Class 2 — iTV Experiments
The ATE

Jake Bowers

0. Get ready to work:

   Today we will be trying to replicate the results of a small field experiment run by my friend Costas Panagopoulos. He ran a field experiment run just prior to the Mayoral elections of 2005. From the experimental pool of 8 cities, he matched them (by hand, based roughly on previous turnout) into 4 pairs. Within each pair 1 city was randomly assigned to receive newspaper advertisements urging citizens to vote and 1 received no intervention.

   Load the data:

```
load(url("http://jakebowers.org/PS590/news.wide.rda"))
```

   If you run the following code you can see the pairs ("s" is the stratum indicator, "z" is the treatment assignment indicator, and "r" is the outcome.)

```
news.exp<-news.wide[1:8,c("s","z","r","pop2000","rpre")]
news.exp
```

1. Now, let us turn to Neyman's solution to the fundamental problem of causal inference. Here is a table showing the treatment assignment status, outcomes, and potential outcomes for this experiment.

i	Z_i	y_i	y_{i1}	y_{i0}
Saginaw	0	16	?	16
Sioux City	1	22	22	?
Battle Creek	0	14	?	14
Midland	1	7	7	?
Oxford	0	23	?	23
Lowell	1	27	27	?
Yakima	0	58	?	58
Richland	1	61	61	?

Table 1: Treatment assignment (Z), observed turnout (y), and unobserved potential outcomes (y_1, y_0) for Cities (i) in the Newspapers Experiment.

   Please write down an expression for defining a simple additive, linear, individual level causal effect (we could call this effect τ_i).

$$\tau_i = y_{i,1} - y_{i,0}$$

2. Now, we cannot actually calculate τ_i. How would Fisher have suggested we try to use what we observe (say, in that table above), to learn about τ_i?

   He would have suggested that we hypothesize about τ_i, in particular, $H_0 : \tau_i = 0$ or $H_0 : y_{i1} = y_{i0}$. As you showed last time, we can link a hypothesis about τ_i to what we do observe via the mechanism of randomization; and randomization allows us to characterize the distribution of test statistics summarizing H_0, and thus we can get a p-value.

3. There is another approach, most commonly associated with Neyman, although it is so wide-spread that many have stopped citing Neyman's 1923 article (which was only translated into English in 1990 anyway). How would Neyman fill in those "?" in the table in order to reason about τ_i? *Hint:* This is a trick question.

   He wouldn't. Rather he would take averages over the entire column. Even if we cannot observe the individual "?" we can know a lot about how the mean of a random sample relates to an unobserved population. We don't have to observe the whole population to draw inferences from the sample to the population. So, Neyman doesn't require us to fill in the individual "?".

4. So, Neyman changed the rules or asked different question.[1] Neyman decided that he would be interested in differences of means rather than in what actually happens to individual units. It turns out that this decision was very smart and convenient from a mathematical perspective. So, his question was about $\bar{\tau} = (1/n)(\sum_{i=1}^{n} y_{1i}) - (1/n)(\sum_{i=1}^{n} y_{0i}) = \bar{y}_1 - \bar{y}_0$ rather than about comparisons of individual units' potential outcomes. That is, Neyman wanted to ask questions about differences of means, and so, since a hypothesis is a kind of question, he would write $H_0 : \bar{\tau} = 0$. Now, one way to get $\bar{\tau} = 0$ is for $y_{1i} = y_{0i}$ but there are many other

[1] For some of you, it will be useful to think about what Neyman did as akin to Kirk's third response to the Kobayashi Maru test.

ways for this to occur where $y_{1i} \neq y_{0i}$ thus, we call Neyman's hypothesis the "weak null hypothesis" in that it says less about the relationships between the potential outcomes than Fisher's "strong null hypothesis" does.

Now, why would a "weak null" also solve the causal inference problem? We still cannot observe y_{1i} for everyone and so cannot calculate $\bar{y}_1$.

We'll see this more vividly when we show that the sample ATE is an unbiased estimator of the difference in means of the potential outcomes.

For now we can just say that our best estimate of average turnout in the population (i.e. the experimental pool, the "finite population") under treatment is the mean turnout among those assigned treatment.

```
with(news.exp,mean(r[z==1]))
```

```
[1] 29.2
```

5. What would be our best estimate of mean turnout in the control group in the population? *Hint:* Remember that by "population" Neyman meant the 8 cities here.

```
with(news.exp,mean(r[z==0]))
```

```
[1] 27.8
```

6. So, what we have said (without proof so far but just using our intuitions about random sampling) is we can use sample means to represent population means. And it would seem to follow that $\hat{\bar{\tau}} = (1/m)\sum_{i=1}^{n} Z_i Y_i - (1/(n-m))\sum_{i=1}^{n}(1-Z_i)Y_i$ is a good estimator of $\bar{\tau}$ (where m is number of treated observations and n is total number of observations). But what do we mean by a good estimator?

   A good estimator is one that approaches the right answer more and more closely as we add more information (i.e. is consistent). And, a good estimator is right on average (i.e. is unbiased) — on average across repeated realizations of the data generation process (here across repeated assignments of treatment to these same cities in these same pairs).

7. So, you might have said, "consistent", but we'll focus on "unbiased" today (especially important for our little experiment where consistency of estimtors is probably not a useful property!) Show that $E_y[\hat{\bar{\tau}}] = \bar{\tau}$ where E_y is "across all ways to draw assignments, Z, from the urn." *Hint:* (1) Ignore the pairing for today. We are just sampling 4 units for treatment from the population of 8 units: this is a simply randomized experiment; (2) recall that only Z is random and that $E_R[Z_i] = m/n$ because no pairs for now; (3) Recall that $y_i = Z_i y_{1i} + (1-Z_i)y_{0i}$ or $y_i = y_{0i}$ when $Z_i = 0$ so we can write y_{1i} for y_i when we see $Z_i Y_i$ and y_{0i} for y_i when we see $(1-Z_i)Y_i$. (4) We've defined $\bar{\tau}$ and $\hat{\bar{\tau}}$ in earlier questions.

   So, earlier we linked what we could observe y_i with what we had social science theory (or at least substantive concerns) about, y_{1i}, y_{0i}. Using the observational identity and hypotheses.

   This little derivation allows you to see how Neyman made the link between the observed outcomes and the potential outcomes.

$$E_y[\hat{\tau}] = E_y\left[\sum_{i=1}^{n} Z_i \frac{Y_i}{m} - \sum_{i=1}^{n}(1-Z_i)\frac{Y_i}{n-m}\right] \tag{1}$$

recall that $y_i = Z_i y_{1i} + (1-Z_i)y_{0i}$ or $y_i = y_{0i}$ when $Z_i = 0$, etc.

$$= E\left[\sum_{i=1}^{n} Z_i \frac{y_{1i}}{m}\right] - E\left[\sum_{i=1}^{n}(1-Z_i)\frac{y_{0i}}{n-m}\right] \tag{2}$$

recall that only Z is random and that $E_R[Z_i] = m/n$ because no pairs for now

$$= \sum_{i=1}^{n} \frac{m}{n}\frac{y_{1i}}{m} - \sum_{i=1}^{n}\left(1-\frac{m}{n}\right)\frac{y_{0i}}{n-m} \tag{3}$$

now $(1-(m/n) = (n-m)/n$

$$= \sum_{i=1}^{n} \frac{y_{1i}}{n} - \sum_{i=1}^{n} \frac{y_{0i}}{n} \tag{4}$$

$$= \bar{y}_{1i} - \bar{y}_{0i} = \bar{\tau} \tag{5}$$

8. Now, it was fun to do a little algebra, but let's show that this is the case with a little simulation study. Explain what is happening here. It may help to run each line and print the results.

```
##First produce the set of possible random assignments within pairs
Omega1 <- combn(8,4,FUN=function(x){ tmp<-rep(0,8); tmp[x]<-1; return(tmp)})
dim(Omega1)
choose(8,4)
##Next set up some fake potential and observed outcomes. We are acting as if we could actually observe both
##potential outcomes.
set.seed(20121102)
news.exp$faker0<-rnorm(8,mean=mean(news.exp$r[news.exp$z==0]),sd=sd(news.exp$r[news.exp$z==0]))
news.exp$faker1<-rnorm(8,mean=mean(news.exp$faker0)+10,sd=sd(news.exp$r[news.exp$z==1]))
news.exp$obsr<-with(news.exp, z*faker1+(1-z)*faker0 )
##Now calculate the true ATE and the estimated ATE from the sample
(trueATE<-with(news.exp,mean(faker1)-mean(faker0)) )
(estATE<-with(news.exp,mean(obsr[z==1])-mean(obsr[z==0])))
##Now define a function which reveals a difference observed outcome and calculates
##a different mean difference given a different treatment vector
make.new.R.and.mean.diff<-function(thez){
  newobsr<-with(news.exp, thez*faker1+(1-thez)*faker0 )
  return(mean(newobsr[thez==1])-mean(newobsr[thez==0]))
}
##For every way possible to run the experiment, calculate this mean difference
dist.sample.mean.diff<-apply(Omega1,2,function(thez){
  make.new.R.and.mean.diff(thez)
})
##Calculate the average of the randomization distribution of the mean difference (i.e. E(\hat{\bar{\tau}}))
(E.estATE<-mean(dist.sample.mean.diff))
##Show that this average is equal to the truth.
all.equal(E.estATE,trueATE)
```

9. Now let us get a little more clear on what the weak null means. So, in the last question the true $\bar{\tau}$ was about 3.85. Does this mean that $y_{1i} > y_{0i}$ for all i? Or that $y_{1i} = y_{0i} + 3.85$? *Hint:* In this simulation we actually created y_{1i} and y_{0i} (`faker1` and `faker0`). Try subtracting them to see what the true treatment effects are for each unit.

   The weak null does not mean those things. In fact, the weak null has nothing to say about any individual city here. In fact, a positive average treatment effect (the object about which the weak null asks questions) is compatible with situations in which some cities have turnout decreased because of the treatment assignment:

```
news.exp$faker1-news.exp$faker0
```

```
[1]  -1.73 -10.20  29.75 -20.32  10.24  21.80 -21.09  22.39
```

10. What if we hypothesized, in Fisher's framework, that the treatment effect was 2 percentage points of turnout for each and every unit. What does an average effect of 2 points mean? List two different sets of unit-level observed outcomes which would produce an average difference of means of 2 points from this experiment. *Hint:* I did something like this to simplify, but you can feel free to setup this problem however you'd like.

```
fakeR1<-c(2,3,3,3,6,7,7,7)
fakeZ<-rep(c(1,0),each=4)
diff(tapply(fakeR1,fakeZ,mean))
```

    We could get an average of 2 in many many many ways here. One way would be this pattern of individual level effects:

```
##One positive effect and 3 negative effects
fakeR1<-c(33,5,5,5,10,10,10,10)
fakeZ<-rep(c(1,0),each=4)
diff(tapply(fakeR1,fakeZ,mean))
```

```
1
2
```

```
##A constant, additive effect
fakeR1<-c(12,4,4,4,10,2,2,2)
diff(tapply(fakeR1,fakeZ,mean))
```

```
1
2
```

11. So, notice another difference between Neyman and Fisher here. Neyman's emphasis was on *estimating* an average causal effect (so, unbiasedness of an estimator is important). Fisher's emphasis was on *testing* a hypothesis about a unit specific causal effect. Now, for paired or stratified data, the ATE is a little different (but still pretty simple): we just calculate the ATE within stratum and then take the weighted average across the strata (often we weight by the the number of treated units in the stratum or perhaps just the total size of the stratum or trying to take into account the ratio of treated to controls in the stratum).

    For example, we might write:

$$\hat{\hat{\tau}}_b = (1/m_b)\sum_{i=1}^{n_b} Z_i Y_i - (1/(n_b - m_b))\sum_{i=1}^{n_b}(1 - Z_i)Y_i$$

and

$$\hat{\hat{\tau}} = (1/B)\sum_{b=i}^{B} \hat{\hat{\tau}}_b$$

The idea of weighting each block/strata equally is appropriate for our current paired situation and is simple to look at. Although there is reason to weight them unequally when we have unequal numbers of observations in the strata and when the ratios of treated to control units varies by stratum.

Let's estimate the average treatment effect taking the pairing into account. Also show that we can get the same estimate from a least squares model with indicator variables for the pairs. *Hint:* Here is a start. Make sure you understand the code.

```
thepairs.list<-split(news.exp[,c("z","r")],news.exp$s)
taub<-sapply(thepairs.list,function(dat){ mean(dat$r[dat$z==1]) - mean(dat$r[dat$z==0]) })
news.exp$sF<-factor(news.exp$s) ##useful for lm

lm1<-lm(r~z+sF,data=news.exp)
```

12. Now we have a linear model object (that you produced in the last question) and so it is hard to resist statistical inference using what you know about linear models.

```
est.sigma<-sum(resid(lm1)^2)/lm1$df
xtxinv<-  solve(t(model.matrix(lm1)) %*% model.matrix(lm1)) ##model.matrix(lm1) just gives me an X matrix easily
varblm1<-est.sigma * xtxinv
seblm1<-sqrt(diag(varblm1))
rbind(b=coef(lm1),seb=seblm1) ##compare to summary(lm1)$coef
coef(lm1)["z"]+c(-1,1)*qt(p=.975,df=lm1$df)*seblm1["z"] ##cf confint(lm1,parm="z")
```

Ack! I just accidentally calculated a confidence interval and displayed a standard error using the default OLS variance formula $\text{Var}\hat{\boldsymbol{\beta}} = \sigma(\mathbf{X}^T\mathbf{X})^{-1}$. What do we have to believe to feel like they reflect something useful about the ways that the ATE might vary from randomization to randomization?

Here we are back to your first year of stats: we have to believe that the observations are iid and we need a large sample story for the Normal distribution with unknown variance (which we estimate and thus arrive at a t-distribution). We have to also believe that the effect is the same across the pairs.

A small dilemma here is that we probably created heteroskedasticity via our treatment (i.e. if the treatment mattered, we imagine that it would not necessarily only matter for the mean but also for the variance of the outcome). A larger dilemma is that the value of the treatment vector are not really independent of each other: we require one and only one treatment within a pair, once one unit is treated, the other must be control. That is, most experiments do not use simple coin flips to assign treatment. The Freedman readings address these points at length — suggesting that while our little regression without pairs might be ok, once we have pair indicators we might be in trouble if we use the OLS assumptions for statistical inference about randomized data.

13. Right, we have the whole set of OLS assumptions to think about here. So, we might not want to just use plain old OLS for statistical inference on randomized experiments that are this small. Specifically, using the same kind of tricks by which we showed that the sample difference of means was an unbiased estimator of the average causal effect, we can show the following:

$$\begin{aligned}\text{Var}(d_{\text{pairs}}) &= \left(\frac{1}{B}\right)^2 \sum_{b=1}^{B} \frac{n_b}{m_b(n_b - m_b)} \sum_{i=1}^{n_b} \frac{(Y_{ib} - \bar{y}_b)^2}{n_b - 1} \\ &= \left(\frac{2}{B^2}\right) \sum_{b=1}^{B} \sum_{i=1}^{2} (Y_{ib} - \bar{y}_b)^2. \end{aligned} \tag{6}$$

If we calculate this we get a different standard error than the one arising from least squares.

```
## see https://github.com/markmfredrickson/RItools/tree/randomization-distribution
## for the next few lines (commented out because only need to run once)
##install.packages("devtools")
## library("devtools")
## .libPaths("~/R/library/")
## install_github("RItools", user = "markmfredrickson", ref = "randomization-distribution",
##                args="--no-multiarch")
library("RItools", lib.loc = "~/R/library")
se.ate.paired<-sqrt((2/4^2)*sum(sapply(thepairs.list,function(dat){ with(dat, sum((r-mean(r))^2) )})))
##compare to what is reported by xBalance as the sd of the null rand dist:
xb1<-xBalance(z~r,strata=list(s=~s),data=news.exp,report="all")
xb1$results[,c("adj.diff","adj.diff.null.sd"),]
```

So, what can we do? It would be nice to use our friend least squares to estimate average treatment effects and also for statistical inference about these estimates even in smallish size samples. What ideas might you have to use least squares to estimate the ATE but make some modification to improve our standard error estimation?

So, "robust" standard errors come immediately to mind. Or just using the variance calculation derived by Neyman/etc.. Bootstrapping would not be a good idea since (1) we do not know how these cities were sampled from larger population and (2) even if we did, sampling with replacement with our small dataset would seriously inflate the standard errors because of multicollinearity. If we did know how the outcome was generated (i.e. Poisson, Gamma Mixture of Poissons [i.e. Negative Binomial], Gaussian, etc..) then we could use maximum likelihood here (and there are some interesting approaches to small sample statistical inference with MLE using higher order asymptotics which could help). Yet, the iid assumption which most MLE derivations require (i.e. represent the likelihood function as a product and thus to get a log likelihood, etc..) again doesn't fly, and the uncertainty here is mainly in the treatment assignment (i.e. uncertainty in exactly which city would get treatment) rather than uncertainty in the distribution of the outcome.

14. Here is one idea, use "robust" standard errors. (Specifically Lin (2011) recently shows that the HC2 version of robust standard errors are equivalent to the ones that Neyman proposed in 1923).

$$HC2 = (\mathbf{X}^T\mathbf{X})^{-1}\text{diag}\left[\frac{e_i^2}{1-h_{ii}^2}\right]\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}$$

where $h_{ii} = x_i(\mathbf{X}^T\mathbf{X})^{-1}x_i$, h is often called the "hat matrix" it captures the distance an individual unit is from all of the others (the centroid of all of the X's). That is, it allows for some sense that particularly distant points might overly influence the SE calculation.

```
library(car) ## we'll ask Fox's code to do this for us.
lm1HC0<-hccm(lm1,type="hc0")
lm1HC1<-hccm(lm1,type="hc1")
lm1HC2<-hccm(lm1,type="hc2")
lm1HC3<-hccm(lm1,type="hc3")
lm1HC4<-hccm(lm1,type="hc4")
##summary(lm1)$coef

round(cbind(pairedse=se.ate.paired,
            ols.iid=seblm1["z"], ##sqrt(diag(vcov(lm1))),
            hc0=sqrt(diag(lm1HC0))["z"],
            hc1=sqrt(diag(lm1HC1))["z"],
            hc2=sqrt(diag(lm1HC2))["z"],
            hc3=sqrt(diag(lm1HC3))["z"],
            hc4=sqrt(diag(lm1HC4))["z"]),4) #$
```

Now we have lots of possibilities here and each might lead to a different confidence interval. How should we adjudicate among them? Here are two ideas:

(1) Use simulation: What is the standard deviation of mean differences across randomizations really? [this is what the standard error means]. Explain the following code (perhaps you can rewrite the sd() part more cleanly.

```
## these next two lines create Omega using as yet undocumented functions in the
## development branch of RItools
Omega.maker <- simpleRandomSampler(z=news.exp$z, b=news.exp$s)
Omega<-Omega.maker(100)$samples ## notice that it tops out at 16 because that is the max
## What is happening here?
sd(apply(Omega,2,function(thez){
  coef(lm(r~thez+sF,data=news.exp))["thez"]
}))
```

15. Another idea (2) is to assess the operating characteristics of statistical inferences made with the different procedures. Specifically, assess the Type I error rate or "coverage". First, explain what the Type I error rate is and why we'd care about it.

The Type I error rate is the "size" of the test — it is the actual proportion of rejections of a true null made by a given test. I highly encourage you to read those parts of Rosenbaum's glossary.

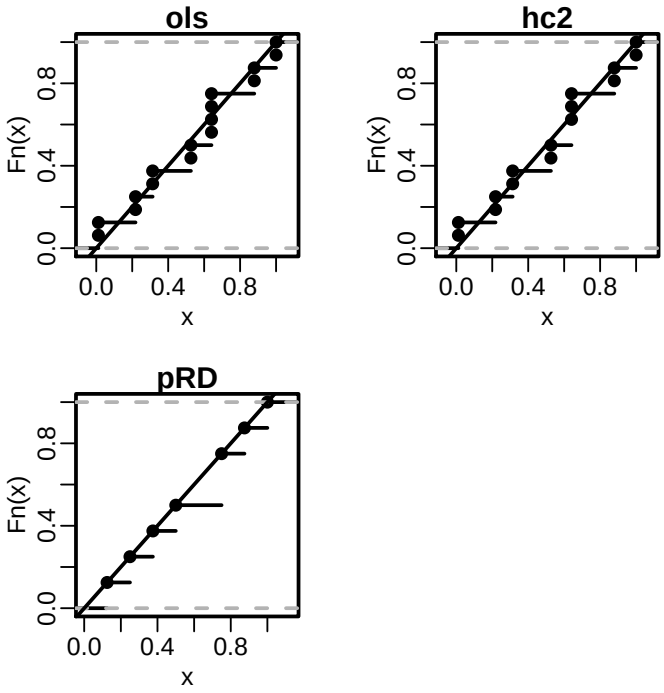
16. Here is the code which produces a two-sided p-value for the test of the weak null (and the strict null is here in comparison) for every way that we could assign treatment here. That is, we take each of the 16 ways to assign treatment given the paired design temporarily as the "observed" assignment and compare it to the other 16 to get a p-value. By doing this we break the relationship between z and y so we should mostly not reject the null. In fact, if these tests "keep their promises" (in Rosenbaum's language), what should we expect from these lists of p-values?

```
simp.lm.p.value<-function(lmobj,se,varname="z"){
  2*(min(pt(coef(lmobj)[varname]/se,df=lmobj$df),1-pt(coef(lmobj)[varname]/se,df=lmobj$df) ))
}
##test it: simp.lm.p.value(lm1,seblm1["z"],"z") versus summary(lm1)$coef["z",]
library(parallel) ## for RItest
get.many.ps<-function(az){
  thelm<-lm(r~az+sF,data=news.exp)
  olsse<-sqrt(diag(vcov(thelm)))["az"]
  hc2se<-sqrt(diag(hccm(thelm,type="hc2")))["az"]

  ols.p<-simp.lm.p.value(thelm,olsse,"az")
  hc2.p<-simp.lm.p.value(thelm,hc2se,"az")

  thexb<-xBalance(az~r,strata=list(s=~s),data=data.frame(az=az,news.exp),report="p.values")
  thexb.p<-thexb$yesults[,"p",] ##uses Normal rather than t because we know the population (from Neyman)
  thepRD<-RItest(y=news.exp$r,
                 z=az,
                 test.stat=mean.difference,
                 sampler=simpleRandomSampler(z=news.exp$z,b=news.exp$s))

  thepRD.p<-thepRD[,"p.value"]

  return(c(ols=ols.p,hc2=hc2.p,xb=thexb.p,pRD=thepRD.p))
}
many.ps<-apply(Omega,2,function(pretend.z){
  get.many.ps(pretend.z)
})
```

We should see 5% of the tests have p-values less than or equal to .05, etc... (same for every α). Or, the p-values should be uniformly distributed.

17. Here are a couple of ways to look at the results of our simulation:

```
par(mfrow=c(2,2))
for(i in row.names(many.ps)){
  plot(ecdf(many.ps[i,]),main=i)
  abline(0,1)
}
```




```
alphas <- seq(.0625,.25,.025)
theecdfs<-apply(many.ps,1,ecdf)
cbind(alphas,sapply(theecdfs,function(thefn){ thefn(v=alphas) }))
```

```
     alphas   ols   hc2   pRD
[1,] 0.0625 0.125 0.125 0.000
[2,] 0.0875 0.125 0.125 0.000
[3,] 0.1125 0.125 0.125 0.000
[4,] 0.1375 0.125 0.125 0.125
[5,] 0.1625 0.125 0.125 0.125
[6,] 0.1875 0.125 0.125 0.125
[7,] 0.2125 0.125 0.125 0.125
[8,] 0.2375 0.250 0.250 0.125
```

What do you think? On this particular randomized experiment, which approach to statistical is least likely to mislead us? Are there other approaches that we've missed?

So, the Fisher randomization approach clearly fulfills its promise: the p-values are uniformly distributed.

None of the other approaches is very good in this particular dataset: None are terrible, and we imagine that they all might become much better as the sample size increases. However, the combination of appeals to Normal and t-distributions do not make hc2 particularly better than ols or xb (etc.) Of course, the pRD approach is testing a different null, here, too.

18. Finally, let us come back to a big conceptual and scientific issue (using the language from the Adcock and Collier (2001) reading): What is the background concept of "causal effect" or "causal relationship" that we care about? How do either of Fisher or Neyman's systematized concepts help us? Are there are other scientifically useful ways to to think about what a causal relationship might mean? If so, how might you link those ways to observation? [No need to invent a whole new framework for statistical inference here, just to ask yourself if all science should always care about averages, or when might averages not be something we care about — and/or alternatively, when would specifying unit-level hypotheses be a un-useful (or useful).]

References

Adcock, Robert and David Collier. 2001. "Measurement Validity: A shared Standard for Qualitative and Measurement Validity: A shared Standard for Qualitative and Quantitative Research." *American Political Science Review* 95(3):529–546.

Lin, Winston. 2011. "Agnostic notes on regression adjustments to experimental data: reexamining Freedman's critique." Unpublished manuscript.