Field Experiments: Design, Analysis and Interpretation
Solutions for Chapter 6 Exercises

Alan S. Gerber and Donald P. Green[*]

Question 1

The following three quantities are similar in appearance but refer to different things. Describe the differences.

- $E[Y_i(d(1))|D_i = 1]$
  Answer:
  This expression refers to the expected potential outcome of Y_i given the treatment received by the assigned treatment group $D_i(1)$ for the subgroup of subjects who actually receive the treatment ($D_i = 1$).

- $E[Y_i(d(1))|d_i(1) = 1]$
  Answer:
  This expression refers to the expected potential outcome of Y_i given the treatment received by the assigned treatment group $D_i(1)$ for the subgroup of subjects who receive the treatment if assigned to it ($D_i(1) = 1$). In the case of one-sided non-compliance, this subgroup is the Compliers. For two-sided non-compliance this is composed of Always-Takers and Compliers.

- $E[Y_i(d(1))|d_i(1) = d_i(0) = 1]$
  Answer:
  This expression refers to the expected potential outcome of Y_i given the treatment received by the assigned treatment group $D_i(1)$ for the subgroup of subjects known as Always-Takers, who always receive the treatment regardless of whether they are assigned to the treatment group ($D_i(1) = D_i(0) = 1$).

Question 2

The following expression appears in the proof of the CACE theorem. Interpret the meaning of each term in the expression, and explain why the expression as a whole is equal to zero: $E[Y_i(d(1))|d_i(1) = d_i(0) = 0] - E[Y_i(d(0))|d_i(1) = d_i(0) = 0]$.
Answer:
The first expression refers to the expected potential outcome of Y_i given the treatment received by the assigned treatment group $D_i(1)$ for the subgroup of subjects known as Never-Takers, who never receive the treatment regardless of whether they are assigned to the treatment group ($D_i(1) = D_i(0) = 0$). The second expression refers to the expected potential outcome of Y_i given the treatment received by the assigned control group $D_i(0)$ for the subgroup of subjects known as Never-Takers, who never receive the treatment regardless of whether they are assigned to the

[*] Solutions prepared by Peter M. Aronow and revised by Alexander Coppock

treatment group ($D_i(1) = D_i(0) = 0$). These are equal (provided that the excludability assumption holds) because Never-Takers reveal the same untreated potential outcome regardless of treatment assignment.

Question 3

Assuming that the excludability and non-interference assumptions hold, are the following statements true or false? Explain your reasoning.

a) Among Compliers, the ITT equals the ATE.
Answer:
True. For Compliers, treatment assigned equals treatment received, and so ITT = ATE.

b) Among Defiers, the ITT equals the ATE.
Answer:
False: For Compliers, treatment assigned is the opposite of treatment received, and so ITT = -ATE.

c) Among Always-Takers and Never-Takers, the ITT and ATE are zero.
Answer:
False. For Always-takers and Never-takers, the ITT is zero because they respond the same to both experimental assignments. The ATE among these subgroups may not be nonzero; the ATE is not revealed empirically.

Question 4

When analyzing experiments with two-sided noncompliance, why is it incorrect to define Compliers as "those who take the treatment if assigned to treatment"?
Answer:
This definition should be "those who take the treatment if AND ONLY IF assigned to treatment." The definition given in the question would also hold for Always-takers, who take the treatment if assigned to treatment.

Question 5

Suppose that a sample contains 30% Always-Takers, 40% Never-Takers, 15% Compliers, and 15% Defiers. What is the ITT_D?
Answer:
Recall from equation (6.19): $ITT_D = \pi_C + \pi_{AT} - (\pi_D + \pi_{AT}) = \pi_C - \pi_D$, which in this case implies that the $ITT_D = 0$.

Question 6

Suppose that, in violation of the monotonicity assumption, a sample contains both Compliers and Defiers. Let π_C be the proportion of subjects who are Compliers, and let π_D be the proportion of subjects who are Defiers. Show that the CACE is nevertheless identified if (i) the ATE among Defiers equals the ATE among Compliers and (ii) $\pi_C \neq \pi_D$.

Answer:
Let's re-write equation (6.20) assuming that the two ATEs are the same and that the denominator is nonzero:

$$\frac{\text{ITT}}{\text{ITT}_\text{D}} = \frac{(\text{ATE} \mid \text{Compliers})\pi_C - (\text{ATE} \mid \text{Defiers})\pi_D}{(\pi_C - \pi_D)} = \frac{(\text{ATE} \mid \text{Compliers})(\pi_C - \pi_D)}{(\pi_C - \pi_D)} = (\text{ATE} \mid \text{Compliers}).$$

Question 7

In experiments with one-sided noncompliance, the ATE among subjects who receive the treatment (sometimes called the average treatment-on-the-treated effect, or ATT) is the same as the CACE, because only Compliers receive the treatment. Explain why the ATT is not the same as the CACE in the context of two-sided noncompliance.
Answer:
Under two-sided noncompliance, both Compliers and Always-takers receive treatment when assigned to the treatment group, and Always-takers receive treatment when assigned to the control group. Therefore, as we move from one-sided to two-sided noncompliance, "the treated" no longer refers to Compliers, and the ATT no longer equals the CACE.

Question 8

In the Milwaukee domestic violence experiment, researchers working in collaboration with police officers randomly assigned one of three treatments when officers responded to an incident involving domestic violence.[1] Officers were instructed to arrest the perpetrator and hold him overnight, arrest the perpetrator but release him after a brief period, or issue a warning. The full breakdown of assigned and actual treatments is presented in Table 6.7, along with observed rates of later arrest in the three treatment conditions.

Table 1: Question 8 Table

		Assigned Treatment		
		Full Arrest	Brief Arrest	Warning
Actual Treatment	Full Arrest	400	13	1
	Brief Arrest	1	384	1
	Warning	3	1	396
	Total N	404	398	398
Subsequent Outcomes	Calls to hotline to report perpetrator	296	301	261
	Perpetrators later arrested	146	157	151

a) Consider a simplified coding of the assigned and actual treatment, dividing subjects into two categories: arrest or non-arrest. Evaluate the plausibility of the non-interference, excludability, and monotonicity assumptions in this application.
Answer:
Non-interference: Potential outcomes reflect only the subject's own treatment status and not the status of other observations. It is possible, though unlikely, that perpetrators discuss their treatments with one another, in which case potential outcomes might be affected not only by

[1] Sherman et al.1992.

one's own treatment but whether it seems severe or lenient vis-a-vis other treatments that other subjects have received.

Excludability: Potential outcomes respond solely to receipt of the treatment and not the random assignment of the treatment or any indirect byproduct of random assignment (e.g., other actions that the responding police officer takes in addition to or instead of issuing a warning or making an arrest; for example, a police officer who is instructed to make no arrest might say/do other threatening things to compensate for the lenient punishment). There is no reason to believe that compensatory actions were taken in this study.

Monotonicity implies that there are no subjects who would be arrested if assigned to non-arrest and not arrested if assigned to arrest. This assumption seems plausible, as it is hard to imagine a scenario by which arrests occurr if and only if no arrest is assigned.

b) Assume that the core assumptions hold, and calculate the $\widehat{ITT_D}$, $\widehat{ITT}$, and $\widehat{CACE}$ given the simplified treatment categorization. Interpret the results.

```
In [1]: clear
        scalar pi_at = 2/398
        scalar pi_nt = (3+1)/(404+398)
        scalar pi_c = 1- pi_at - pi_nt
        scalar itt_d = pi_c
        disp %8.2f itt_d

    0.99
```

```
In [2]: scalar itt_hotline = (296+301)/(404+398) - (261/398)
        disp %8.5f itt_hotline

 0.08861
```

```
In [3]: scalar cace_hotline = itt_hotline/itt_d
        disp %8.5f cace_hotline

 0.08951
```

```
In [4]: scalar itt_arrest = (303)/(802) - (151/398)
        disp %8.6f itt_arrest

-0.001591
```

```
In [5]: scalar cace_arrest = itt_arrest/itt_d
        disp %8.6f cace_arrest

-0.001608
```

Let's first consider the effects on hotline calls. The estimated CACE of 0.09 means that among Compliers (those who are arrested if and only if assigned to arrest), an actual arrest appears to increase the probability of subsequent hotline calls by 9 percentage points. The estimated CACE of -0.002 for subsequent arrests means that among Compliers, actual arrest decreases the probability of subsequent arrest by a mere 0.2 percentage points.

c) Suppose monotonicity were not assumed. What do the results of the simplified treatment suggest about the maximum and minimum values of π_{NT}, π_{AT}, and π_C?
Answer:
Without assuming monotonicity, we return to what the observable quantities imply about the latent groups' shares of the subject pool.

Subjects who were treated when assigned to control: $\hat{\pi_{AT}} + \hat{\pi_D} = \frac{2}{398} = 0.00503$
Subjects who were not treated when assigned to control: $\hat{\pi_{NT}} + \hat{\pi_C} = \frac{396}{398} = 0.99497$
Subjects who were treated when assigned to treatment: $\hat{\pi_{AT}} + \hat{\pi_C} = \frac{401+397}{404+398} = 0.99501$
Subjects who were not treated when assigned to treatment: $\hat{\pi_{NT}} + \hat{\pi_D} = \frac{4}{404+398} = 0.00499$

The lower bounds for shares of Defiers, Always-takers, or Never-takers is 0. The last line enables us to put an upper bound on the share of Defiers and Never-takers: 0.00499, or 0.499%. The first line enables us to put an upper bound on the share of Always-takers: 0.503%. If Never-takers were as high as 0.499%, the share of Compliers cannot be lower than 0.99497 - 0.00499 = 0.98998. If there were no Never-takers, the maximum share of Compliers is 99.497%.

d) More complexity is introduced when we consider the full array of three treatment assignments and three forms of actual treatment. In the case of two assigned treatments and two actual treatments, we have four types of subjects (Compliers, Defiers, Never-Takers, and Always-Takers). How many types of subjects are there with three treatment assignments and three forms of actual treatment?
Answer:
There are 27 possible types. See table below.

e) How many types are there if you make the following "monotonicity" stipulations:

(i) Anyone who is fully arrested if assigned to be warned would also be fully arrested if assigned to be briefly arrested or fully arrested

(ii) Anyone who is fully arrested if assigned to be briefly arrested would also be fully arrested if assigned to be fully arrested

(iii) Anyone who is briefly arrested if assigned to be warned would also be briefly arrested if assigned to be briefly arrested

(iv) Anyone who is warned if assigned to be arrested would also be warned if assigned to be warned.

The table below shows all 27 possible combinations of potential outcomes for treatment. After accounting for all four restrictions, 10 types remain.

Table 2: Question 8 Table

	Z = Warning	Z = Brief Arrest	Z = Full Arrest	Monotonicity Constraints				
Type	D(0)	D(1)	D(2)	i	ii	iii	iv	all four
1	0	0	0					
2	0	0	1					
3	0	0	2					
4	0	1	0					
5	0	1	1					
6	0	1	2					
7	0	2	0		X			X
8	0	2	1		X			X
9	0	2	2					
10	1	0	0			X	X	X
11	1	0	1			X	X	X
12	1	0	2			X	X	X
13	1	1	0				X	X
14	1	1	1					
15	1	1	2					
16	1	2	0		X	X	X	X
17	1	2	1		X	X		X
18	1	2	2			X		X
19	2	0	0	X			X	X
20	2	0	1	X			X	X
21	2	0	2	X			X	X
22	2	1	0	X			X	X
23	2	1	1	X				X
24	2	1	2	X				X
25	2	2	0	X	X		X	X
26	2	2	1	X	X			X
27	2	2	2					
Total Types	27	27	27	19	21	21	17	10

Question 9

In their study of the effects of conscription on criminal activity in Argentina, Galiani, Rossi, and Schargrodsky use official records of draft lottery numbers, military service, and prosecutions for a cohort of men born between 1958 and 1962.[2] Draft eligibility is scored 1 if an individual had a draft lottery number that caused him to be drafted, and 0 otherwise. Draft lottery numbers were selected randomly by drawing balls from an urn. Military service is scored 1 if the individual actually served in the armed services, and 0 otherwise. Subsequent criminal activity is scored 1 if the individual had a judicial record of prosecution for a serious offense. For a sample of 5,000 observations, the authors report an $\widehat{ITT_D}$ of 0.6587 (SE = 0.0012), an $\widehat{ITT}$ of 0.0018 (SE = 0.0006), and a $\widehat{CACE}$ of 0.0026 (SE = 0.0008). The authors note that the $\widehat{CACE}$ implies a 3.75% increase in the probability of criminal prosecution with military service.

[2] Galiani, Rossi, and Schargrodsky 2010.

a) Interpret the $\widehat{ITT_D}$, $\widehat{ITT}$, $\widehat{CACE}$, and their standard errors.
Answer:
The $\widehat{ITT_D}$ refers to the difference in rates of military service between the treatment and control groups. Evidently, the treatment group was 65.87 percentage points more likely to serve in the military than the control group. The $\widehat{ITT}$ refers to the difference in prosecution rates between the assigned treatment and control groups (irrespective of whether a subject actually served). The estimate of 0.0018 implies that the treatment group was 0.18 percentage points more likely to be prosecuted than the assigned control group. The $\widehat{CACE}$ is the estimated ATE among Compliers, those who serve in the military if and only if they have a draft-eligible number. This estimate is 0.0026, which implies that Compliers become 0.26 percentage points more likely to be prosecuted as a result of serving in the military. The standard errors are a measure of statistical uncertainty, and a rule of thumb is that a 95% confidence interval may be formed by adding and subtracting +/- 2SEs. In this case, the 95% interval for $\widehat{ITT_D}$ is 65.87 +/- 0.0024; for $\widehat{ITT}$ is 0.0018 +/- 0.0012; for $\widehat{CACE}$, it is 0.0026 +/- 0.0016. The margin of uncertainty for the $\widehat{ITT}$ and $\widehat{CACE}$ is fairly wide, but the intervals are on the positive side of zero, suggesting that military service (if the exclusion restriction holds) has a criminogenic effect.

b) The authors note that 4.21% of subjects who were not draft eligible nevertheless served in the armed forces. Based on this information and the results shown above, calculate the proportion of Never-Takers, Always-Takers, and Compliers under the assumption of monotonicity.
Answer:
Monotonicity means that the proportion of Defiers is zero. The 4.21% who served without being drafted implies that Always-takers are 4.21% of the subject pool. From the $\widehat{ITT_D}$ of 0.6587 we infer the Compliers are 65.87% of the subject pool. That leaves 1 - 4.21% - 65.87% = 29.9% who are Never-takers.

c) Discuss the plausibility of the monotonicity, non-interference, and excludability assumptions in this application. If an assumption strikes you as implausible, indicate whether you think the $\widehat{CACE}$ is biased upward or downward.
Answer:
Let's analyze each assumption. Monotonicity implies no Defiers. Defiers are those who serve in the military if and only if they are not drafted. Given that one ordinarily think of people who join the military on their own volution as being willing to go if drafted, it is difficult to imagine that many people fit this description, so this assumption seems plausible. Random assignment implies that treatment assignment is independent of the potential outcomes. Although some lotteries are implemented incompetently or corruptly, we are given no reason to suspect that here. Non-interference means that potential outcomes reflect only the treatment or control status of the subject in question and do not depend on the status of other observations. In this case the potential outcome is whether a subject will be prosecuted. It seems possible that one's criminal career could be shaped by whether one's friends are or aren't drafted, but it is not clear how this violation of non-interference would bias the results, since if my friends are drafted it might make me more likely to engage in criminal conduct regardless of whether I am assigned to treatment or control. Excludability means that potential outcomes respond solely to receipt of the treatment (military service) and not the random assignment of the treatment or any indirect byproduct of random assignment (e.g., draft dodging). If citizens that are drafted are more easily monitored (e.g., their finger prints are recorded) then there might be an upward bias in the measurement of the crime committed by those assigned to treatment simply because it is easier to solve a crime committed by them.

Question 10

In her study of election monitoring in Indonesia, Hyde randomly assigned international election observers to monitor certain polling stations.[3] Here, we consider a subset of her experiment where approximately 20% of the villages were assigned to the treatment group. Because of difficult terrain and time constraints, observers monitored 68 of the 409 polling places assigned to treatment. Observers also monitored 21 of the 1,562 stations assigned to the control group. The dependent variable here is the number of ballots that were declared invalid by polling station officials.

a) Is monotonicity a plausible assumption in this application?
Answer:
Monotonicity implies no Defiers. Defiers in this context are polling stations that are monitored if and only if they are assigned to the control group. Monotonicity seems a plausible assumption if one imagines that polling stations are monitored in the control group because monitors are tourists who like to monitor spots that are close to tourist attractions. These attractions would also draw their attention if the polling stations in question were in the treatment group.

b) Under the assumption of monotonicity, what proportion of subjects (polling locations) would you estimate to be Compliers, Never-Takers, and Always-Takers?
Answer:
Monotonicity implies that Defiers make up 0% of the subject pool. Always-takers make up $\pi_{\hat{A}T} = \frac{21}{1562} = 0.013$ or 1.3%. Compliers make up $\hat{\pi_C} = \frac{68}{409} - \frac{21}{1562} = 0.153$ of 15.3%. Therefore, Never-takers make up 83.4%.

c) Explain what the non-interference assumption means in the context of this experiment.
Answer:
Non-interference means that each polling station's potential outcomes respond only to whether the polling station itself is treated. This assumption would be jeopardized if monitors, for example, displace corruption when they monitor nearby polling stations. Under that scenario, the "untreated" potential outcome may rise when neighboring stations are treated, causing bias when treated and untreated stations are compared in order to gauge the ATE of treatment vs. no treatment.

d) Download the sample dataset at http://isps.research.yale.edu/FEDAI and estimate the ITT and the CACE. Interpret the results.

```
In [1]: qui import delim ./data/chapter06/Hyde_POP_2012, clear

In [2]: rename sample Z
        rename invalidballots Y
        rename observed D

In [3]: qui regress Y Z
        scalar ITT = _b[Z]
        qui regress D Z
        scalar ITTD = _b[Z]
        scalar CACE = ITT/ITTD
```

[3] Hyde2010.

```
In [4]: disp %8.3f ITT

   4.824


In [5]: disp %18.2f CACE

    31.57
```

The ITT is estimated by comparing means in the assigned control 81.33 and treatment groups 86.16, for a difference of 4.82: assignment to monitoring appears to increase the number of invalid ballots by 4.82 per polling station. The CACE is estimated by dividing the ITT by the ITTd, calculated above: 31.57. Assuming non-interfernce, excludability, and monotonicity, this estimate implies that an actual visit by observers causes an increase of 31.57 invalid ballots among Compliers (polling stations that are observed if and only if assigned to treatment).

e) Use randomization inference to test the sharp null hypothesis that there is no intent-to-treat effect for any polling location. Interpret the results. Explain why testing the null hypothesis that the ITT is zero for all subjects serves the same purpose as testing the null hypothesis that the ATE is zero for all Compliers.

```
In [6]: ritest Z _b[Z], reps(10000) nodots: ///
        regress Y Z

  res. var(s):  Z
   Resampling:  Permuting Z
Clust. var(s):  __000006
     Clusters:  1971
Strata var(s):  none
       Strata:  1

------------------------------------------------------------------------------
T            |     T(obs)       c       n   p=c/n   SE(p) [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _pm_1 |   4.824097    4858   10000  0.4858  0.0050  .4759595   .4956488
------------------------------------------------------------------------------
Note: Confidence interval is with respect to p=c/n.
Note: c = #{|T| >= |T(obs)|}


In [7]: di %8.4f el(r(p), 1,1)

  0.4858
```

Testing the null hypothesis of zero ITT for all subjects, we generated 10,000 randomizations and compared the observed ITT to the sampling distribution of simulated ITTs. We obtained a two-tailed p-value of 0.5 because the observed value was larger than approximately 50% of the simulated

ITTs. We therefore cannot reject the sharp null hypothesis of no effect for any unit. Testing the null that the ITT is zero for all subjects is the same as testing the null that the CACE is zero for all compliers because the ITT is in the numerator of the CACE.

Question 11

A large-scale experiment conducted between 2002 and 2005 assessed the effects of Head Start, a preschool enrichment program designed to improve school readiness.[4] The assigned treatment encouraged a nationally representative sample of eligible (low-income) parents to enroll their four-year-olds in Head Start. Of the 1,253 children assigned to the Head Start treatment, 79.8% actually enrolled in Head Start; 855 of the children assigned to the control group (13.9%) nevertheless enrolled in Head Start. One of the outcomes of interest is pre-academic skills, as manifest at the end of the yearlong intervention. The principal investigators report that scores averaged 365.0 among students assigned to the treatment group and 360.5 among students assigned to the control group, with a two-tailed p-value of .041. Two years later, students completed first grade. Their first grade scores on a test of academic skills averaged 447.7 in the treatment group and 449.0 in the control group, with a two-tailed p-value of 0.380.

a) Estimate the CACE for this experiment, using pre-academic skills scores as the outcome.
Answer:
The estimated CACE is: $\widehat{CACE} = \frac{365-360.5}{0.798-0.139} = 6.82$

b) Estimate the CACE for this experiment, using academic skills in first grade as an outcome.
Answer:
The estimated CACE is: $\widehat{CACE} = \frac{447.7-449.0}{0.798-0.139} = -1.97$

c) Estimate the average downstream effect of pre-academic skills on first grade academic skills. Hint: Divide the estimated ITT (from a regression of first grade academic skills on assigned treatment) by the estimated ITT_D (from a regression of pre-academic skills on assigned treatment). Interpret your results. Are the assumptions required to identify this downstream effect plausible in this application? If not, would you expect the apparent downstream effect to be overestimated or underestimated?
Answer:
The estimated downstream CACE is: $\widehat{CACE} = \frac{447.7-449.0}{365-360.5} = -0.29$
The results suggest, surprisingly, that an improvement in pre-academic skills among Compliers (those whose pre-academic skills change if they are exposed to the treatment) led to a deterioration of academic skills in first grade. For every one-point gain in pre-academic skills, there was a 0.29 drop in first grade skills. Ordinarily, one would expect a positive relationship (building early skills help build skills later on). One possible explanation for this anomalous result is sampling variability. Another is a violation of the exclusion restriction. Suppose, for the sake of argument, that Head Start teachers were coaching students to help them perform better on tests of pre-academic skills. (One could define this sort of teaching-to-the-test as the effect of Head Start, in which case there would be no excludability violation.) Suppose that coaching boosts pre-academic skills scores but lowers first grade scores because the same tricks that are used on the pre-academic skills test lower grades on the first grade test. The excluded factor of coaching boosts the denominator and lowers the numerator, and so the net bias is difficult to predict.

[4] Puma et al. 2010. We focus here on one part of the study, the sample of four-year-old subjects.