IX: Randomized Field Experiments in Practice III

Alexandra Avdeenko

Evaluation

2022

Overview

Field Experiment Example

- Road construction in Indonesia
- Treatments
- Randomization
- Measuring corruption
- Estimation and results
- Interpreting experimental results

Designing Experiments: 'Theory' and Practice

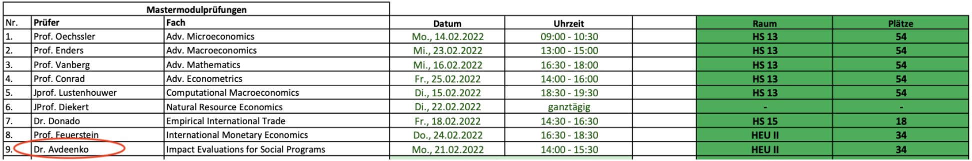
- Randomization and treatment design
- Sample size and statistical power
- Implementation issues

Selected Challenges

- Imperfect compliance
- Other potential problems

Concluding Remarks

Exam



Mastermodulprüfungen							
Nr.	Prüfer	Fach	Datum	Uhrzeit		Raum	Plätze
1.	Prof. Oechssler	Adv. Microeconomics	Mo., 14.02.2022	09:00 - 10:30		HS 13	54
2.	Prof. Enders	Adv. Macroeconomics	Mi., 23.02.2022	13:00 - 15:00		HS 13	54
3.	Prof. Vanberg	Adv. Mathematics	Mi., 16.02.2022	16:30 - 18:00		HS 13	54
4.	Prof. Conrad	Adv. Econometrics	Fr., 25.02.2022	14:00 - 16:00		HS 13	54
5.	Jprof. Lustenhouwer	Computational Macroeconomics	Di., 15.02.2022	18:30 - 19:30		HS 13	54
6.	JProf. Diekert	Natural Resource Economics	Di., 22.02.2022	ganztägig		-	-
7.	Dr. Donado	Empirical International Trade	Fr., 18.02.2022	14:30 - 16:30		HS 15	18
8.	Prof. Feuerstein	International Monetary Economics	Do., 24.02.2022	16:30 - 18:30		HEU II	34
9.	Dr. Avdeenko	Impact Evaluations for Social Programs	Mo., 21.02.2022	14:00 - 15:30		HEU II	34

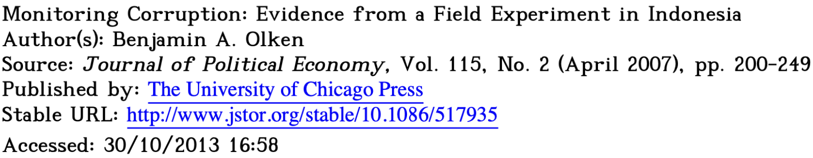
Content

90 minutes

- Part 1: About 15 Multiple choice questions
- Part 2: Questions on impact evaluation methods (book + lecture slides; if needed additional read-up from bibliography) and studies discussed in class (with focus on understanding the methods)
- Part 3: In depth questions on three problem set studies.

Recommendation: Read the book! Review the slides, see that you understand the content, and the three studies in more detail. Do not forget: Send questions prior to QA sessions with Charlotte and myself.

Study



Monitoring Corruption: Evidence from a Field Experiment in Indonesia
Author(s): Benjamin A. Olken
Source: *Journal of Political Economy*, Vol. 115, No. 2 (April 2007), pp. 200-249
Published by: The University of Chicago Press
Stable URL: http://www.jstor.org/stable/10.1086/517935
Accessed: 30/10/2013 16:58

Link to PDF

A prominent field experiment: Monitoring corruption (Olken 2007)

Topic, issue, questions

Topic	corruption in developing countries
Issue	how to curb diversion of public funds in local construction
Q 1	do audits lower the share of diverted funds in grant-funded local construction projects?
Q 2	does enhanced grass-root participation lower the diversion of funds?

Setting or 'case'

- 15,000 Indonesian villages in Subdistrict Development Project (KDP)
- here, 608 receive central government grants for local construction projects, primarily village roads
- around 9,000USD per village, which is large compared to other public expenditure
- implementation of projects in the hands of village leadership
- audits by development agency (BPKP) not unheard of (4% baseline probability)
- village meetings serve as forum for planning and monitoring

Monitoring corruption (2/7)

Three treatments

T1:	*audits*	100% audit probability announced (prior) and implemented (during or after)
T2:	*invites*	substantial increase of written invitations to village meetings (plus 300-500)
T3:	*inv. + comments*	treatment 2 plus anonymous comment forms (sent along with invite)

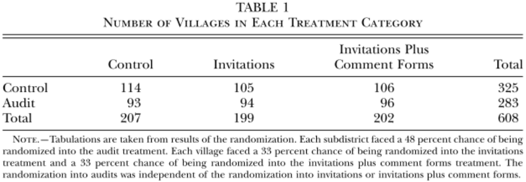


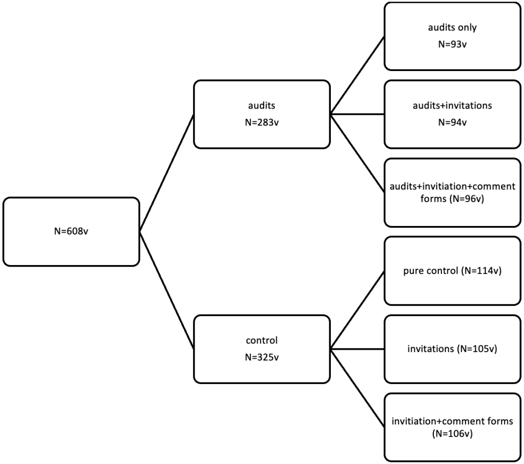
TABLE 1
NUMBER OF VILLAGES IN EACH TREATMENT CATEGORY

	Control	Invitations	Invitations Plus Comment Forms	Total
Control	114	105	106	325
Audit	93	94	96	283
Total	207	199	202	608

NOTE.—Tabulations are taken from results of the randomization. Each subdistrict faced a 48 percent chance of being randomized into the audit treatment. Each village faced a 33 percent chance of being randomized into the invitations treatment and a 33 percent chance of being randomized into the invitations plus comment forms treatment. The randomization into audits was independent of the randomization into invitations or invitations plus comment forms.

→ in this study assignment of T1 independent from T2 and T3

Monitoring corruption (2/7): Treatment Arms



N=608v

audits N=283v

control N=325v

audits only N=93v

audits+invitations N=94v

audits+invitiation+comment forms (N=96v)

pure control (N=114v)

invitations (N=105v)

invitiation+comment forms (N=106v)

Monitoring corruption (3/7)

Randomization issues

Randomization Level

- audit treatment at subdistrict level
  - means: either all or no village in subdistrict gets treated
  - reason: to avoid spill-over effect from audits in one village to others

Stratification

- by subdistrict (invites + comments), by district and duration in KDP (audits)
  - ensures that share of treated villages is equal in all subdistricts (i + c)
  - ensures that share of treated subdistricts with a given time in KDP is equal in all districts (audits)
  - *Important:* stratification ensures this to be true ex post, not just in expectation as randomization would

Monitoring corruption (4/7)

Measuring corruption

- perception-based indicators common but unrealiable
- here, measure difference between reported and real road construction cost
- main drivers of real cost are
  - construction material → road composition samples
  - amount of paid labour → ask workers and foremen
  - input prices → ask workers, suppliers, procurers, etc.
- these data are very difficult to obtain and estimate

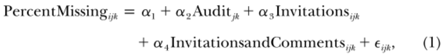
⇒ corruption measure in Olken's study is

$$\textit{percent missing} = \log(\textit{reported expenditure}) - \log(\textit{real expenditure})$$

⇒ use several variants of that measure (only road cost, cost of all grant-funded projects, only labour cost, only material cost)

Monitoring corruption (5/7)

Estimation



$$\text{PercentMissing}_{ijk} = \alpha_1 + \alpha_2\text{Audit}_{jk} + \alpha_3\text{Invitations}_{ijk} + \alpha_4\text{InvitationsandComments}_{ijk} + \epsilon_{ijk}, \qquad (1)$$

- cluster standard errors at subdistrict level because of level of randomization
- additional controls:
    - engineering team fixed effects
    - stratum fixed effects

Monitoring corruption: Table (6/7)

Main results: 'percent missing' (OLS)

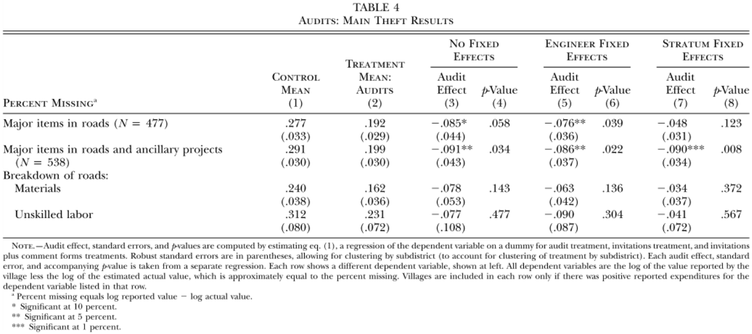


TABLE 4
AUDITS: MAIN THEFT RESULTS

			No Fixed Effects		Engineer Fixed Effects		Stratum Fixed Effects	
Percent Missing[a]	Control Mean (1)	Treatment Mean: Audits (2)	Audit Effect (3)	p-Value (4)	Audit Effect (5)	p-Value (6)	Audit Effect (7)	p-Value (8)
Major items in roads (N = 477)	.277	.192	−.085*	.058	−.076**	.039	−.048	.123
	(.033)	(.029)	(.044)		(.036)		(.031)	
Major items in roads and ancillary projects (N = 538)	.291	.199	−.091**	.034	−.086**	.022	−.090***	.008
	(.030)	(.030)	(.043)		(.037)		(.034)	
Breakdown of roads:								
Materials	.240	.162	−.078	.143	−.063	.136	−.034	.372
	(.038)	(.036)	(.053)		(.042)		(.037)	
Unskilled labor	.312	.231	−.077	.477	−.090	.304	−.041	.567
	(.080)	(.072)	(.108)		(.087)		(.072)	

NOTE.—Audit effect, standard errors, and p-values are computed by estimating eq. (1), a regression of the dependent variable on a dummy for audit treatment, invitations treatment, and invitations plus comment forms treatments. Robust standard errors are in parentheses, allowing for clustering by subdistrict (to account for clustering of treatment by subdistrict). 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.

[a] Percent missing equals log reported value − log actual value.

* Significant at 10 percent.

** Significant at 5 percent.

*** Significant at 1 percent.

Monitoring corruption: Findings (6/7)

- audits reduce 'percent missing' in road construction by ca 8 percentage points
- effects somewhat larger when considering all parts of grant-funded project
- given that 'percent missing' is about 27.7% in control villages, raising audit probability from 4% to 100% reduces level of diverted funds by 29%
- decomposition into effects on material and labour costs inconclusive

Monitoring corruption (7/7)

Graphical illustration of results

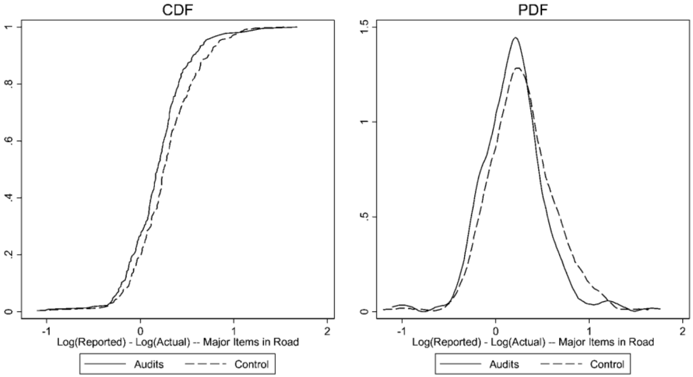


Fig. 1.—Empirical distribution of missing expenditures. The left-hand figure shows the empirical CDF of missing expenditures for the major items in a road project, separately for villages in the audit treatment group (solid line) and the control group (dashed line). The right-hand figure shows estimated PDFs of missing expenditures for both groups; PDFs are estimated using kernel density regressions using an Epanechnikov kernel.

Interpreting experimental results on 'complex' processes

Overall versus ceteris paribus effects

- effect of D contains effect of treatment itself as well as effects of any responses to the treatment
- example of cash transfers to schools $\rightarrow$ parents may respond by lowering education-related expenditure

Implications

- ATE from experimental treatment estimates overall effect, not ceteris paribus effect
- mechanism decomposition difficult ex post
  - consider nepotism results in controlling corruption example
  - we only know that audits reduced fund-diversion, and that they enhanced employment of family members
  - we do not know if the overall effect is partially due to trusted family members being less corrupt, or it is the positive net effect of audits despite increased nepotism

→ disentangling that would have required ex ante theorizing and collection of respective data

Excursus: Nepotism results from Olken (2007)

Working for pay (linear probability)

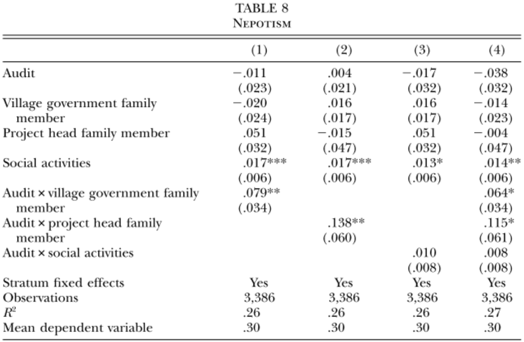


TABLE 8
NEPOTISM

	(1)	(2)	(3)	(4)
Audit	−.011 (.023)	.004 (.021)	−.017 (.032)	−.038 (.032)
Village government family member	−.020 (.024)	.016 (.017)	.016 (.017)	−.014 (.023)
Project head family member	.051 (.032)	−.015 (.047)	.051 (.032)	−.004 (.047)
Social activities	.017*** (.006)	.017*** (.006)	.013* (.006)	.014** (.006)
Audit × village government family member	.079** (.034)			.064* (.034)
Audit × project head family member		.138** (.060)		.115* (.061)
Audit × social activities			.010 (.008)	.008 (.008)
Stratum fixed effects	Yes	Yes	Yes	Yes
Observations	3,386	3,386	3,386	3,386
R^2	.26	.26	.26	.27
Mean dependent variable	.30	.30	.30	.30

Types of field experiments and treatment design

- classic randomization ('trial')
  - → some get treated, some don't, no questions asked, no fussing around

Types of field experiments and treatment design

- classic randomization ('trial')
  - → some get treated, some don't, no questions asked, no fussing around
- randomized piloting or phase-in
  - → some groups are treated earlier, but eventually all groups get the treatment
  - ⇒ often lowers ethical concerns
  - ⇒ beware of anticipation effects

Types of field experiments and treatment design

- classic randomization (‘trial’)
  - → some get treated, some don’t, no questions asked, no fussing around
- randomized piloting or phase-in
  - → some groups are treated earlier, but eventually all groups get the treatment
  - ⇒ often lowers ethical concerns
  - ⇒ beware of anticipation effects
- encouragement designs
  - → useful if availability of treatment is universal but take-up is not
  - → administered ‘treatment’ is only encouragement to opt for actual treatment
  - ⇒ difference between those receiving encouragement and the actually treated leads to complications in estimation (see below)

Types of field experiments and treatment design

- classic randomization ('trial')
  - → some get treated, some don't, no questions asked, no fussing around
- randomized piloting or phase-in
  - → some groups are treated earlier, but eventually all groups get the treatment
  - ⇒ often lowers ethical concerns
  - ⇒ beware of anticipation effects
- encouragement designs
  - → useful if availability of treatment is universal but take-up is not
  - → administered 'treatment' is only encouragement to opt for actual treatment
  - ⇒ difference between those receiving encouragement and the actually treated leads to complications in estimation (see below)
- oversubscription designs
  - → randomly admit some marginal rejects in addition to those treated anyhow
  - ⇒ minimal ethical concerns or interference with existing measures

Assignment modes with multiple treatments (I)

- suppose there are two treatments of interest, A and B
- three possibilities to form groups

Assignment modes with multiple treatments (II)

	$D^A = 0$	$D^A = 1$
Joint Treatments		
$D^B = 0$	control group	
$D^B = 1$		treatment group ($T^{A\ \mathit{and}\ B}$)
Multiple Treatments		
$D^B = 0$	control group	$T^{A\ \mathit{only}}$ group
$D^B = 1$	$T^{B\ \mathit{only}}$ group	
Cross-cutting Treatments		
$D^B = 0$	control group	$T^{A\ \mathit{only}}$ group
$D^B = 1$	$T^{B\ \mathit{only}}$ group	$T^{A\ \mathit{and}\ B}$ group

Assignment modes with multiple treatments (III)

- multiple treatment option allows to assess relative effectiveness of each treatment
- cross-cutting design allows additionally to investigate interactions between treatments

The level of randomization

- is not always smallest unit of observation
    - risk of spill-overs
    - e.g. envy leading to non-compliance
    - individual-level randomization may be unfeasible or uneconomical due to fixed cost
- randomization at a higher level
    - can minimize spill-overs
    - but affects statistical power, and thus sample size and budget
- in the corruption study earlier
    - the audit treatment was randomized at the subdistrict level because of feared spill-overs
    - the invites+comments treatment was randomized at the village level

The statistical power of an experiment...

- is the probability that we reject the H_0 (‘no effect’) for a given real effect size and significance level
- alternatively, think of power in terms of the *minimum detectable effect size (MDE)*
- when designing an experiment, power calculation is crucial to determine required number of subjects and randomization strategy
- won’t cover it in detail here
  - → there are online calculators
  - → still involves a lot of guess work (how to gauge intra-group correlation of outcome)

Design factors that affect power are. . .

- number of subjects and share of T versus C groups
- group-level treatment
  - → rule of thumb: increasing number of groups better than increasing number of subjects per group
- imperfect compliance
  - → rule of thumb: rate of non-compliance lowers power by more than number of observations increases it
- control variables
  - → trade-off between reducing variance versus losing degrees of freedom
  - → baseline value of Y is always a good control to have
- stratification
  - → generally more effective than including controls

Issues of implementation. . .

- often reliance on partner organizations
  - governments
  - NGOs
  - firms
- pilots are often windows of opportunity
  - partners usually motivated and funding in place
  - however, often less control over design
  - beware of opportunistic phasing-in (non-random piloting)
- biggest problems are money and good, reliable staff
- one option is to buy 'randomista' expertise (J-PAL at MIT, IPA at Yale, ...)

Imperfect compliance

Two main reasons for imperfect compliance

1. it may be that subjects in the control group receive the treatment
   - spill-overs (envy or 'desperation')
   - strategic action (defiance if subjects resent being experimented with)
   - the treatment might be available elsewhere

Imperfect compliance

Two main reasons for imperfect compliance

1. it may be that subjects in the control group receive the treatment
   - spill-overs (envy or 'desperation')
   - strategic action (defiance if subjects resent being experimented with)
   - the treatment might be available elsewhere
2. it may also be that some subjects in the treatment group to not receive treatment
   - they refuse
   - the mistakenly miss out
   - maybe implementation was disturbed
   - encouragement designs do not even attempt to treat all
   - encouragement designs only aim at affecting the probability of subjects to receive the actual treatment

⇒ there is thus a difference between 'intention to treat' and actual treatment

⇒ has implications for estimation of causal effects

Intention to treat

Let's therefore distinguish between

- the randomized action, Z ⇒ denote being in treatment group with $Z = T$, $Z = C$ otherwise
- and the actual treatment, D ⇒ denote actual treatment received with $D = 1$, $D = 0$ otherwise

By comparing the means of the treatment versus control groups, we obtain

$$E[Y|Z = T] - E[Y|Z = C].$$

This is the intention to treat estimate (ITT), but not the ATE, which is

$$E[Y|D = 1] - E[Y|D = 0].$$

→ The ITT is often highly policy relevant.

Policy makers affect Z, not D; what is rolled out as a programme is Z, not D; → but we may still want to estimate the causal effect of D; ⇒ the Wald estimator can do that

The Wald estimator

The shares of actually treated in the T and C groups are ex ante $E[D|Z=T]$ and $E[D|Z=C]$, or ex post π^T and π^C.

The Wald estimator

$$\hat{\beta}_{Wald} = \frac{E[Y|Z=T] - E[Y|Z=C]}{E[D|Z=T] - E[D|Z=C]} = \frac{'ITT'}{\pi^T - \pi^C}.$$

The Wald estimator

The shares of actually treated in the T and C groups are ex ante $E[D|Z = T]$ and $E[D|Z = C]$, or ex post π^T and π^C.

The Wald estimator

$$\hat{\beta}_{Wald} = \frac{E[Y|Z = T] - E[Y|Z = C]}{E[D|Z = T] - E[D|Z = C]} = \frac{'ITT'}{\pi^T - \pi^C}.$$

Under three assumptions

1. $E[d_i|z_i = T] \geq E[d_i|z_i = C]$ for every individual, or $E[d_i|z_i = T] \leq E[d_i|z_i = C]$ for every individual
2. any difference $(Y|Z = T) - (Y|Z = C)$ is due to Z
3. outcome Y is not directly affected by Z, only through D

the Wald estimator gives us the so called *local average treatment effect* (LATE)

See Duflo, Glennerster, Kremer (2006), pp. 48 ff. for the proof.

LATE and IV

1. first assumption requires that not all subjects need to be affected by Z, but those who are all need to be affected in the same 'direction' (*monotonicity*)
2. the second assumption requires that Z is ('as if') randomly assigned (thus, sometimes also called the *independence assumption*)
3. the third assumption is the same as the *exclusion restriction* for IVs

LATE and IV

1. first assumption requires that not all subjects need to be affected by Z, but those who are all need to be affected in the same 'direction' (*monotonicity*)
2. the second assumption requires that Z is ('as if') randomly assigned (thus, sometimes also called the *independence assumption*)
3. the third assumption is the same as the *exclusion restriction* for IVs

→ thus, with imperfect compliance, assignment to the treatment group ($Z = T$) works the same way as an instrumental variable

→ it 'exogenously' pushes marginally non-treated individuals into treatment

→ the Wald estimator gives us the causal effect of the treatment for those marginal individuals, which is the LATE

The local average treatment effect (LATE)

... is the effect of the treatment on those whose treatment status is changed by the instrument (the so called 'compliers'). Neither does it apply to all treated or untreated, nor to the entire sample (like the ATE does).

Other sources of problems with experiments

- probability of treatment differs by stratum
  - e.g. when a fixed number of treatments is assigned to strata with different numbers of subjects
  - implies that treatment is not random overall but random within each stratum
  - → conditioning and averaging over strata (weighted by treatment probabilities) can solve this (see lecture 1)
  - → use of OLS with a *saturated model* works, too
- externalities / spill-overs
  - means that SUTVA is violated (i's treatment effect independent of j's assignment)
  - rule of thumb: when spill-overs from T to C group are positive → underestimation of effects
- non-random attrition
  - over-time reduction in ability to collect data on certain subjects
  - problem even with equal attrition rates in T and C groups

Pros and cons of field experiments

Pro

- experiments are a powerful and arguably the cleanest way to 'identify' and estimate effects of causes
- most design problems are by now well-understood; large method toolbox
- social sciences hard to imagine without field experiments
- not only used in development research, also in firms, in public opinion, in campaigning and elections, ...

Pros and cons of field experiments

Con

- experiments are no panacea
- they are bad at uncovering causes of effects
- require careful ex ante theorizing (ex post mechanisms decomposition rarely possible)
- extremely resource-intensive to implement / design mistakes are costly
- often huge discrepancy between tested measures and rolled-out measures
- cost-benefit analysis?
- external validity / can't repeat every experiment everywhere