the respective villages and (iii) formal invitations plus the opportunity to fill in an anonymous survey form ahead of the meetings.

Corruption was measured as the percent of missing expenditures between the official cost on the road project and the cost estimated from independent engineering teams. Findings indicated that on the one hand, increased audit monitoring, from the baseline 4 percent to 100 percent, had a negative impact on missing expenditures in road projects by an average of 8 percentage points. On the other hand, increased bottom-up monitoring in terms participation on meetings had a small and, in most cases, insignificant negative effect on missing expenditures. From the study it was concluded that top-down monitoring can be effective in reducing corruption, while bottom-up monitoring, however, appears to be limited in comparison.

For the remainder of this paper, what follows is a replication of the main results from Olken (2007). Apart from attempting to replicate parts of the main results from the study, the purpose of this paper is to also comment on some of the aspects in the analysis that were not addressed as well as provide extensions. Furthermore, the econometric specifications and main assumptions for the models are not explicitly addressed in this replication work. These are instead discussed in the original paper by Olken (2007) and so will not be repeated here.

What follows is the outline of this paper. Section 2 replicates the main tables of Olken (2007) and presents additional information to the analysis that was not included in some of the tables. Section 3 presents an extension that was never addressed in the study. Here I investigate the potential effect of nepotism through bottom-up monitoring. Section 4 presents the second extension where I address adjusted Roman-Wolf p-values for the multiple hypothesis in Olken (2007). Section 5 concludes this paper.

2. Replication analysis

Running the author's codes:

I was unable to display from the author's Stata code due to outdated coding related to the outreg command. Instead of replacing the original codes with outreg2 commands (or similar), I instead checked that the regression output from the Stata code matched with what was actually presented in the study itself,

The output that was not coded with the outreg command was instead coded with the outsheet command, and was therefore saved in .OUT format. The software program that was able to read these .OUT files did not give any clear indication what it was I was looking at. Instead of reading regression tables of the main results, I was instead looking at just pure numbers that did not tell anything about the results from the author's codes. In other words, I was unable to decipher some of the results from the author's codes and see whether it matched with the original study.

Replication of the paper:

Without success of running the do-file from the author's code to compare his results from Stata with what he had presented in the study, I proceeded by trying to replicate the main results from the study using my own code. This section thus reports the results from the replication of the tables from the study and further comments on them. Additionally, robustness checks that were not discussed in the paper are added for some of the tables. Before presenting the replication results, I provide a brief summary of the replication results. Overall the replication for all tables was successful, and most results came out very similarly to that of the original study. For many of the result, my coding resulted in standard errors that did not match entirely with the errors that the author reported.

However, despite these very small differences, in the majority of instances the difference in standard errors were not large enough as to change the results nor the general conclusions of the author. For the remaining part of this section I present the replication, table by table. I present the tables that I think were the most relevant to the main results of the study. All other tables are excluded from this paper.

Table 2

Table 2 in Olken (2007) reports on the relationships between the treatment mechanism and village characteristics. Column 1-3 are those that are included in the original paper, and they were successfully replicated in terms of coefficients and standard errors. The mostly insignificant estimates of the covariates on audit, invitations and comments (conditional on invitations) indicate that there is no selection bias between the groups. The randomization into treatment and control group thus seem to have worked, which is further confirmed by the large p-values for the joint test on all covariates.

Interesting to note is that in the datafile the author used the variable ”undfpm” as the dependent variable for invitations in column 2. However, ”undfpm” explains *invitations or comments*, and not invitations. Therefore, in table 2 replicated here I include column 4 which reports on the actual invitations variable, ”und”. The estimates from column 4 gives the same conclusion regarding randomization, which leads us to believe that using the ”undfpm” variable instead of ”und” probably was a simple coding mistake from the author rather than an intentional decision.

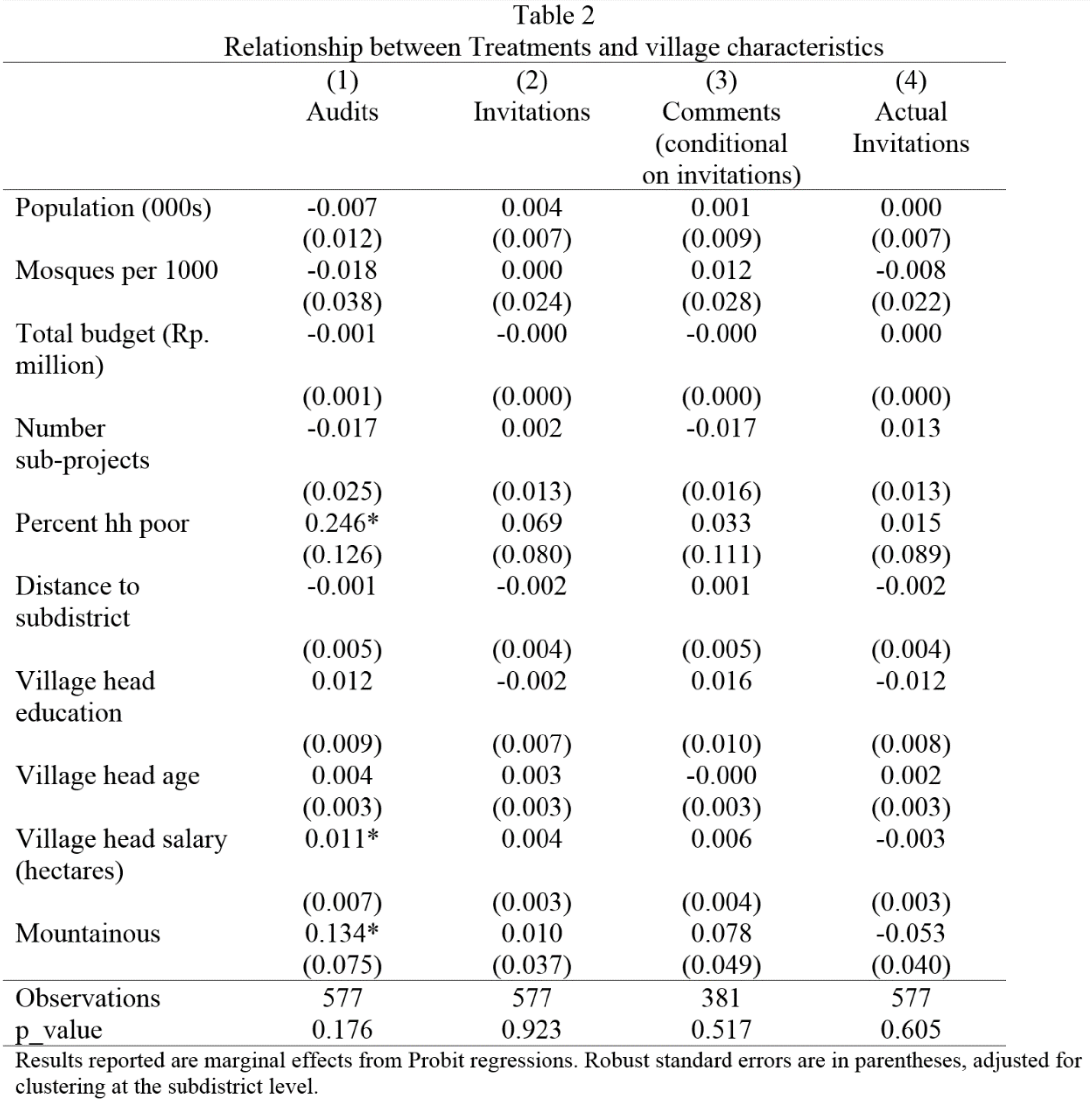


Table 2
Relationship between Treatments and village characteristics

	(1) Audits	(2) Invitations	(3) Comments (conditional on invitations)	(4) Actual Invitations
Population (000s)	-0.007	0.004	0.001	0.000
	(0.012)	(0.007)	(0.009)	(0.007)
Mosques per 1000	-0.018	0.000	0.012	-0.008
	(0.038)	(0.024)	(0.028)	(0.022)
Total budget (Rp. million)	-0.001	-0.000	-0.000	0.000
	(0.001)	(0.000)	(0.000)	(0.000)
Number sub-projects	-0.017	0.002	-0.017	0.013
	(0.025)	(0.013)	(0.016)	(0.013)
Percent hh poor	0.246*	0.069	0.033	0.015
	(0.126)	(0.080)	(0.111)	(0.089)
Distance to subdistrict	-0.001	-0.002	0.001	-0.002
	(0.005)	(0.004)	(0.005)	(0.004)
Village head education	0.012	-0.002	0.016	-0.012
	(0.009)	(0.007)	(0.010)	(0.008)
Village head age	0.004	0.003	-0.000	0.002
	(0.003)	(0.003)	(0.003)	(0.003)
Village head salary (hectares)	0.011*	0.004	0.006	-0.003
	(0.007)	(0.003)	(0.004)	(0.003)
Mountainous	0.134*	0.010	0.078	-0.053
	(0.075)	(0.037)	(0.049)	(0.040)
Observations	577	577	381	577
p_value	0.176	0.923	0.517	0.605

Results reported are marginal effects from Probit regressions. Robust standard errors are in parentheses, adjusted for clustering at the subdistrict level.

Table 3

Table 3 reports the summary statistics from the paper. This table was successfully replicated, with exact means and standard deviations as in the study. This table illustrates the average project size per village in expenditures, shares of the expenditures into different projects, shares of reported expenses , and missing expenditures which are the dependent variables for most of the study.

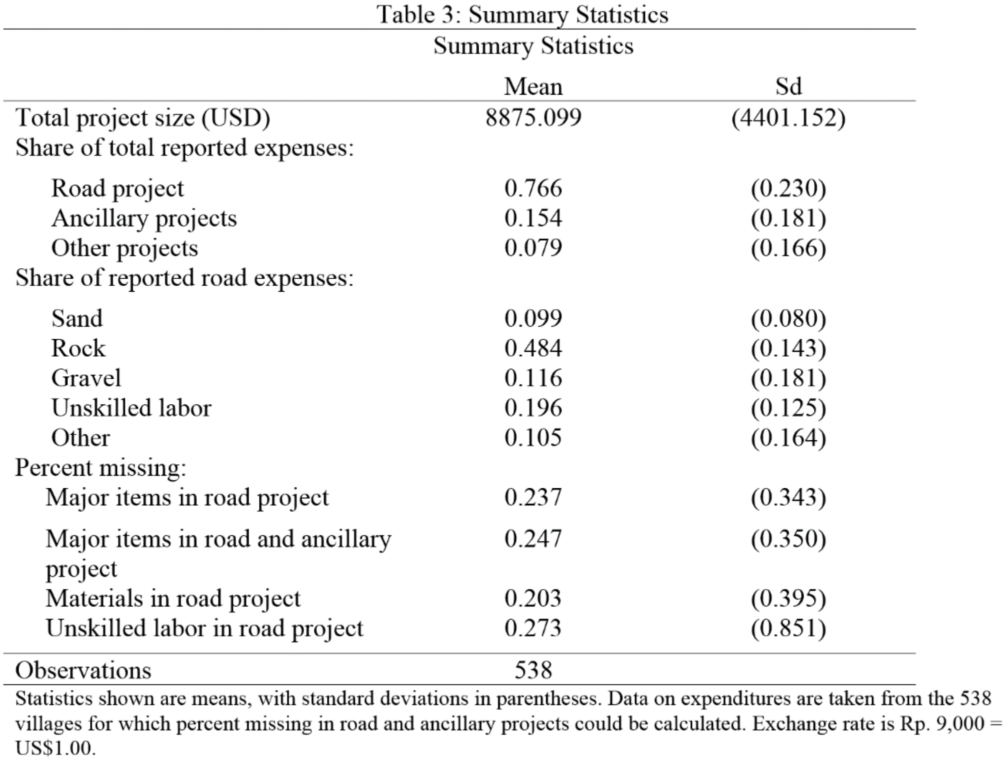


Table 3: Summary Statistics

	Summary Statistics	
	Mean	Sd
Total project size (USD)	8875.099	(4401.152)
Share of total reported expenses:		
Road project	0.766	(0.230)
Ancillary projects	0.154	(0.181)
Other projects	0.079	(0.166)
Share of reported road expenses:		
Sand	0.099	(0.080)
Rock	0.484	(0.143)
Gravel	0.116	(0.181)
Unskilled labor	0.196	(0.125)
Other	0.105	(0.164)
Percent missing:		
Major items in road project	0.237	(0.343)
Major items in road and ancillary project	0.247	(0.350)
Materials in road project	0.203	(0.395)
Unskilled labor in road project	0.273	(0.851)
Observations	538	

Statistics shown are means, with standard deviations in parentheses. Data on expenditures are taken from the 538 villages for which percent missing in road and ancillary projects could be calculated. Exchange rate is Rp. 9,000 = US$1.00.

Table 4

The main results from audit monitoring on missing expenditures, under three separate fixed effects specifications, are presented in table 4. I managed to replicate all coefficients and most of the standard errors from the table. The errors that do differ from those in the study do so only to a very small extent, and not enough to actually change the conclusion. From the replication I thus find a similar conclusion as the author: audit monitoring is related to a decrease in missing expenditures, by 7.6 to 8.5 percentage points in major items in road, and around 8.6 to 9 percentage points when including missing expenditures in ancillery projects. The last two dependent variables, for missing expenditures in road materials and in unskilled labor, show insignificant coefficients for all specifications, similar to in the study.

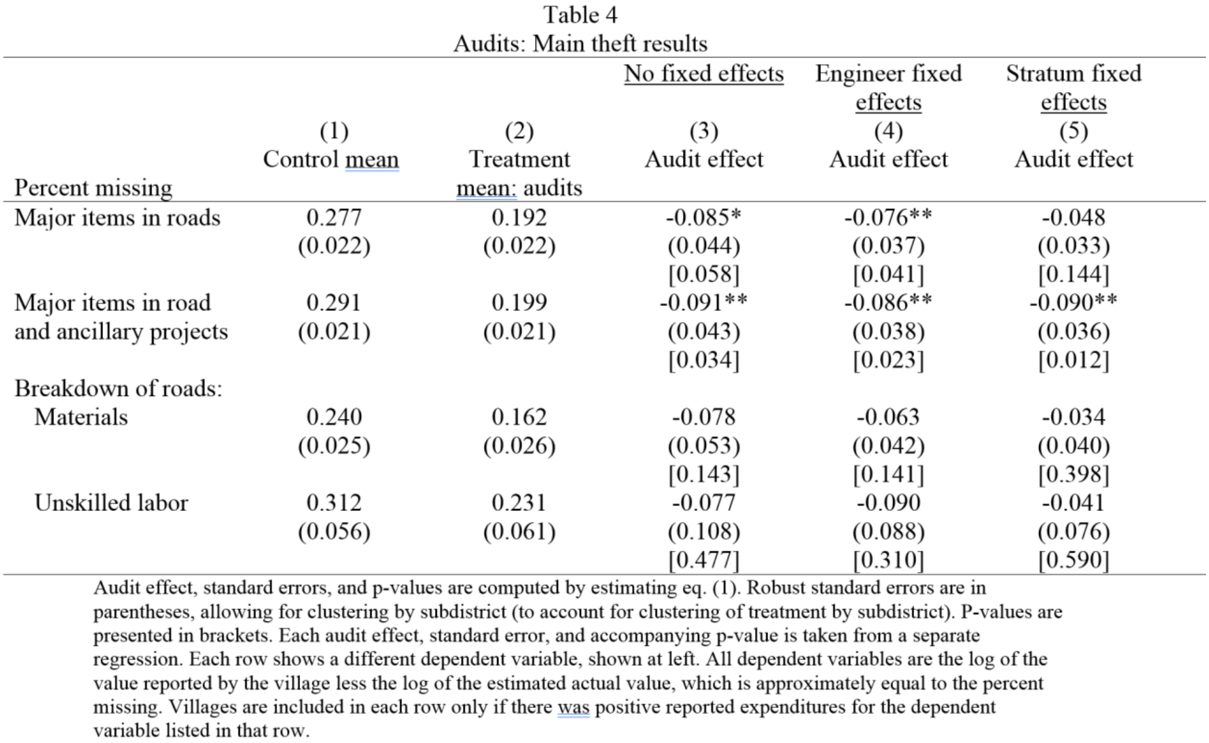


Table 4
Audits: Main theft results

			No fixed effects	Engineer fixed effects	Stratum fixed effects
	(1)	(2)	(3)	(4)	(5)
Percent missing	Control mean	Treatment mean: audits	Audit effect	Audit effect	Audit effect
Major items in roads	0.277	0.192	-0.085*	-0.076**	-0.048
	(0.022)	(0.022)	(0.044)	(0.037)	(0.033)
			[0.058]	[0.041]	[0.144]
Major items in road and ancillary projects	0.291	0.199	-0.091**	-0.086**	-0.090**
	(0.021)	(0.021)	(0.043)	(0.038)	(0.036)
			[0.034]	[0.023]	[0.012]
Breakdown of roads:					
Materials	0.240	0.162	-0.078	-0.063	-0.034
	(0.025)	(0.026)	(0.053)	(0.042)	(0.040)
			[0.143]	[0.141]	[0.398]
Unskilled labor	0.312	0.231	-0.077	-0.090	-0.041
	(0.056)	(0.061)	(0.108)	(0.088)	(0.076)
			[0.477]	[0.310]	[0.590]

Audit effect, standard errors, and p-values are computed by estimating eq. (1). Robust standard errors are in parentheses, allowing for clustering by subdistrict (to account for clustering of treatment by subdistrict). P-values are presented in brackets. Each audit effect, standard error, and accompanying p-value is taken from a separate regression. Each row shows a different dependent variable, shown at left. All dependent variables are the log of the value reported by the village less the log of the estimated actual value, which is approximately equal to the percent missing. Villages are included in each row only if there was positive reported expenditures for the dependent variable listed in that row.

Table 6

The replication for table 6 was fairly successful, however the standard errors in the replication was higher than in Olken (2007). This caused the estimates for *auditor admin score* on *engineering team admin score* (column 2) and *percent missing in road* (column 3) to decrease from a 95 percent significance level to a 90 percent level. Otherwise the results look similar. Specifically, we find a positive correlation between the auditor team physical score and the engineering team physical score (column 1), as well as a positive correlation between auditor administrative score and engineering administrative score (column 2). The positive correlation implies that both the auditor team and the independent engineering teams graded the project, both in terms of infrastructure (physical score) and administrative score, fairly similar. This finding could in itself indicate that the auditors might not have been entirely corrupt.

What is noteworthy from this table though, is that the author only included column 1-3 in the paper. While column 1 and 2 reports on engineering team scores as the dependent variables, column 3 instead reports one of the missing expenditures variables. Therefore, in this table the author only presented results for *one* of the four dependent variables of corruption. On the coefficient for this dependent variable, *percent missing in road*, he reported a negative and significant estimate from auditor admin score (column 3). He further uses this significant estimate to validate the underlying conclusion of the table, that: ”*the auditors were not completely corrupt […] and that the administrative aspects investigated by the auditors were in fact correlated with missing expenditures*”.

In the replication we extended table 6 to also include the regressions for the other three dependent variables: *percent missing in road and ancillary projects*, *percent missing in road materials* and *percent missing in unskilled labor*, each displayed in column 4, 5 and 6. The coefficients show insignificant estimates of auditor admin score for each one of them, compared to the significant estimate for the one that the author had chosen to include in the study.

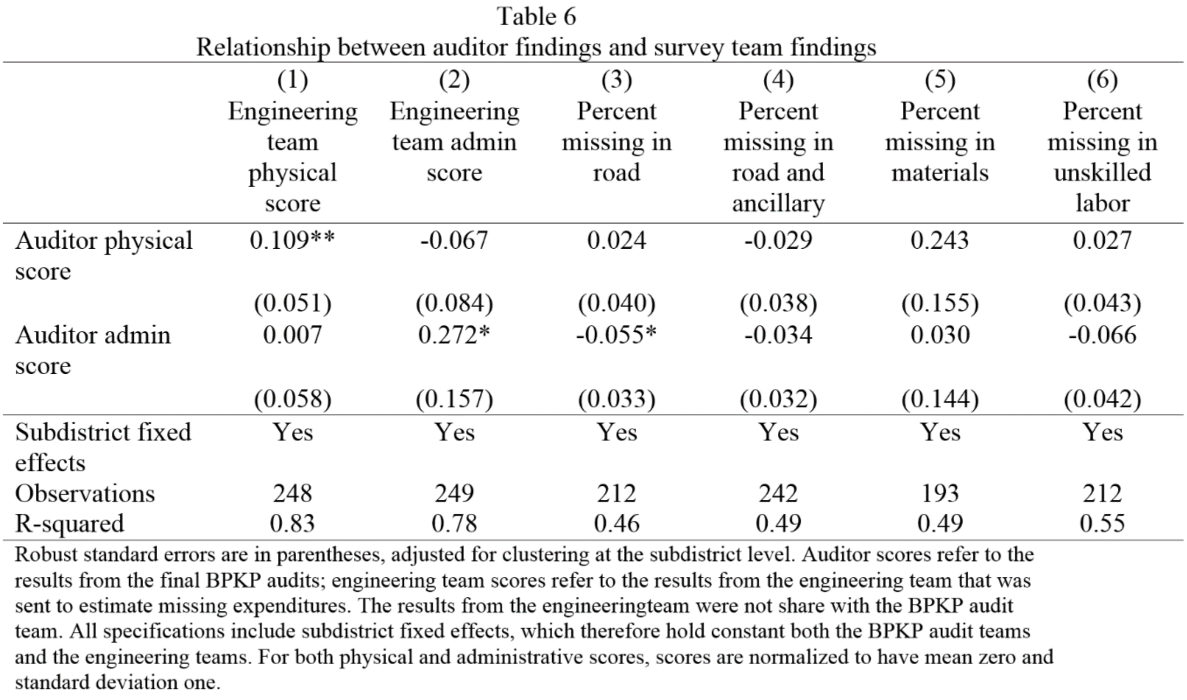


Table 6
Relationship between auditor findings and survey team findings

	(1) Engineering team physical score	(2) Engineering team admin score	(3) Percent missing in road	(4) Percent missing in road and ancillary	(5) Percent missing in materials	(6) Percent missing in unskilled labor
Auditor physical score	0.109**	-0.067	0.024	-0.029	0.243	0.027
	(0.051)	(0.084)	(0.040)	(0.038)	(0.155)	(0.043)
Auditor admin score	0.007	0.272*	-0.055*	-0.034	0.030	-0.066
	(0.058)	(0.157)	(0.033)	(0.032)	(0.144)	(0.042)
Subdistrict fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Observations	248	249	212	242	193	212
R-squared	0.83	0.78	0.46	0.49	0.49	0.55

Robust standard errors are in parentheses, adjusted for clustering at the subdistrict level. Auditor scores refer to the results from the final BPKP audits; engineering team scores refer to the results from the engineering team that was sent to estimate missing expenditures. The results from the engineeringteam were not share with the BPKP audit team. All specifications include subdistrict fixed effects, which therefore hold constant both the BPKP audit teams and the engineering teams. For both physical and administrative scores, scores are normalized to have mean zero and standard deviation one.

Alone, our estimate in column 3 would lead us to draw the same conclusion as Olken (2007) did. However, including any other of the dependent variables would instead cast doubt whether we can actually draw the same conclusion as the author did (due to not being able to reject the null hypothesis for the other ones). Where does this finding lead us? One would think that there is a reason for why the missing expenditures variable with significant result is the only one presented in table 6 in the study. It could of course just be a random occurance that the author forgot to include any of the other variables in the table.

It could also have been an intentional decision to only include *missing expenditures in road projects* since road projects make up the largest portion of the share of total reported expenditures (see table 2). However, as reported in table 4, there would be little reason not to also include *missing expenditures in road and ancillary projects*, due to its heavily significant negative effect from audits (see table 4; row 2; columns 3–5). The fact that he did not do this however, would further suggest that the author has cherrypicked the estimate that yields a significant value (and which validates his results) while excluding any others from the table.

While this general finding does not change the overall conclusion of the article, it does however contest the conclusion that the author does from this table. Thus, if one includes the other missing expenditure variables, one can not easily draw the same conclusion anymore that higher administrative score from the auditor is in fact correlated with less missing expenditures.

Table 8

The author investigated if there existed alternative forms of corruption for audited villages relative control villages. This is reported in table 8, which reports if village heads and project heads employed family members (nepotism) to the projects due to the increased audit monitoring. The replication of this table was successful, but with small changes for some of the coefficients. In the study, the variable for project head family member was insignificant, however in the replication this variable is significant on a 10 percent level in column 1 and 3. The reason for this is unclear however. Apart from this variable, all other estimates point towards the same conclusion as in the study. Audited villages are correlated with nepotism: a higher share of project workers in audited villages, compared to control villages, are related to the village or project head.

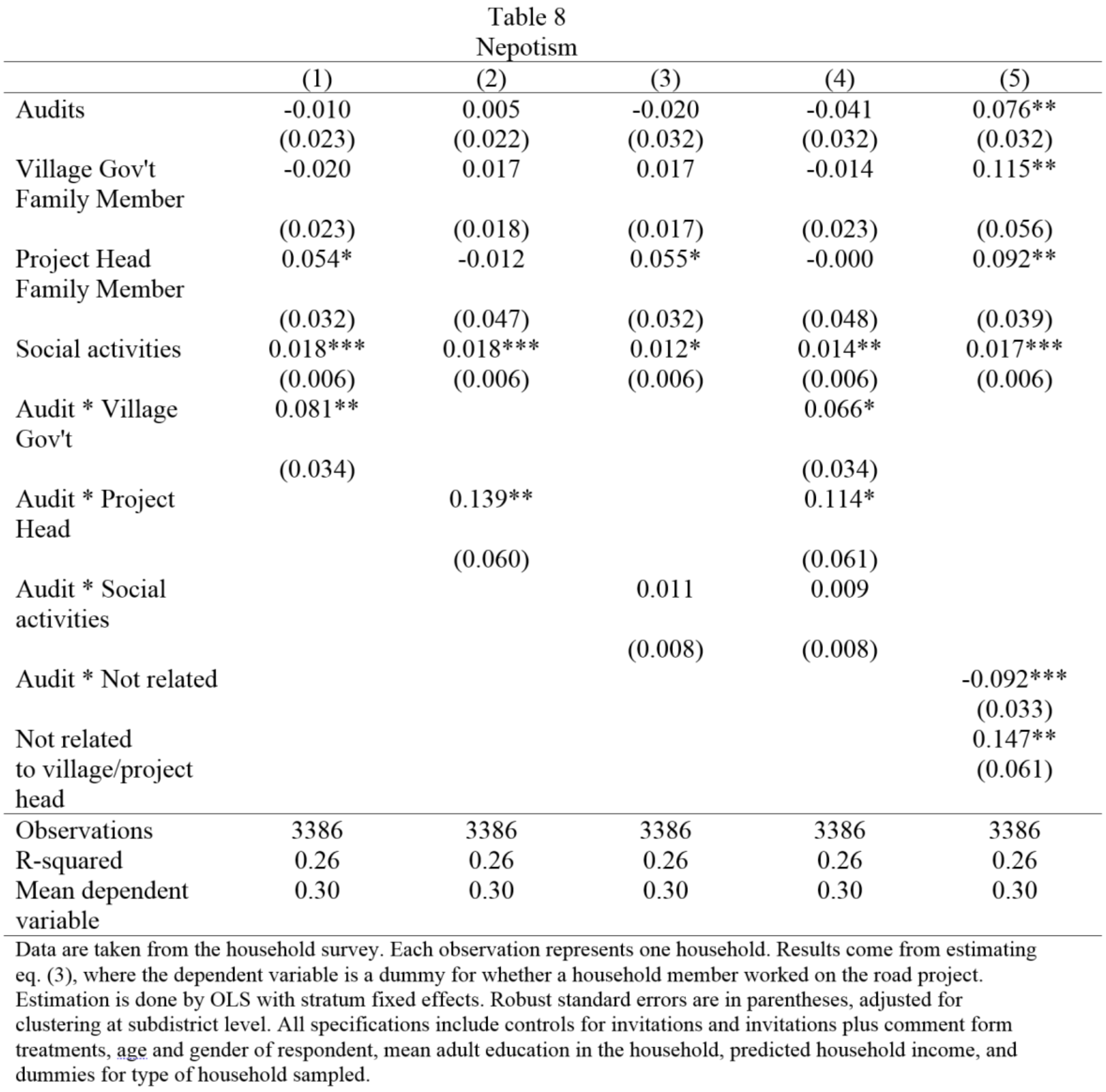


Table 8
Nepotism

	(1)	(2)	(3)	(4)	(5)
Audits	-0.010	0.005	-0.020	-0.041	0.076**
	(0.023)	(0.022)	(0.032)	(0.032)	(0.032)
Village Gov't Family Member	-0.020	0.017	0.017	-0.014	0.115**
	(0.023)	(0.018)	(0.017)	(0.023)	(0.056)
Project Head Family Member	0.054*	-0.012	0.055*	-0.000	0.092**
	(0.032)	(0.047)	(0.032)	(0.048)	(0.039)
Social activities	0.018***	0.018***	0.012*	0.014**	0.017***
	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)
Audit * Village Gov't	0.081**			0.066*	
	(0.034)			(0.034)	
Audit * Project Head		0.139**		0.114*	
		(0.060)		(0.061)	
Audit * Social activities			0.011	0.009	
			(0.008)	(0.008)	
Audit * Not related					-0.092***
					(0.033)
Not related to village/project head					0.147**
					(0.061)
Observations	3386	3386	3386	3386	3386
R-squared	0.26	0.26	0.26	0.26	0.26
Mean dependent variable	0.30	0.30	0.30	0.30	0.30

Data are taken from the household survey. Each observation represents one household. Results come from estimating eq. (3), where the dependent variable is a dummy for whether a household member worked on the road project. Estimation is done by OLS with stratum fixed effects. Robust standard errors are in parentheses, adjusted for clustering at subdistrict level. All specifications include controls for invitations and invitations plus comment form treatments, age and gender of respondent, mean adult education in the household, predicted household income, and dummies for type of household sampled.

As a robustness check, I ran a regression to see if the probability of *not being related* to the village or project head, on the contrary, was lower in audited villages compared to control villages. I would claim that if workers that are related to village heads have a higher probability of working in audited villages, then the opposite would also be true for the same villages if one is instead not related to the village head. The results from this robustness check is presented in column 5, where an added variable for *not being related to the village or project head* is included, plus an interaction between this variable and audits. The control variable for *not related* shows a positive and significant estimate. But more importantly, the negative coefficient of the interaction (that is statistically significant on a 99 percent level) confirms my previous statement in this section.

Furthermore, this finding strengthens the already established hypothesis of the author regarding nepotism in audited villages.[1]

One last thing to mention is that table 8 reports estimates from a linear probability model (LPM). In the code I ran probit regressions for the same specifications to see if the LPM coefficients were consistent. The probit estimates for table 8 are presented below in table 8b. Column 5 from table 8 is also included. Overall, the estimates from the probit show a slightly larger positive effect on the probability of working, with the most remarkable effect being observed is in column 2 for the interaction coefficient of interest. In the LPM, this coefficient showed a positive effect of 0.139 on a 5 percent level, while in the probit it is 0.166 and at a 1 percent level. This could imply that the estimates in the LPM are slightly underestimated. Otherwise, the probit results are reasonably similar to the ones of the LPM, and the overall conclusion thus stays the same.

[1] I also ran a regression that, similar to column 4, included all interaction terms (plus the new interaction of not related). This regression (not included in the paper) turned all interaction coefficients insignificant. This is most likely due to the high correlation between the norelated interaction and village head interaction on the one hand, and the norelated interaction and project head interaction on the other.

Table 8b
Nepotism: Probit

	(1)	(2)	(3)	(4)	(5)
Audits	-0.011	0.006	-0.024	-0.049	0.086**
	(0.028)	(0.025)	(0.036)	(0.037)	(0.036)
Village Gov't Family Member	-0.019	0.023	0.022	-0.011	0.107*
	(0.028)	(0.020)	(0.020)	(0.028)	(0.060)
Project Head Family Member	0.056*	-0.026	0.056*	-0.014	0.087**
	(0.034)	(0.052)	(0.034)	(0.052)	(0.039)
Social activities	0.019***	0.019***	0.013*	0.014**	0.019***
	(0.006)	(0.006)	(0.007)	(0.007)	(0.006)
Audit * Village Gov't Family Member	0.089**			0.070*	
	(0.039)			(0.039)	
Audit * Project Head Family Member		0.166***		0.140**	
		(0.064)		(0.065)	
Audit * Social activities			0.013	0.011	
			(0.009)	(0.009)	
Audit * Not related to village/project head					-0.104***
					(0.038)
Not related to village/project head					0.140**
					(0.066)
Observations	3365	3365	3365	3365	3365

Check note for table 8. Probit estimates of table 8 that checked for nepotism. Robust standard errors are in parentheses, adjusted for clustering at subdistrict level. R-squared was not successfully included in this table unfortunately. Otherwise, all controls and interactions remain the same as in table 8.

Table 9 and 10

Table 9 and 10 are presented here very briefly. The replication of both of them was successful. Table 9 shows that both participation mechanisms (invitations and invitations plus comments) had a positive and significant effect on overall attendance at accountability meetings, as well as on the number of peolpe that talked. Table 10 indicates that the participation mechanisms had a positive effect on the number of corruption-related problems being discussed. Increased participation did however not have any clear effect on the number of problems being discussed.

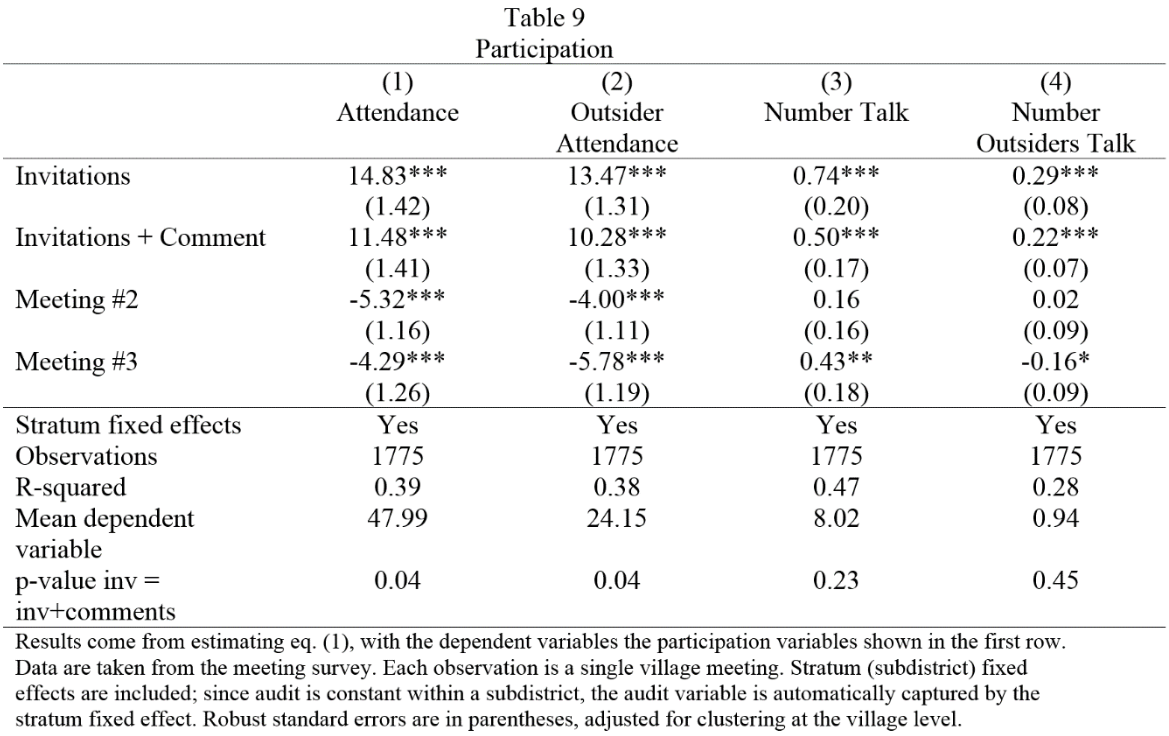


Table 9
Participation

	(1) Attendance	(2) Outsider Attendance	(3) Number Talk	(4) Number Outsiders Talk
Invitations	14.83***	13.47***	0.74***	0.29***
	(1.42)	(1.31)	(0.20)	(0.08)
Invitations + Comment	11.48***	10.28***	0.50***	0.22***
	(1.41)	(1.33)	(0.17)	(0.07)
Meeting #2	-5.32***	-4.00***	0.16	0.02
	(1.16)	(1.11)	(0.16)	(0.09)
Meeting #3	-4.29***	-5.78***	0.43**	-0.16*
	(1.26)	(1.19)	(0.18)	(0.09)
Stratum fixed effects	Yes	Yes	Yes	Yes
Observations	1775	1775	1775	1775
R-squared	0.39	0.38	0.47	0.28
Mean dependent variable	47.99	24.15	8.02	0.94
p-value inv = inv+comments	0.04	0.04	0.23	0.45

Results come from estimating eq. (1), with the dependent variables the participation variables shown in the first row. Data are taken from the meeting survey. Each observation is a single village meeting. Stratum (subdistrict) fixed effects are included; since audit is constant within a subdistrict, the audit variable is automatically captured by the stratum fixed effect. Robust standard errors are in parentheses, adjusted for clustering at the village level.

Table 10
Participation: Impact on meetings

	(1) Number problems	(2) Corruption-related problem	(3) Serious response
Invitations	0.072	0.027**	-0.003
	(0.066)	(0.013)	(0.008)
Invitations + Comment	0.104	0.026**	0.015*
	(0.067)	(0.013)	(0.008)
Meeting #2	-0.187***	0.002	-0.020**
	(0.069)	(0.013)	(0.010)
Meeting #3	-0.428***	-0.036***	-0.029***
	(0.077)	(0.012)	(0.010)
Stratum fixed effects	Yes	Yes	Yes
Observations	1783	1783	1783
R-squared	0.50	0.31	0.22
Mean dependent variable	1.18	0.07	0.03
p-value inv = inv+comments	0.62	0.96	0.02

Results come from estimating eq. (1), with the dependent variables the participation variables shown in the first row. Data are taken from the meeting survey. Each observation is a single village meeting. Serious response is defined as agreeing to replace a supplier or village office, agreeing that money should be returned, agreeing to an internal village investigation, asking for help from district project officials, or requesting an external audit. Estimation is by OLS. Robust standard errors are in parentheses, adjusted for clustering by village.

Table 11

Table 11 highlights the main results from the participation monitoring on missing expenditures. The replication points towards the same conclusion as in the paper: looking at the difference between column (1) and (2), grassroot monitoring is related to a small decrease in missing expenditures between 1.5 to 3 percentage points for most dependent variables. The coefficients for all specifications are however insignificant except for missing expenditures in unskilled labor due to invitations, which shows a decrease up to 21 percentage points in column (4) with engineer fixed effects.

Table 11
Participation: main theft results

			No fixed effects	Engineer fixed effects	Stratum fixed effects
	(1)	(2)	(3)	(4)	(5)
	Control mean	Treatment mean	Treatment effect	Treatment effect	Treatment effect
			A. invitations		
Major items in roads	0.252	0.230	-0.021	-0.030	-0.026
	(0.026)	(0.030)	(0.035)	(0.035)	(0.034)
			[0.5569]	[0.391]	[0.448]
Major items in roads and ancillary projects	0.268	0.236	-0.030	-0.032	-0.029
	(0.025)	(0.028)	(0.032)	(0.032)	(0.032)
			[0.360]	[0.324]	[0.356]
Breakdown of roads:					
Materials	0.209	0.221	0.014	0.008	0.005
	(0.028)	(0.036)	(0.038)	(0.038)	(0.037)
			[0.725]	[0.841]	[0.882]
Unskilled labor	0.369	0.180	-0.187*	-0.215**	-0.143*
	(0.065)	(0.074)	(0.098)	(0.095)	(0.086)
			[0.058]	[0.026]	[0.098]
			B. invitations plus comments		
Major items in roads	0.252	0.228	-0.022	-0.024	-0.015
	(0.026)	(0.025)	(0.030)	(0.029)	(0.030)
			[0.455]	[0.416]	[0.601]
Major items in roads and ancillary projects	0.268	0.238	-0.026	-0.025	-0.027
	(0.025)	(0.025)	(0.032)	(0.031)	(0.031)
			[0.409]	[0.411]	[0.385]
Breakdown of roads:					
Materials	0.209	0.180	-0.028	-0.022	-0.010
	(0.028)	(0.030)	(0.034)	(0.033)	(0.033)
			[0.414]	[0.501]	[0.754]
Unskilled labor	0.369	0.267	-0.099	-0.132	-0.090
	(0.065)	(0.075)	(0.087)	(0.088)	(0.091)
			[0.255]	[0.136]	[0.323]

See the note to table 4. Results come from estimating eq. (1), a regression of the dependent variable on a dummy for audit treatment, invitations treatment, and invitations plus comment forms treatments. Each Invitations effect and invitations plus comments effect comes from a separate regression, with the dependent variable listedin the row and the fixed effects specification listed in the column heading. Robust standard errors are in parentheses. Regressions without stratum fixed effects include a variable for audits and allow for clustering of standard errors by subdistrict.

Table 12

Table 12 discusses the same effects as in table 11, but it separates the effects of participation by distribution via neighborhood heads and distribution via schools. Estimates from this table shows similar effects as in table 10: mostly small effects on missing expenditures between the treatment and control villages from increased participatory monitoring. In all cases we observe insignificant estimates from invitations no matter the method of distribution, For invitations plus comments however, we observe a small yet significant decrease on missing expenditures while distributed via schools.

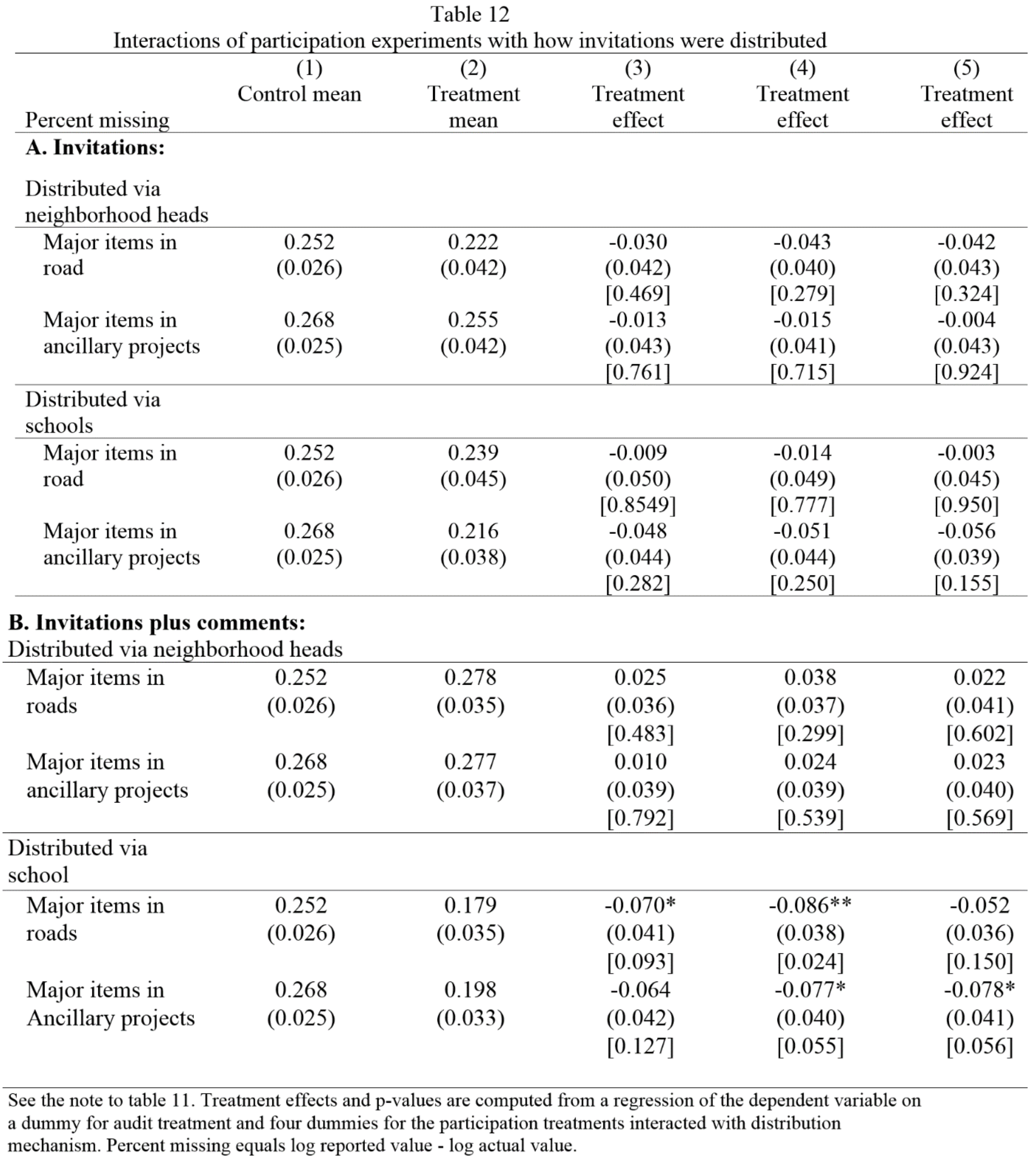


Table 12
Interactions of participation experiments with how invitations were distributed

	(1) Control mean	(2) Treatment mean	(3) Treatment effect	(4) Treatment effect	(5) Treatment effect
Percent missing					
A. Invitations:					
Distributed via neighborhood heads					
Major items in road	0.252 (0.026)	0.222 (0.042)	-0.030 (0.042) [0.469]	-0.043 (0.040) [0.279]	-0.042 (0.043) [0.324]
Major items in ancillary projects	0.268 (0.025)	0.255 (0.042)	-0.013 (0.043) [0.761]	-0.015 (0.041) [0.715]	-0.004 (0.043) [0.924]
Distributed via schools					
Major items in road	0.252 (0.026)	0.239 (0.045)	-0.009 (0.050) [0.8549]	-0.014 (0.049) [0.777]	-0.003 (0.045) [0.950]
Major items in ancillary projects	0.268 (0.025)	0.216 (0.038)	-0.048 (0.044) [0.282]	-0.051 (0.044) [0.250]	-0.056 (0.039) [0.155]
B. Invitations plus comments:					
Distributed via neighborhood heads					
Major items in roads	0.252 (0.026)	0.278 (0.035)	0.025 (0.036) [0.483]	0.038 (0.037) [0.299]	0.022 (0.041) [0.602]
Major items in ancillary projects	0.268 (0.025)	0.277 (0.037)	0.010 (0.039) [0.792]	0.024 (0.039) [0.539]	0.023 (0.040) [0.569]
Distributed via school					
Major items in roads	0.252 (0.026)	0.179 (0.035)	-0.070* (0.041) [0.093]	-0.086** (0.038) [0.024]	-0.052 (0.036) [0.150]
Major items in Ancillary projects	0.268 (0.025)	0.198 (0.033)	-0.064 (0.042) [0.127]	-0.077* (0.040) [0.055]	-0.078* (0.041) [0.056]

See the note to table 11. Treatment effects and p-values are computed from a regression of the dependent variable on a dummy for audit treatment and four dummies for the participation treatments interacted with distribution mechanism. Percent missing equals log reported value - log actual value.

3. Extension 1: bottom-up monitoring on nepotism

In table 8 from the original study, Olken estimated the effect of audit monitoring on nepotism to see whether the decrease in missing expenditures led to an increase in alternative forms of corruption. From this he found that audited villages do observe an increase in nepotism in the sense of the probability of working if one is related to the village head or project head. Olken didn't, however, observe whether there is an overall increase in nepotism from bottom-up monitoring. In this extension of the replication paper I thus test whether there was an oberved effect on alternative forms of corruption from the increased bottom-up monitoring, regardless whether these bottom-up monitoring had an effect on missing expenditures or not.

My hypothesis regarding this is that, with grassroot monitoring there should be less nepotism from being related to village heads. When the monitoring comes directly from the other village members and them only, it will be harder for the village elite to get away with favourably employing their own family members ahead of others. For this hypothesis to hold, I would expect a negative effect from being related to any heads on the probability of working.

The results from this extension is presented in table 13 below. There is no clear effect of nepotism from bottom-up monitoring when workers are related to village heads, since the corresponding coefficients (column 1 and 3) are both insignificant. Surprisingly though is that, while the effects from being related to the project head are both significant on a five percent level, the effects goes in opposite directions depending on if it is through invitations (column 2) or invitations plus comments (column 4). For villages with invitations, being related to the project head is associated with a 13 percent lower probability of working in the project. For villages with invitations plus comments however, this probability is instead positive at 14 percent. While both effects are significant, the fact that they move in different directions make it unclear what conclusions to actually draw from the effect of participation monitoring on nepotism. I am further unsure what specifically could explain the mechanism behind the individual significant effects found, and what specifically explains the opposite effects of them.

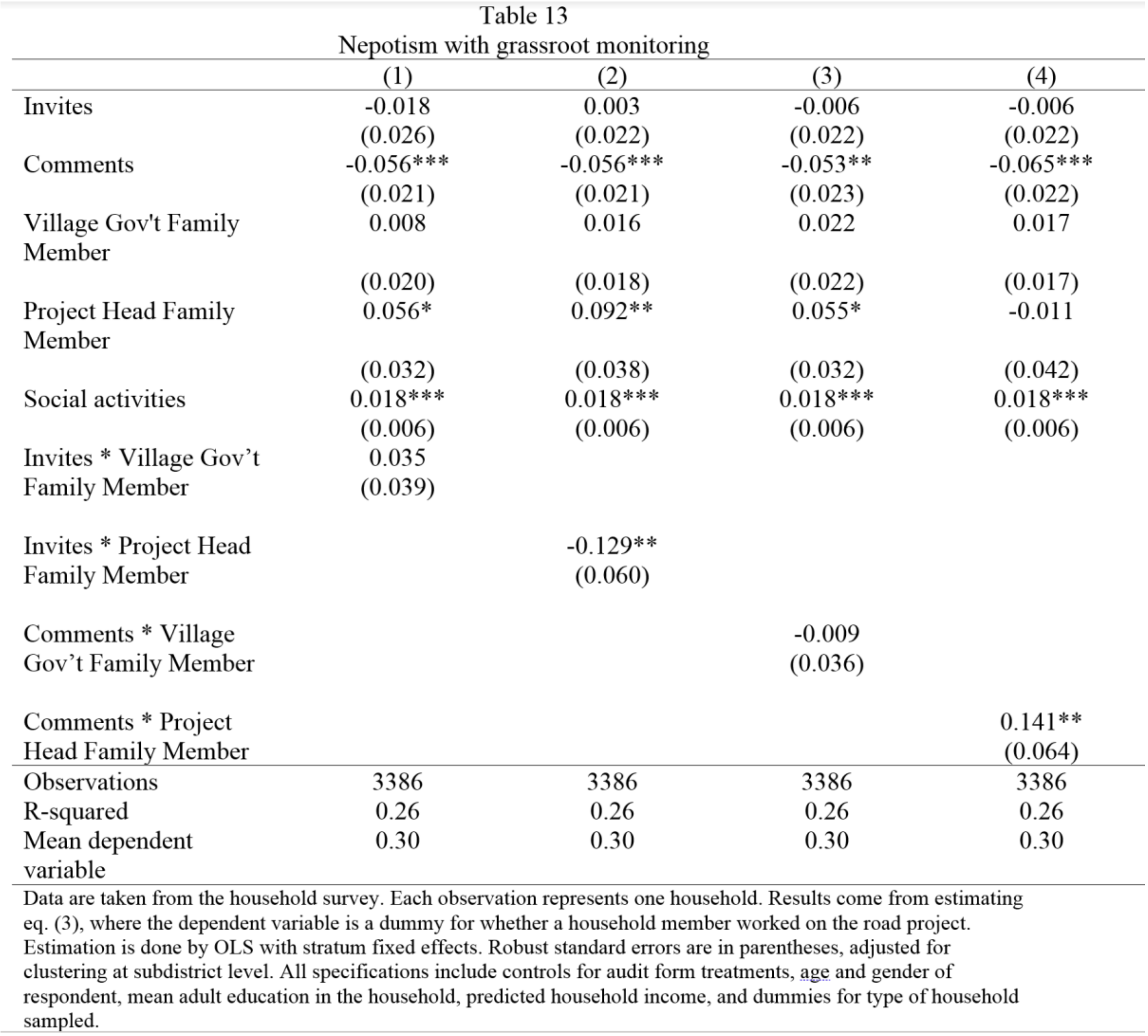


Table 13
Nepotism with grassroot monitoring

	(1)	(2)	(3)	(4)
Invites	-0.018	0.003	-0.006	-0.006
	(0.026)	(0.022)	(0.022)	(0.022)
Comments	-0.056***	-0.056***	-0.053**	-0.065***
	(0.021)	(0.021)	(0.023)	(0.022)
Village Gov't Family Member	0.008	0.016	0.022	0.017
	(0.020)	(0.018)	(0.022)	(0.017)
Project Head Family Member	0.056*	0.092**	0.055*	-0.011
	(0.032)	(0.038)	(0.032)	(0.042)
Social activities	0.018***	0.018***	0.018***	0.018***
	(0.006)	(0.006)	(0.006)	(0.006)
Invites * Village Gov't Family Member	0.035			
	(0.039)			
Invites * Project Head Family Member		-0.129**		
		(0.060)		
Comments * Village Gov't Family Member			-0.009	
			(0.036)	
Comments * Project Head Family Member				0.141**
				(0.064)
Observations	3386	3386	3386	3386
R-squared	0.26	0.26	0.26	0.26
Mean dependent variable	0.30	0.30	0.30	0.30

Data are taken from the household survey. Each observation represents one household. Results come from estimating eq. (3), where the dependent variable is a dummy for whether a household member worked on the road project. Estimation is done by OLS with stratum fixed effects. Robust standard errors are in parentheses, adjusted for clustering at subdistrict level. All specifications include controls for audit form treatments, age and gender of respondent, mean adult education in the household, predicted household income, and dummies for type of household sampled.

To try and find more answers regarding this, I ran the same regressions again for how being family member of a head might affect the probability of working for villages with bottom-up monitoring. However, this time I instead *combined the two separate bottom-up monitoring variables*, ”invitations” and ”invitations plus comments”, *into one single variable* that I call ”invitations or comments”. Apart from this change, the regressions are exactly the same as the ones presented in table 13. The strength of doing this new procedure is that it increases power in the estimation, while the drawback is that using one variable instead of two separate ones will decrease the variation that is explained in the coefficient.

The results for using one combined variable for grassroot monitoring is presented in table 14. The coefficients for all interaction terms are insignificant and I am thus unable to draw any further conclusion whether there actually is an effect from grassroot monitoring on nepotism. Furthermore, in the original study the author only reported any results of nepotism from audit monitoring, while ignoring the mechanism from grassroot monitoring whatsoever. The reason why he only reported on the effect from audits alone could possibly be explained by three things. (i) he either forgot to check for the corresponding effect from grassroot monitoring, (ii) since there was only a small decreased effect of bottom-up monitoring on missing expenditures, it might not have been that worthwile to test if participation actually had an effect on nepotism, or (iii) simply because he ran similar tests as the ones I just did, and decided to not report on it once he might have realized that there is no discernible overall effect from participation.

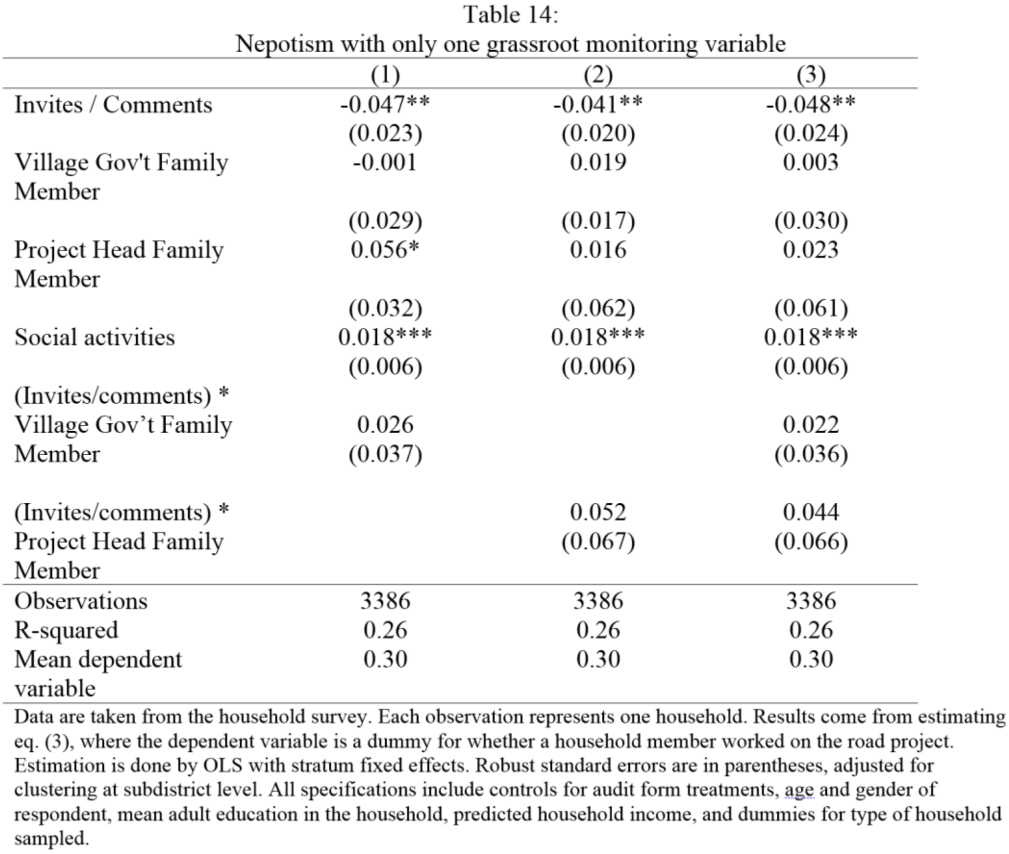


Table 14:
Nepotism with only one grassroot monitoring variable

	(1)	(2)	(3)
Invites / Comments	-0.047**	-0.041**	-0.048**
	(0.023)	(0.020)	(0.024)
Village Gov't Family Member	-0.001	0.019	0.003
	(0.029)	(0.017)	(0.030)
Project Head Family Member	0.056*	0.016	0.023
	(0.032)	(0.062)	(0.061)
Social activities	0.018***	0.018***	0.018***
	(0.006)	(0.006)	(0.006)
(Invites/comments) * Village Gov't Family Member	0.026 (0.037)		0.022 (0.036)
(Invites/comments) * Project Head Family Member		0.052 (0.067)	0.044 (0.066)
Observations	3386	3386	3386
R-squared	0.26	0.26	0.26
Mean dependent variable	0.30	0.30	0.30

Data are taken from the household survey. Each observation represents one household. Results come from estimating eq. (3), where the dependent variable is a dummy for whether a household member worked on the road project. Estimation is done by OLS with stratum fixed effects. Robust standard errors are in parentheses, adjusted for clustering at subdistrict level. All specifications include controls for audit form treatments, age and gender of respondent, mean adult education in the household, predicted household income, and dummies for type of household sampled.

4. Extension 2: Romano-Wolf multiple hypothesis correction

The risk of family-wise error rate (FWER) (that is, the probability of having at least one type I error in a series of hypothesis) increases when dealing with multiple hypotheses testing. In these cases, unless one considers the multiplicity of the testing procedure, there might be over-rejection of the null hypotheses. One way of controlling for the FWER is through The Romano-Wolf multiple hypothesis correction.

The study of Olken (2007) deals with multiple hypotheses in many of the specifications. Most notoriously is equation (1) in the study (presented in table 4 and 11), which uses four separate dependent variables for missing expenditures, while also using three different inpependent variables. One could argue for specifically how many hypotheses that would be needed before multiple hypotheses correction procedures becomes suitable to use. However, I would argue here that the use of the multiple dependent variables (thus also the many null hypotheses that arise from these) invites the use for multiple hypothesis correction as a way of controlling for the FWER.

More specifically, the Romano-Wolf correction deals with the FWER through resampling process (similar to the bootstrapping procedure (Westfall and Young, 1993)) and step-down procedures (similar to the Holm and Bonferroni correction (Holm, 1979). Compared to these other correction procedures however, the Romano-Wolf correction is considered to have more power (that is, the capacity to correctly reject false null hypotheses) while also controlling for dependence among p-values in the multiple test. These reasons motivates the use of Romano-Wolf over other correction procedures (Clarke et al., 2020).

In models, we generally test the probability of rejecting the null hypothesis against the bounded significance level alpha (usually set at alpha=0.05). When reporting on p-values, such as in in the case of the original study, one typically test the probability of rejection under one null hypothesis at a time. The Romano-Wolf procedure differs here from the general procedure since it extends this setting to instead test for multiple null hypothesis at the same time. Therefore, the probability of having at least one type I error among the multiple hypotheses will be controlled for when accounting controlling for many nulls, as opposed to when controlling one null at a time (Clarke et al., 2020).

The Romano-Wolf correction is run only for eq. (1) in Olken (2007), whose results are the ones presented in table 4 (the main results of audits on missing expenditures). I chose to focus on the p-values from table 4 specifically, since it is from this table the author draws his main conclusions of corruption in Indonesia. Moreover, I ran the test only for column 3 (no fixed effects) and 4 (engineer fixed effects). The final column (with stratum fixed effects) I excluded due to the already very high p-values of the coefficients. The results of the adjusted p-values from the Romano-Wolf correction is displayed below in table 15.

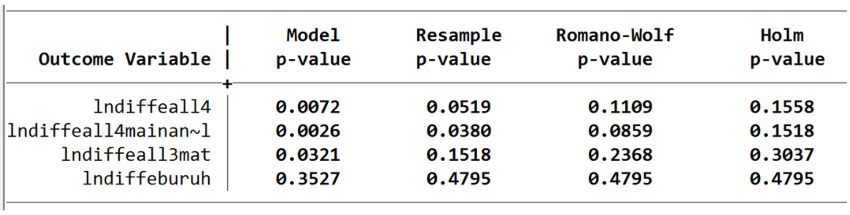
Panel A reports the p-values from column 3 in table 4, with no fixed effects, while panel B displays from column 4 in table 4, with engineer fixed effects. All four missing expenditure variables were run (however not properly labeled here), using audits as the independent variable, and having invitations and invitations plus comments as controls. I used 1000 bootstrap resamples for each correction test, as recommended in Clarke et al. (2020). One can of course use another number for the amount of resampling, which will to some extent yield other p-values.

The first column, *model p-value*, presents the regular p-values from the model specified in Olken (2007). The next column, *resample p-value*, lists the adjusted p-values from the resampling process, which however do not correct for multiple testing. The column *Romano-Wolf p-value* list the p-values that are both resampled and corrected for multiple hypothesis testing. The final column, *Holm p-value*, report on p-values that are corrected for multiple hypotheses but that are not resampled though.

Table 15: Romano-Wolf correction

Panel A: no fixed effects

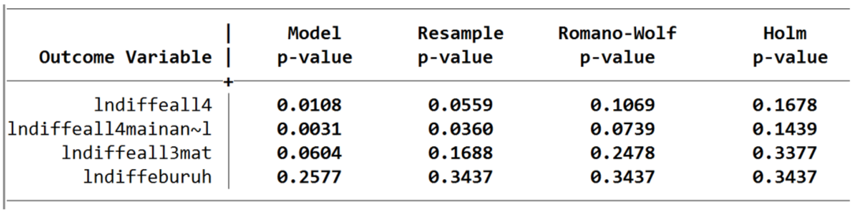
Controls for invitations and invitations plus comments were included in the -rwolf- command.



```
                     |     Model       Resample     Romano-Wolf        Holm
  Outcome Variable   |    p-value       p-value       p-value         p-value
---------------------+--------------------------------------------------------
          lndiffeall4|    0.0072        0.0519        0.1109          0.1558
lndiffeall4mainan~l  |    0.0026        0.0380        0.0859          0.1518
        lndiffeall3mat|    0.0321        0.1518        0.2368          0.3037
         lndiffeburuh|    0.3527        0.4795        0.4795          0.4795
```

Panel B: engineer fixed effects

Controls for invitations and invitations plus comments were included in the -rwolf- command.



```
                     |     Model       Resample     Romano-Wolf        Holm
  Outcome Variable   |    p-value       p-value       p-value         p-value
---------------------+--------------------------------------------------------
          lndiffeall4|    0.0108        0.0559        0.1069          0.1678
lndiffeall4mainan~l  |    0.0031        0.0360        0.0739          0.1439
        lndiffeall3mat|    0.0604        0.1688        0.2478          0.3377
         lndiffeburuh|    0.2577        0.3437        0.3437          0.3437
```

First thing to note is that, under the column *model p-value* , the displayed model p-values do not actually match those in table 4 of the original study nor in the replication. For the first two rows in panel A, the audit coefficient is significant on a 99 percent level, while they in the study were on a 5 percent level at most (p=.034 and p=.058). Moreover, the third row coefficient is here significant on a 5 percent level, while in the original study it was not significant at all (p=.143). Therefore, since the p-values do not match completely between this test and in the replication, they should be interpreted as speculative rather than facts. Another thing to note is that the p-values in the column *resample p-value* are those that are most similar to those on the original study. For these reasons, I ran additional specifications of the Romano-Wolf correction (these runs are presented in Appendix A).

As shown in the table when comparing the p-values to the Romano-Wolf ones, three of the model p-values are significant. But for the adjusted Romano-Wolf p-values most turn insignificant, even at the 10 percent level. The only coefficient that remains significant, in both panels, is the one for missing expenditures in road and ancillary projects (1ndiffeall4mainancil). These values indicate that the coefficients lose most of their significance when controlling for the FWER. Moreover, the probabilty of rejecting at least one true null hypothesis oversteps the limit for the 10 percent significance level for three of the four dependent variables. It is however doubtful whether these adjusted p-values actually tell us anything conclusive regarding the main results of the study. Most notably since the model p-values do not match those in the actual study. They do however imply that the probability of having at least one type I error increases, and quite significantly, when testing for multiple nulls at the same time.

I further present graphs for the null distributions for the multiple tests (figure 1 and 2). The most noteworthy information they add, apart from the list of p-values already presented, is that they give a picture of how much more demanding the Romano-Wolf correction is compared to the unadjusted p-value test. The figures display the distribution from a one-sided test[2]: the solid red line corresponds to the observed coefficient, the grey area show the actual distribution of the null hypotheses, and the dotted line is the theoretical distribution of null hypotheses. As shown in the top left panel, the mismatch between the actual distribution and the theoretical distributions of the nulls implies that the first null distribution is more demanding than the theoretical one. In the top right, and after in the bottom left, the match between the distributions is better since the previously tested variables are removed after each test, and thus each test become less demanding.

[2] In Clarke et al. (2020), they labeled these as two-sided tests however.

The last test, shown in the bottom right, presents similar distributions due to the variable only being tested from bootstrap replications of itself.

Figure 1

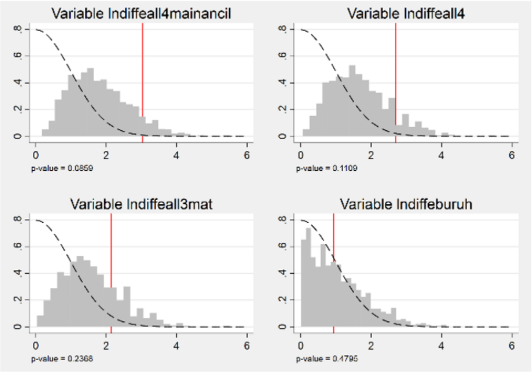


Figure 2

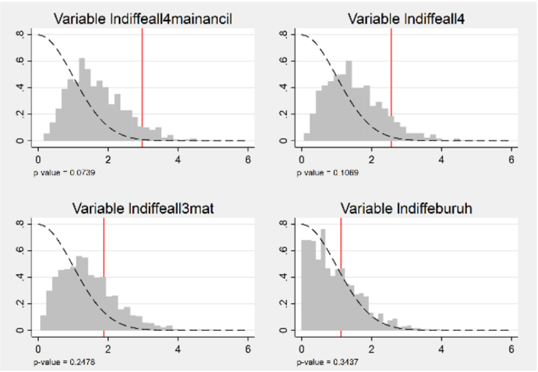


Figure 1 (left): Null distributions of one-sided tests with no fixed effects
Figure 2 (right): Null distributions of one-sided tests with engineer team fixed effects

5. Conclusion

The purpose of this paper was to replicate the results from Olken (2007) and check the robustness of the study. There were coding mistakes present and a couple of variables that were mislabeled in the author's code which led to confusion while working on this paper. Despite this, the replication was successful. Furthermore, this paper included some robustness checks, primarily for table 4, 6 and 8.

The overall replication and the added robustness checks point toward similar results of the original study, which is that on the one hand audit monitoring is associated with a decrease in missing expenditures in road projects, but also in road and ancillary projects. On the other hand, participatory monitoring through invitations and comment forms distributed in villages does result in a small, although not significant, decrease in missing expenditures. As the author concluded, this may indicate that top-down monitoring could be more effective than a bottom-up approach in a setting where the purpose is to decrease corruption. From the findings of this replication paper, I would therefore confirm that the results of Olken (2007) are robust.

Furthermore, it is worth mentioning that the decrease in corruption due to audits is followed by an increase in nepotism. Since the author did not include a similar test for whether bottom-up monitoring affects nepotism, I ran such tests in this replication paper. Although these tests did involve significant results, the opposed direction of these individual effects causes uncertainty how to interpret this, or whether there actually is such an effect present for bottom-up monitoring.

References

Clarke, D., Romano, J.P. and Wolf, M., 2020. The Romano–Wolf multiple-hypothesis correction in Stata. *The Stata Journal*, *20*(4), pp.812-843.

Holm, S., 1979. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pp.65-70.

Olken, B.A., 2007. Monitoring corruption: evidence from a field experiment in Indonesia. *Journal of political Economy*, *115*(2), pp.200-249.

Westfall, P.H. and Young, S.S., 1993. *Resampling-based multiple testing: Examples and methods for p-value adjustment* (Vol. 279). John Wiley & Sons.

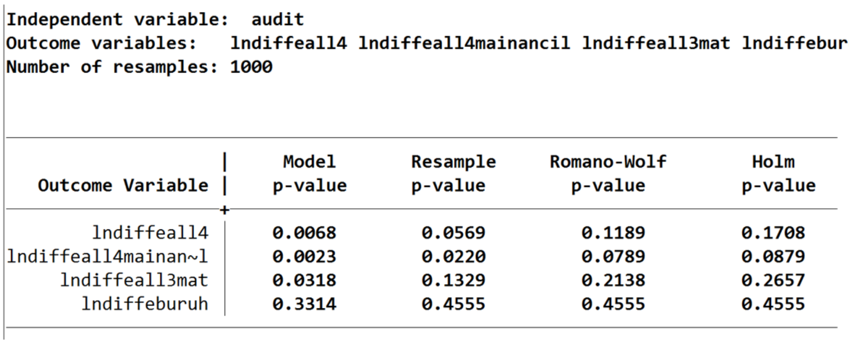
Appendix A:

Romano-Wolf multiple hypothesis correction, other specifications

Panel A: no fixed effects

Controls for invitations and invitations plus comments were included in the -rwolf- command.

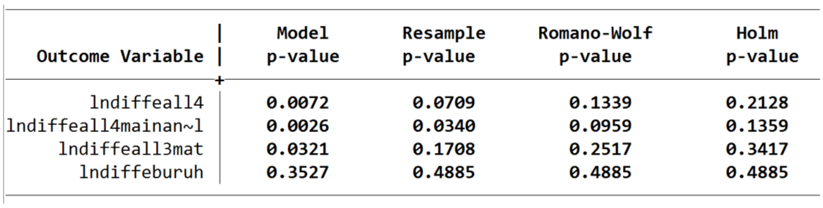
a. no controls included in the -rwolf- command (4 hypotheses)



```
Independent variable:  audit
Outcome variables:   lndiffeall4 lndiffeall4mainancil lndiffeall3mat lndiffeburuh
Number of resamples: 1000


------------------------------------------------------------------------------------
                    |    Model         Resample       Romano-Wolf        Holm
   Outcome Variable |   p-value        p-value          p-value         p-value
--------------------+---------------------------------------------------------------
        lndiffeall4 |   0.0068         0.0569           0.1189          0.1708
lndiffeall4mainan~l |   0.0023         0.0220           0.0789          0.0879
     lndiffeall3mat |   0.0318         0.1329           0.2138          0.2657
       lndiffeburuh |   0.3314         0.4555           0.4555          0.4555
------------------------------------------------------------------------------------
```

b. invitations and invitations plus comments included as independent variables in the command (12 hypotheses)

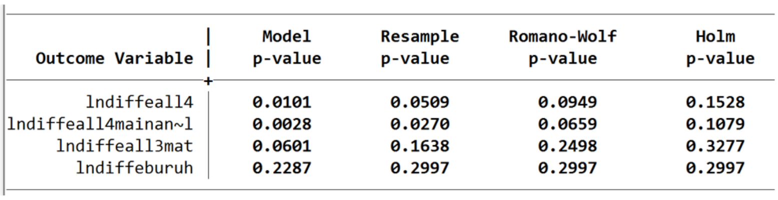


```
------------------------------------------------------------------------------------
                    |    Model        Resample      Romano-Wolf        Holm
   Outcome Variable |   p-value        p-value        p-value         p-value
--------------------+---------------------------------------------------------------
        lndiffeall4 |   0.0072         0.0709         0.1339          0.2128
lndiffeall4mainan~l |   0.0026         0.0340         0.0959          0.1359
     lndiffeall3mat |   0.0321         0.1708         0.2517          0.3417
       lndiffeburuh |   0.3527         0.4885         0.4885          0.4885
------------------------------------------------------------------------------------
```

Panel B: engineer fixed effects

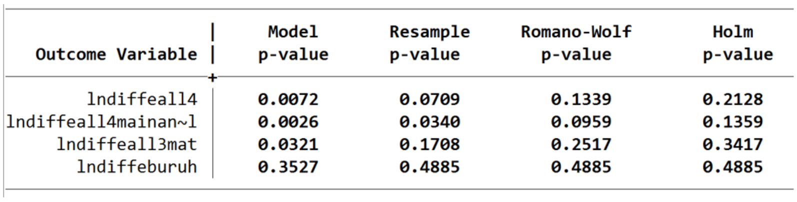
Controls for invitations and invitations plus comments were included in the -rwolf- command.

a. no controls included in the command (4 hypotheses)



```
------------------------------------------------------------------------------------
                    |    Model        Resample      Romano-Wolf        Holm
   Outcome Variable |   p-value        p-value        p-value         p-value
--------------------+---------------------------------------------------------------
        lndiffeall4 |   0.0101         0.0509         0.0949          0.1528
lndiffeall4mainan~l |   0.0028         0.0270         0.0659          0.1079
     lndiffeall3mat |   0.0601         0.1638         0.2498          0.3277
       lndiffeburuh |   0.2287         0.2997         0.2997          0.2997
------------------------------------------------------------------------------------
```

b. invitations and invitations plus comments included as independent variables in the command (12 hypotheses)



```
------------------------------------------------------------------------------------
                    |    Model        Resample      Romano-Wolf        Holm
   Outcome Variable |   p-value        p-value        p-value         p-value
--------------------+---------------------------------------------------------------
        lndiffeall4 |   0.0072         0.0709         0.1339          0.2128
lndiffeall4mainan~l |   0.0026         0.0340         0.0959          0.1359
     lndiffeall3mat |   0.0321         0.1708         0.2517          0.3417
       lndiffeburuh |   0.3527         0.4885         0.4885          0.4885
------------------------------------------------------------------------------------
```