Experimental Methods: Lecture 3

Raymond Duch

May 15, 2018

Director CESS Nuffield/Santiago

Road Map to Presentation

- Power Calculations
- Cluster Random Assignment
- Covariates
- Placebo Design
- Efficient Design Strategies
- Conjoint Experiments
- Mode effects

Power Analysis

Statistical Power

- What is the power of a statistical test? H_0: null hypothesis
- Apply estimator to test some alternative H_A
- Type I error: False positive
  - If the null is true, how likely does the estimated effect (or greater) occur by chance?
  - Our tolerance for these errors is set by α
  - When $\alpha = 0.05$, 95% of the CIs we construct from repeated sampling will contain the true parameter

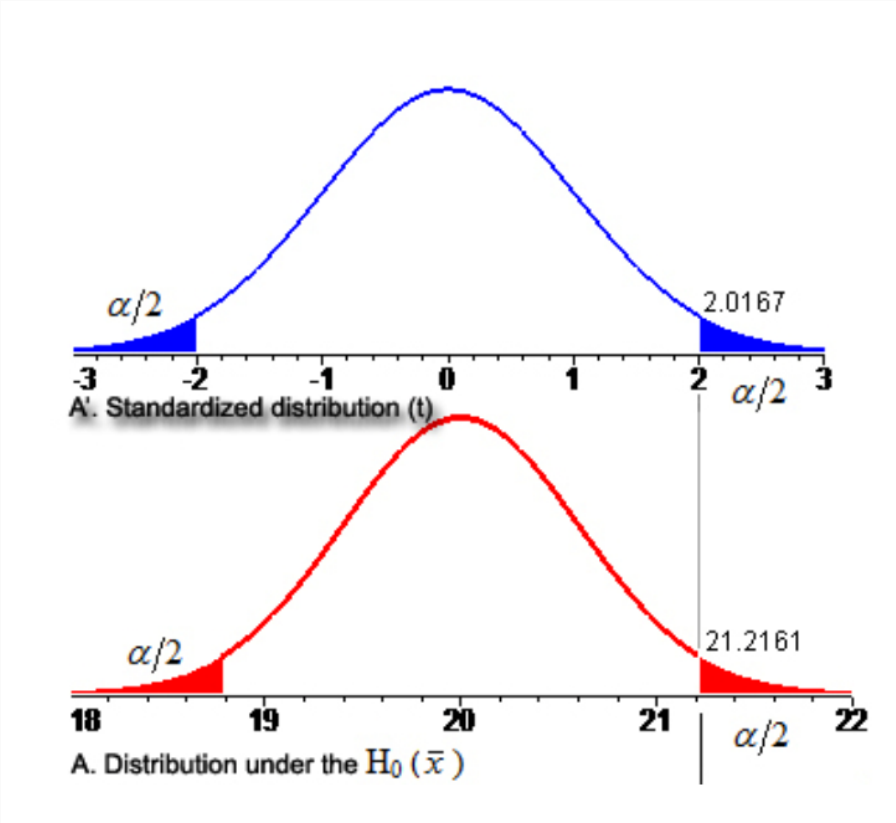
Statistical Power

- Type II error: False negative
  - If the null is not true, how often can we reject the null successfully?
  - Probability or rate of Type II error, β
- Power of a test: probability that the test rejects H_0, $1 - \beta$

Basic Inference Revisited

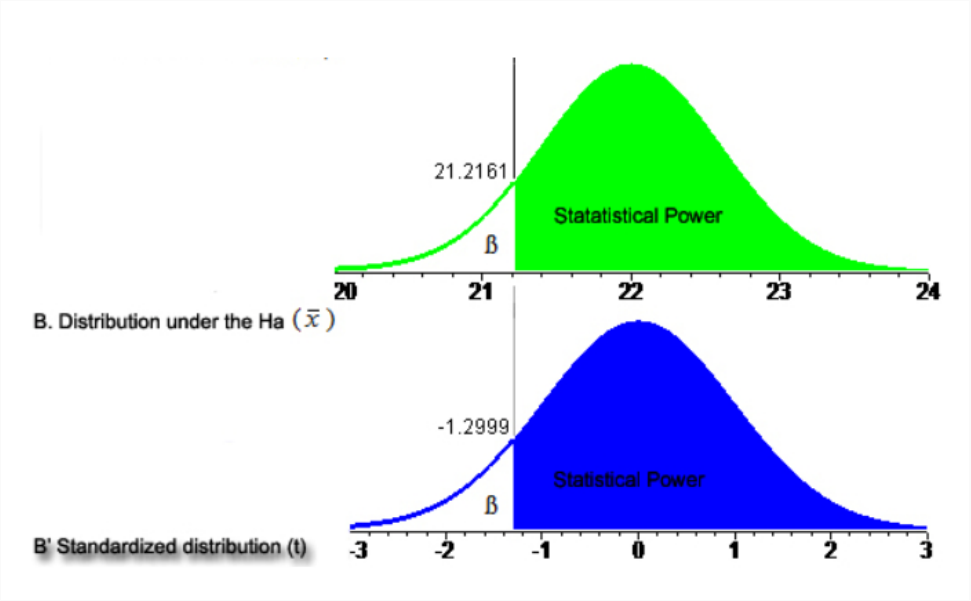
- What is the effect of losing Medicaid on infant mortality?
- $H_0 = 20$ deaths per 1,000 live births (assumed known without uncertainty here)
- True effect is an increase of 2 deaths per 1,000 live births
- Standard deviation in population is 4, we have N=44 observations; sampling distribution yields a standard error of 0.60
- $\hat{x}$ is our estimate of the new infant mortality rate
- Let's say we get an estimate right at the true estimate, $\hat{x} = 22$
- How unlikely is it we get this estimate, if the null is actually true?

Sampling Distribution Under Null



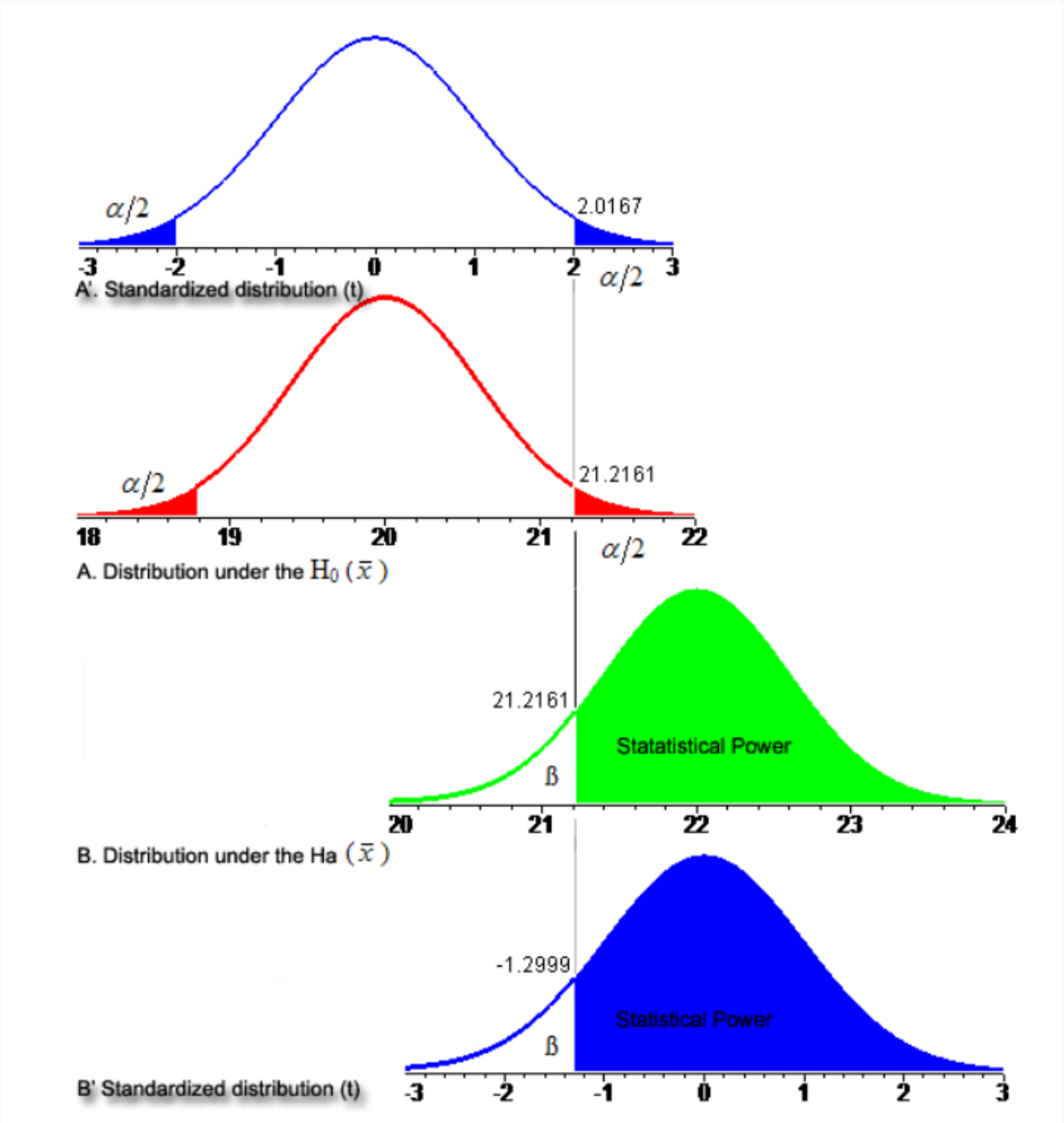
- Say for our test $\alpha = 0.05$
- Can rescale via Z-transformation
- What does this graphic mean?
- For $\hat{x} = 22$,
- t-stat=3.32, $p < 0.01$

Sampling Distribution of $\hat{x}$



- Interpret this graphic
- $1 - \beta$ is fraction of estimates that reject null hypothesis
- Power of the test
- What $x_t rue$ yields $1 - \beta = 0.5$?
- What parameters are needed?

The Relationship Between α and β



Sample Size Increases Power

- Of primary interest because it can be manipulated
- Law of large numbers: for independent data, statistical precision of estimates increases with the square root of the sample size, $\sqrt{n}$
- Test statistics often have the form $T = \hat{\theta}/\sqrt{\hat{V}(\hat{\theta})}$
- Example: Mean of normal distribution θ, data $y = (y_1, ..., y_n)$, iid

$$\hat{\theta} = n^{-1} \sum_{i=1}^{n} y_i = \bar{y}$$

$$\hat{V}(\hat{\theta}) = V(y)/n \text{ and } \sqrt{\hat{V}(\hat{\theta})} = s_y/\sqrt{n}$$

$$T = \bar{y}/(s_y/\sqrt{n})$$

- This logic extends to two-sample case (e.g., treated vs control in an experiment), regression, logistic regression, etc.

Reverse Engineer T to Determine Sample Size

- How much sample do I need to give myself a "reasonable" chance of rejecting H_0, given expectations as to the magnitude of the "effect"
- Example:

  A proportion $\theta \in [0, 1]$ estimated as $\hat{\theta}$

  Variance is $\theta(1-\theta)/n$, maxes at 0.5

  A 95% CI at $\theta = 0.5$ is $0.5 \pm 2\sqrt{0.25/n}$

  Width of that interval is $W = 4\sqrt{0.25/n} \rightarrow n = 4/W^2$

- Typical use: how big must a poll be to get reasonable MOE?
- For researchers, how big must a poll be to detect a campaign effect?
  - Answer depends on beliefs about likely magnitude of campaign effects

Example 2: campaign effect

- In R, power.prop.test()
- Researcher thinks effects that move a proportion (i.e. vote support) from 50% to 52% are likely
- Would like to be able to detect effects of this size at conventional levels of statistical significance
- ($p = 0.05$; 95% confidence interval for the effect excludes zero), with power $(1 - \beta)$ equal to 0.50
- $H_0 : \delta = \theta_1 - \theta_2 = 0$; $H_A : \delta \neq 0$ (two-sided alternative)

Power Estimate for 2 Point Effect

Two-sided alternative at conventional levels of significance

```
>power.prop.test(p1 = 0.5, p2 = 0.52, power
    = 0.5)
```

Two-sample comparison of proportions power calculation

n = 4799.903

p1 = 0.5

p2 = 0.52

sig.level = 0.05

power = 0.5

alternative = two.sided

NOTE: n is number in *each* group

Power Estimate for 2 Point Effect

One-sided alternative at conventional levels of significance

```
> power.prop.test(p1 = 0.5, p2 = 0.52,
    power = 0.5,
alternative=  one.sided")
```

Two-sample comparison of proportions power calculation

n = 3380.577

p1 = 0.5

p2 = 0.52

sig.level = 0.05

power = 0.5

alternative = one.sided

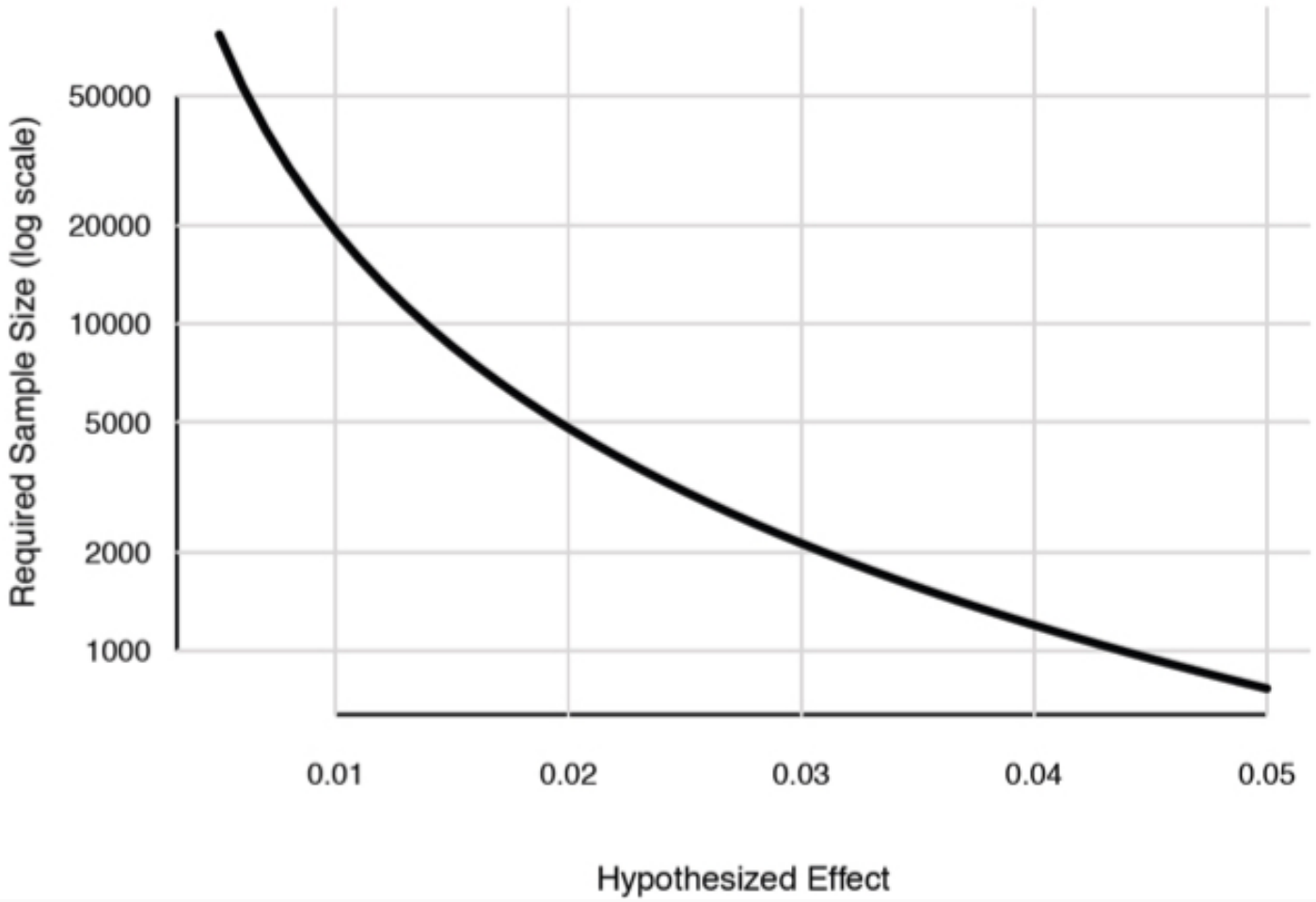
NOTE: n is number in *each* group

Power Curves

```
> p> effects <- seq(0.005, 0.05, by =
    0.001)

> base <- 0.5
> m <- length(effects)
> n <- rep(NA, m)
> for (i in 1:m) {
n[i] <- power.prop.test(p1 = base, p2 =
    base + effects[i],
+ power = 0.5)$n
+})
```

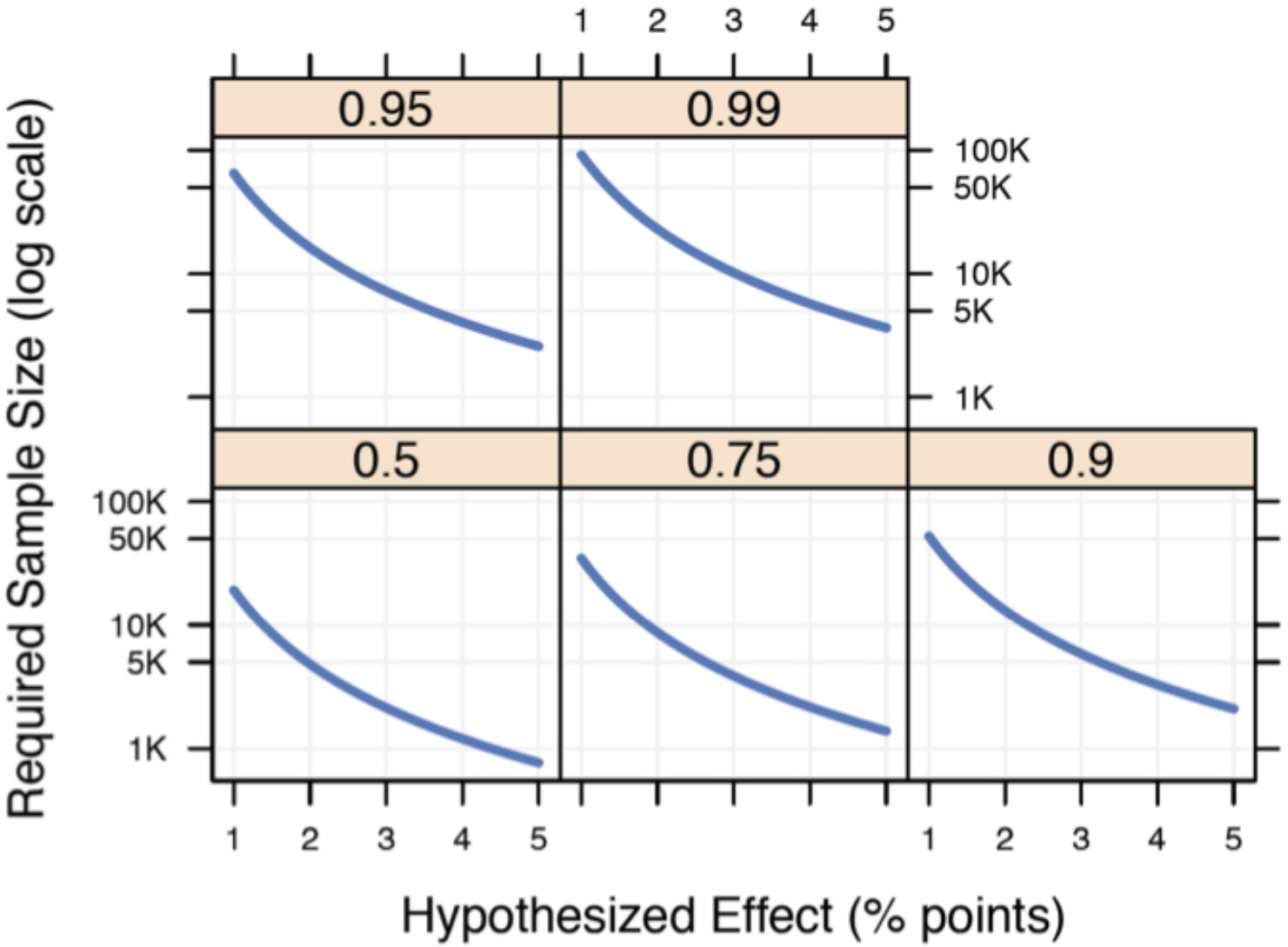
Power Curves



Looping over Power Curves

```
> power <- c(0.5, 0.75, 0.9, 0.95, 0.99)
> effects <- seq(0.01, 0.05, by = 0.001)
> base <- 0.5
> m <- c(length(power), length(effects))
> n <- matrix(NA, m[1], m[2])
> for (i in 1:(m[1])) {
+ for (j in 1:(m[2])) {
+ n[i, j] <- power.prop.test(p1 = base, p2
    = base + effects[j],
+ power = power[i])$n
+ }
+ }
```

Power Curves: different power levels



Beramendi Duch Example

Table 1: Tax Compliance Experimental Treatments and Hypothesized Treatment Outcomes

		Invest		Sacrifice	
Tax Progressivity	**Benefits Regime**	**Poor**	**Rich**	**Poor**	**Rich**
Low Tax (25%)					
High (Rich contribute all/poor zero)	Regressive (Equal Division)	30%	**100%**	30%	**70%**
High (Rich contribute all/poor zero)	Progressive (Poor all/rich zero)	50%	80%	50%	90%
Low (Proportional contribution)	Regressive (Equal Division)	60%	90%	60%	90%
Low (Proportional contribution)	Progressive (Poor all/rich zero)	80%	**50%**	80%	**100%**
High Tax (50%)					
High (Rich contribute all/poor zero)	Regressive (Equal Division)	30%	**100%**	30%	**50%**
High (Rich contribute all/poor zero)	Progressive (Poor all/rich zero)	80%	**50%**	90%	**70%**
Low (Proportional contribution)	Regressive (Equal Division)	30%	**20%**	30%	**70%**
Low (Proportional contribution)	Progressive (Poor all/rich zero)	100%	**10%**	100%	**90%**

Beramendi Duch Example

Table 2: Summary of Tax Compliance Experimental Treatments

Tax Progressivity	Benefits Regime	Chile	UK
Low Tax (25%)			
High (Rich contribute all/poor nothing)	Regressive (Equal Division)	50	50
High (Rich contribute all/poor nothing)	Progressive (All to poor/zero to rich)	50	50
Low (Proportional contribution)	Regressive (Equal Division)	50	50
Low (Proportional contribution)	Progressive (All to poor/zero to rich)	50	50
High Tax (50%)			
High (Rich contribute all/poor nothing)	Regressive (Equal Division)	50	50
High (Rich contribute all/poor nothing)	Progressive (All to poor/zero to rich)	50	50
Low (Proportional contribution)	Regressive (Equal Division)	50	50
Low (Proportional contribution)	Progressive (All to poor/zero to rich)	50	50

Beramendi Duch Example

Table 3: Summary of 80% Power Calculations for Experimental Treatments

Tax Progressivity	Benefits Regime	Effect	N
Invest Rich: Low versus High Tax			
High (Rich contribute all/poor nothing)	Regressive (Equal Division)	0.00	N/A
High (Rich contribute all/poor nothing)	Progressive (All to poor/zero to rich)	0.30	30
Low (Proportional contribution)	Regressive (Equal Division)	0.70	5
Low (Proportional contribution)	Progressive (All to poor/zero to rich)	0.40	30
Sacrifice Rich: Low versus High Tax			
High (Rich contribute all/poor nothing)	Regressive (Equal Division)	0.20	71
High (Rich contribute all/poor nothing)	Progressive (All to poor/zero to rich)	0.20	49
Low (Proportional contribution)	Regressive (Equal Division)	0.20	49
Low (Proportional contribution)	Progressive (All to poor/zero to rich)	0.10	57

Practical Advice on Power

- What is "typical" size for effects, and how might we guess?
  - Some thoughts on later example
- Generally, experiments require $1 - \beta > 0.8$ to get funding
- Zaller's maxim: "Do your power analysis, figure out your sample size, then double it"

Practical Advice on Power

- Cost considerations: Gerber and Green turnout experiment
  - One component involved canvassing
  - $40 per hour for a pair of students, 6,000 treated
  - If 6 houses an hour, need 1000 hours, so $40k right there alone
  - Implications based on power curve slide
- In particular costs high for general population experiments
- Anyone have guesses how much surveys cost?
- How much value?

Cluster Random Assignment

- Thus far we have focussed on individual-level treatments
- Individuals are often embedded in clusters where they receive either the treatment or control
    - Media market
    - Voting district or precinct
    - Classroom
- May be unavoidable or the level at which the intervention realistically takes place
- If potential outcomes differ across clusters it will lead to imprecise estimates

PO: High Sampling Variability

		Classroom-level		*Cluster-level mean*	
School	Classroom	Y_i^c	Y_i^t	Y_i^c	Y_i^t
A	A-1	0	4		
A	A-2	1	5	1	5
A	A-3	2	6		
B	B-1	2	6		
B	B-2	3	7	3	7
B	B-3	4	8		
C	C-1	3	7		
C	C-2	4	8	4	8
C	C-3	5	9		
D	D-1	7	11		
D	D-2	8	12	8	12
D	D-3	9	13		

PO: Low Sampling Variability

School	Classroom	*Classroom-level* Y_i^c	*Classroom-level* Y_i^t	*Cluster-level mean* Y_i^c	*Cluster-level mean* Y_i^t
A	A-1	0	4		
A	A-2	3	7	4	8
A	A-3	9	13		
B	B-1	2	6		
B	B-2	3	7	4	8
B	B-3	7	11		
C	C-1	1	5		
C	C-2	4	8	3.3	7.3
C	C-3	5	9		
D	D-1	4	8		
D	D-2	8	12	4.7	8.7
D	D-3	2	6		

Cluster Random Assignment

- When clusters are the same size the ATE can be estimated the usual way
  - 4 in both cases
- Different standard errors
  - High variability $= 2.9$
  - Complete randomization $= 1.6$
  - Low variability $= 0.57$
- Penalty associated with clustering depends on the variability of the cluster-level means

SE with Clustered Design

$$SE(\widehat{ATE}) = \sqrt{\frac{1}{k-1}\left\{\frac{mVar(\bar{Y}_i^c)}{N-m} + \frac{(N-m)Var(\bar{Y}_j^t)}{m} + 2Cov(Y_j^c, Y_j^t)\right\}}$$

Clusters with very similar mean Y_j^c and mean Y_j^t would lead to smaller variances and lead to a more precise estimate of ATE.

Cluster Random Assignment

- Likely stuck with highly variable clusters based on
  - Geography
  - Institutions
  - Age groups
- Improving precision
  - Increase the number of clusters
  - Increasing the number of subjects per cluster will not have much of an effect on between cluster variance
  - Include covariates
  - `Sampclus` in Stata allows you to play with the number of clusters and cluster size necessary to achieve adequate power

Covariate Adjustment?

- Random assignment ensures unbiased estimation of *ATE*
- Omitted variables is addressed by random assignment
- Including controls is not required

Covariates Useful

- Rescale dependent variable
  - Change from pre-test to post-test (diff-in-diff)
  - Potential outcomes have less variance
- Include in a regression analysis
  - Eliminate observed differences between treatment and control group
  - Reduce variability in outcomes
  - Results in more precise estimate of the treatment effect
- Check randomization process
- Construct blocks

Covariate Rescaling

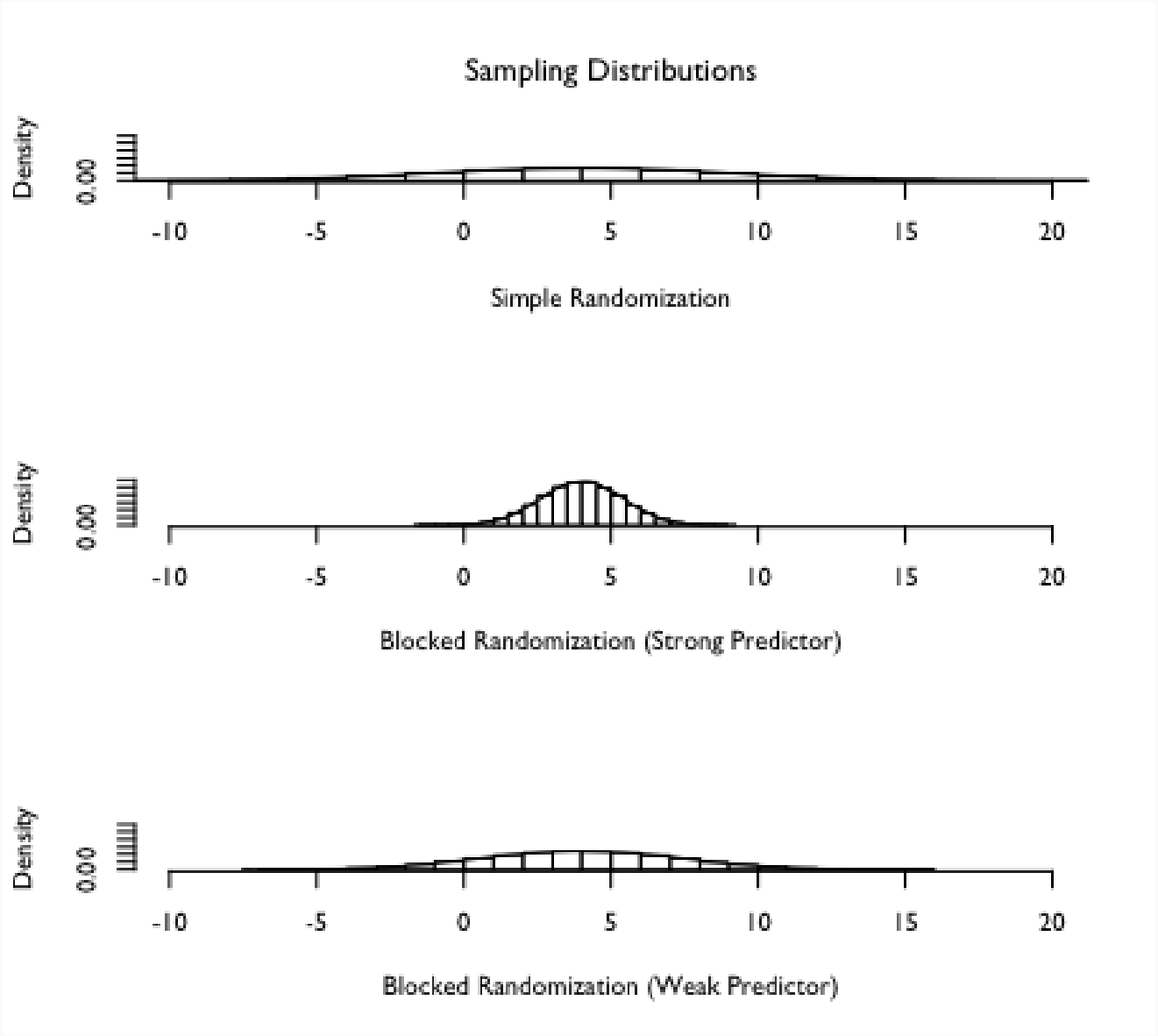
- Pre-treatment covariates
  - Fixed constants that are observed prior to random assignment
  - Unaffected by treatment assignment
- Concerns
  - Budget constraints
  - Pre-test changes the way participants respond to the treatment
  - This violates excludability assumption

Difference-in-differences

$$
\begin{aligned}
E(\widehat{ATE}) &= E[Y_i - X_i | D_i = 1] - E[Y_i - X_i | D_i = 0] \\
&= E[Y_i | D_i = 1] - E[X_i | D_i = 1] \\
&- E[Y_i | D_i = 0] + E[X_i | D_i = 0] \\
&= E[Y_i(1)] - E[Y_i(0)]
\end{aligned} \tag{1}
$$

Covariate Adjustment

Observation	$Y_i(1)$	$Y_i(0)$	D_i	X_i	x_{weak}
1	5	5	0	6	25
2	15	5	1	8	12
3	12	6	1	5	25
4	19	9	0	13	27
5	17	10	0	9	10
6	18	11	0	15	24
7	24	12	0	16	21
8	11	13	0	17	25
9	16	14	0	19	35
10	25	19	1	23	28
⋮	⋮	⋮	⋮	⋮	⋮


Sampling Distributions
Density
0.00
-10
-5
0
5
10
15
20
Simple Randomization
Density
0.00
-10
-5
0
5
10
15
20
Blocked Randomization (Strong Predictor)
Density
0.00
-10
-5
0
5
10
15
20
Blocked Randomization (Weak Predictor)

Covariates Using Regression

$$\begin{aligned} Y_i^* &= Y_i - X_i = Y_i(0)(1-d_i) + Y_i(1)d_i - X_i \\ &= a + bd_i + u_i - X_i \\ &= a + bd_i + u_i^* \end{aligned} \tag{2}$$

$$\begin{aligned} Y_i &= Y_i(0)(1-d_i) + Y_i(1)d_i \\ &= a + bd_i + cX_i + (u_i - cX_i) \end{aligned} \tag{3}$$

Advantages of Covariates

- Can include several covariates as right-hand-side variables
- Reduces disturbance variability more effectively than rescaling
- Weakly predictive covariates do not reduce sampling variability
- Which pre-treatment covariates?
  - Previous research
  - Pilot testing
  - Theoretical intuition

Caution

- Experimental outcomes should not be used to decide which covariates are included
- i.e., running several regressions and choosing one that makes *ATE* look best
- Plan in advance of the experiment
- Present difference-in-means and covariate-adjusted estimates

Covariate Imbalance

- Randomization may create imbalance in some covariates
  - Controlling for covariates reestablishes balance
- Imbalance also may alert problems with randomization
  - Retrace randomization process
  - Check with third parties executing the randomization
  - Correct possible errors

Balance Test (Panagopoulos et al. 2014)

Experimental Conditions	N	Voted (Nov 08)	Voted (Nov 06)	Voted (Nov 04)	Age (years)	Male	Partisan
Self + community high	1,000	64.4	29.2	43.1	27.5	39.9	80.4
Self + community low	1,000	66.8	30.2	44.4	28.0	38.4	83.2
Self only	1,000	64.0	26.9	40.0	28.0	40.1	78.9
Community high only	1,000	64.7	29.3	42.6	26.8	41.1	81.0
Community low only	1,000	64.7	27.8	41.3	27.0	41.9	80.3
Control	13,482	64.7	27.8	42.7	27.5	41.7	81.2
$p > F^a$		.83	.45	.43	.55	.31	.24

Figures in columns represent mean percentages unless otherwise indicated.
Test statistics generated using one way ANOVA to evaluate whether mean turnout levels differ across categories of random assignment. In all cases, we cannot reject the hypothesis of equal means at standard significance levels ($p < .05$), implying balance across groups.

Block Random Assignment

- Subjects are partitioned into blocks
- Complete random assignment within each block
- Example
  - Split the sample by gender
  - Select 5 subjects from the male group for the treatment
  - Select 5 subjects from the female group for the treatment

Block Randomization

- Previous discussion demonstrated that blocking can improve precision
- It is possible to block on several variables simultaneously
- Biggest improvements in precision come from variables that strongly predict the outcome
    - Hard to know before experiment has been run
    - Look to previous studies for clues
- Blocking and covariate adjustment yield similar results with large sample sizes
    - Treatment and control groups of more than 100 subjects

Bertrand and Mullanathan (2004)

Panel A: Subjective Measure of Quality
(Percent Callback)

	Low	High	Ratio	Difference (p-value)
White names	8.50 (1,212)	10.79 (1,223)	1.27	2.29 (0.0557)
African-American names	6.19 (1,212)	6.70 (1,223)	1.08	0.51 (0.6084)

Panel B: Predicted Measure of Quality
(Percent Callback)

	Low	High	Ratio	Difference (p-value)
White names	7.18 (822)	13.60 (816)	1.89	6.42 (0.0000)
African-American names	5.37 (819)	8.60 (814)	1.60	3.23 (0.0104)

Why Block Random Assignment: Practical Concerns

- Program requirements may restrict number of subjects allowed to receive treatment
- E.g. summer reading program concerned about students with low levels of preparedness: 60% of the admitted students must pass basic skills test
- If 50 students are admitted, randomly select 20 from the applicants that failed and 30 from those who passed
- Fairness concerns require each treatment of demographic groups
- Resource constraints mean you are only able to sample a certain number of subjects from certain groups

Why Block Random Assignment: Statistical Concerns

- Reduces sampling variability
- Subjects in blocks likely to have similar potential outcomes (those who fail and those who pass)
- Especially effective in small samples
- Ensures the ability to do subgroup analysis, e.g. women and men
- Complete random assignment may lead to imbalance

Potential Outcomes

Village	Block	$Y_i(0)$	$Y_i(1)$
1	A	0	0
2	A	1	0
3	A	2	1
4	A	4	2
5	A	4	0
6	A	6	0
7	A	6	2
8	A	9	3
9	B	14	12
10	B	15	9
11	B	16	8
12	B	16	15
13	B	17	5
14	B	18	17
⋮	⋮	⋮	⋮

Schedule of potential outcomes for public works projects when audited ($Y(1)$) and not audited ($Y(0)$)

Village	Block	All subjects $Y(0)$	All subjects $Y(1)$	Block A subjects $Y(0)$	Block A subjects $Y(1)$	Block B subjects $Y(0)$	Block B subjects $Y(1)$
1	A	0	0	0	0		
2	A	1	0	1	0		
3	A	2	1	2	1		
4	A	4	2	4	2		
5	A	4	0	4	0		
6	A	6	0	6	0		
7	A	6	2	6	2		
8	A	9	3	9	3		
9	B	14	12			14	12
10	B	15	9			15	9
11	B	16	8			16	8
12	B	16	15			16	15
13	B	17	5			17	5
14	B	18	17			18	17
	Mean	9.14	5.29	4.00	1.00	16.0	11.0
	Variance	40.41	32.49	7.75	1.25	1.67	17.0
	$Cov(Y(0), Y(1))$	31.03		2.13		1.00	

	All subjects		Block A		Block B	
	Y_i^c	Y_i^t	Y_i^c	Y_i^t	Y_i^c	Y_i^t
Mean	9.14	5.29	4.00	1.00	16.00	11.00
Variance	40.41	32.49	7.75	1.25	1.67	17.00
Covariance	31.03		2.13		1.00	

Estimating ATE with Block Random Assignment

$$ATE = \sum_{j=1}^{J} \frac{N_j}{N} ATE_j$$

- Where J is the number of blocks and $\frac{N_j}{N}$ is the share of all subjects in block j
- Weighted average of the block-specific ATEs

Observed Outcomes

Village	Block	$Y_i(0)$	$Y_i(1)$
1	A	0	?
2	A	1	?
3	A	?	1
4	A	4	?
5	A	4	?
6	A	6	?
7	A	6	?
8	A	?	3
9	B	14	?
10	B	?	9
11	B	16	?
12	B	16	?
13	B	17	?
14	B	?	17
⋮	⋮	⋮	⋮

SE with Clustered Design

$SE(\widehat{ATE})$ with complete random assignment

$$= \sqrt{\frac{1}{k-1}\left\{\frac{mVar(\bar{Y}_i^c)}{N-m} + \frac{(N-m)Var(\bar{Y}_j^t)}{m} + 2Cov(Y_j^c, Y_j^t)\right\}}$$

$$= \sqrt{\frac{1}{13}\left\{\frac{4(40.41)}{10} + \frac{(10)(32.49)}{4} + 2(31.03)\right\}}$$

$$= 3.50$$

$SE(\widehat{ATE})$ with block random assignment

$$= \sqrt{SE_1^2\left(\frac{N_1}{N}\right)^2 + SE_2^2\left(\frac{N_2}{N}\right)^2}$$

$$= \sqrt{(1.23)^2\left(\frac{8}{14}\right)^2 + (2.71)^2\left(\frac{6}{14}\right)^2}$$

$$= 1.36$$

Estimating ATE with Block Random Assignment

$$\begin{aligned}
\widehat{ATE} &= (\widehat{ATE}_1)\left(\frac{N_1}{N}\right) + (\widehat{ATE}_2)\left(\frac{N_2}{N}\right) \\
&= (-1.5)\left(\frac{8}{14}\right) + (-2.75)\left(\frac{6}{14}\right) \\
&= -2.04
\end{aligned}$$

Standard Error of the Estimated *ATE*

$$\widehat{SE}(\widehat{ATE}) = \sqrt{\widehat{SE}_1^2 \left(\frac{N_1}{N}\right)^2 + \widehat{SE}_2^2 \left(\frac{N_2}{N}\right)^2}$$

where for each of the two blocks:

$$\widehat{SE} = \sqrt{\frac{\widehat{Var}(Y_i^c)}{N - m} + \frac{\widehat{Var}(Y_i^t)}{m}}$$

Placebo Design

- Researchers attempt to contact individuals assigned to receive the treatment
- Those reached are then randomly allocated to two different groups
  - Treatment group
  - Placebo group receiving a "non-treatment"
- Nickerson (2008) canvassing experiment
  - GOTV (treatment)
  - Recycling (placebo)
- CACE estimated by comparing the outcomes for those in the treatment group to those in the placebo group
  - Random sample of Compliers whose untreated potential outcomes can be measured

Nickerson 2008

	Denver		Minneapolis		Pooled	
	Direct	Secondary	Direct	Secondary	Direct	Secondary
Percent Voting in	47.7%	42.4%	27.1%	23.6%		
GOTV Group	(3.0)	(2.9)	(3.1)	(3.0)		
Percent Voting in	39.1%	36.9%	16.2%	17.3%		
Recycling Group	(2.9)	(2.9)	(2.7)	(2.7)		
Estimated Treatment	**8.6%**	**5.5%**	**10.9%**	**6.4%**	**9.8%**	**6.0%**
Effect	**(4.2)**	**(4.1)**	**(4.1)**	**(4.1)**	**(2.9)**	**(2.9)**
P-Value	0.02	0.09	<0.01	0.06	<0.01	0.02

Note. Numbers in parentheses represent standard errors. P-values test the one-tailed hypothesis. Pooled estimates are weighted averages of results for both cities.

Placebo Design

- Logic is that placebo design screens out Never-Takers
- Compliers in the treatment group are compared directly to Compliers in the untreated group
- Reduces noise from Never-Takers in both treatment and control groups
- Moves us to a world of "full compliance"

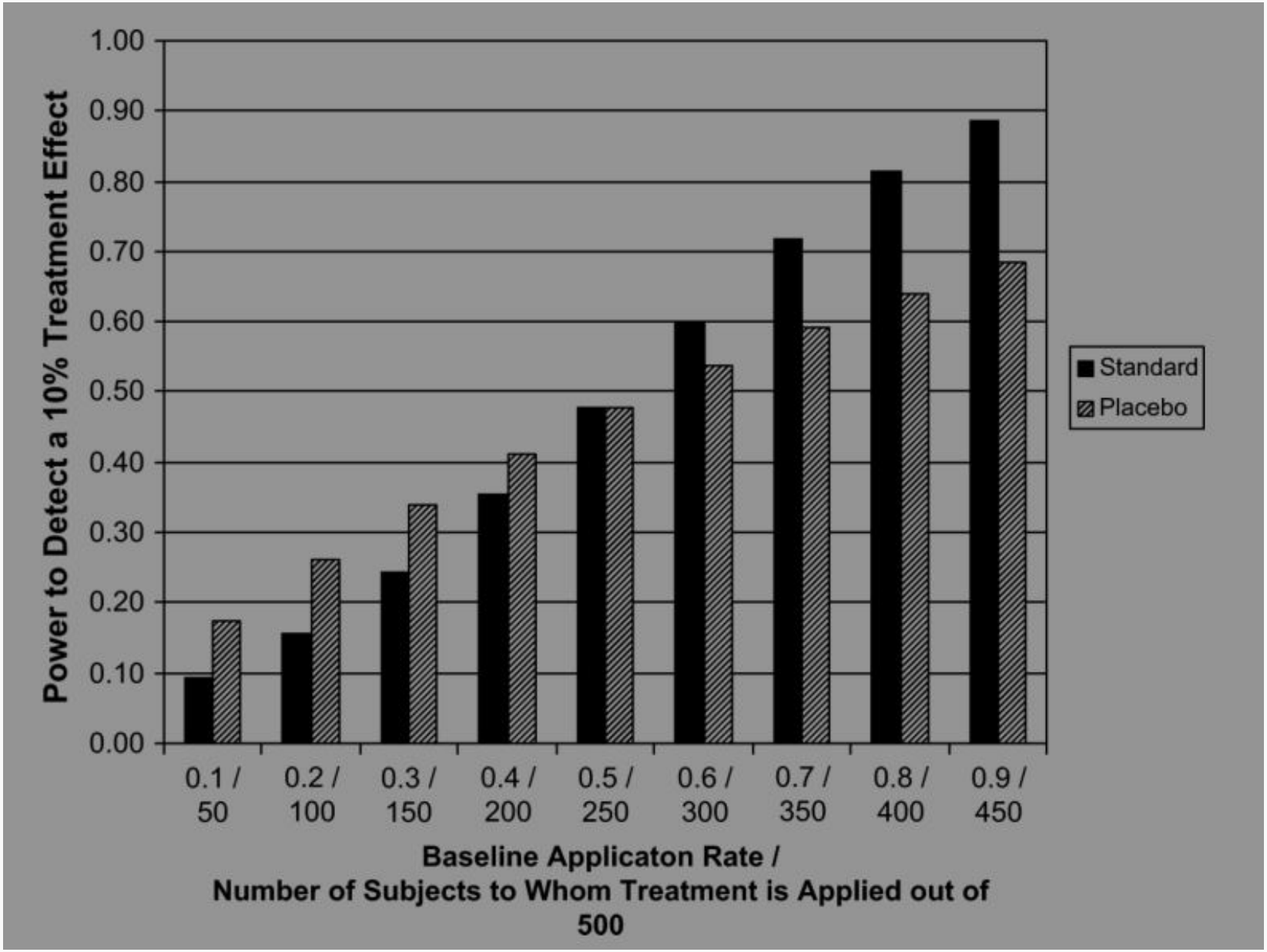
Placebo Design

- Downside is that not all Compliers receive the treatment
- Resources are wasted on those receiving the placebo
- Opportunity to collaborate with someone studying an unrelated topic

Placebo Design

- The placebo and conventional design both allow estimation of the CACE
- Choice depends on the budget and compliance rate
- Under a fixed budget, the conventional design is preferable if compliance rate $> 50\%$
- Canvassing studies often have a lower rate
- A pilot study may give a better idea of the expected compliance rate

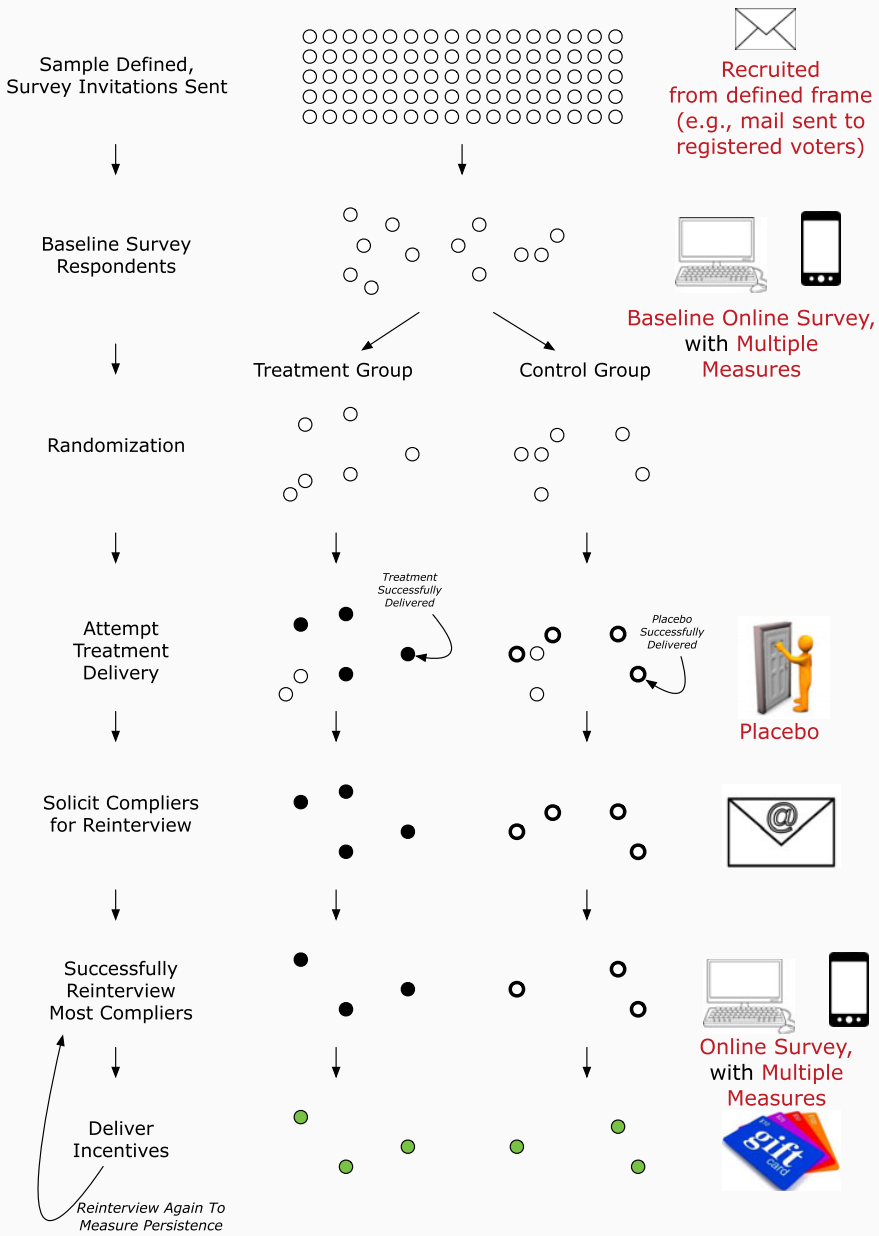
Nickerson 2008



Efficient Design Strategies

- Defined sampling frame
- Placebo
- Pre-treatment baseline survey
- Multiple measures
- Online survey mode

Broockman et al 2017



Challenges

- Failure to treat
- Survey non-response
- Limited pre-treatment covariates

Defined Sampling Frame

- sample recruited from a list sampling frame of the target population units – e.g., a public list of registered voters, a civil registry, membership lists of an activist group, etc.
- Examples:
  - Barber et al 2014
  - Iyengar and Vavreck 2012
  - Himelein 2015

Broockman et al 2017

Notation	Definition	Value used in examples
	Design parameters	
σ^2	True variance of potential outcomes	1
V^*	Target variance of a prospective study	0.002
F	Number of rounds of posttreatment follow-up surveys	2
	Treatment parameters	
N	Number of subjects assigned to treatment and control or placebo in total, with $\frac{N}{2}$ assigned to each condition	
A	Proportion of subjects attempted for treatment that are successfully treated	$\frac{1}{4}$
T	Marginal cost of attempting treatment or placebo contact	\$3
	Survey parameters	
$S_{\text{Mode}\in\{O,T\},\text{Measures}\in\{S,M\}}$	Marginal cost of completed survey; with either Online or Telephone mode and Single or Multiple measures	\$5, except $S_{T,M}$ = \$10
$R_{\text{Wave}\in\{1,2\},\text{Mode}\in\{O,T\}}$	Response rate to a first (1) or second (2) round of surveys, collected Online (O) or by Telephone (T). A first round of surveys could refer to a baseline survey before treatment or an endline survey after treatment when there has been no baseline survey. A second round implies only subjects who answered a first round of surveys are solicited.	$R_{1,O} = 0.07$, $R_{1,T} = 0.07$, $R_{2,O} = 0.75$, $R_{2,T} = 0.35$
$\rho^2_{\text{Mode}\in\{O,T\},\text{Measures}\in\{S,M\}}$	R^2 of regression of outcome at follow-up on pretreatment covariates at baseline; with either Online or Telephone mode and Single or Multiple measures	$\rho^2_{O,S} = 0.25$, $\rho^2_{O,M} = 0.81$, $\rho^2_{T,S} = 0.16$, $\rho^2_{T,M} = 0.33$

Placebo Efficiencies

$$C_{p=1,b=0}(V^*, T) = 4(\frac{\sigma^2}{V^*})(\frac{1}{A})T \tag{4}$$

Placebo is more efficient when...

$$4(\frac{\sigma^2}{V^*})(\frac{1}{A})T < 2(\frac{\sigma^2}{V^*})(\frac{1}{A^2})T \tag{5}$$

Which reduces to... $A < \frac{1}{2}$

- group receiving the placebo can serve as the baseline for comparison for the treatment group
- subjects not contacted in the treatment or placebo groups/all noncompliers do not need to be surveyed.

Pre-treatment baseline survey

- capture pretreatment covariates that analysts can use to increase precision
- can also decrease treatment costs by identifying subjects who are more likely to be interviewed after treatment

$$C_{p=1,b=1}(V^*, F, T, S) = 4(1-\rho_2)(\frac{\sigma^2}{V^*})(FS_0 + \frac{T+S_0}{AR_{2,0}}) \quad (6)$$

Multiple Measure Index

- with multiple measures, baselines can reduce sampling error tremendously
- multiple measures can increase precision even when one item is stable, such as vote choice or partisanship can be; for example, increasing ρ from 0.9 to 0.95 would decrease costs by roughly half

Online Survey Mode

- online surveys can increase reinterview rates after baseline surveys, increasing R_2.
- we have observed $R_{2,O} > R_{2,T}$
- Surveys that collect multiple measures are cheaper to implement online

Sequential Randomised Experiments

Types of Sequential Randomised Experiments

- Non-adaptive - assignment probabilities fixed
- Treatment-adapative - change based on number of subjects in treatment
- Covariate-adaptive - change based on covariate profiles of new and previous subjects
- Responsive-adaptive - change as function of previous units' outcomes

Biased coin method with discrete covariates

- t, letting $t \in (1, 2)$ corresponds to treatment control and treatment conditions
- A unit enters the experiment but prior to randomisation
- J discrete covariates are measured, indexed by j
- The jth covariate has l_j levels, indexed by $i \in (1, ..., l_j)$
- When the current subject arrives, all previous subjects' covariate values and treatment assignments are known

Randomisation protocol

$$S_p = sgn(n_{ij1} - n_{ij2})$$

- If $S_p < 0$ assign the current subject to treatment with some probability $\pi > \frac{1}{2}$
- Some consensus that $\frac{2}{3}$ or $\frac{3}{4}$

$$\text{If } S_p = 0 \text{ let } \pi = \frac{1}{2} \text{If } S_p < 0 \text{ let } \pi < \frac{1}{2}$$

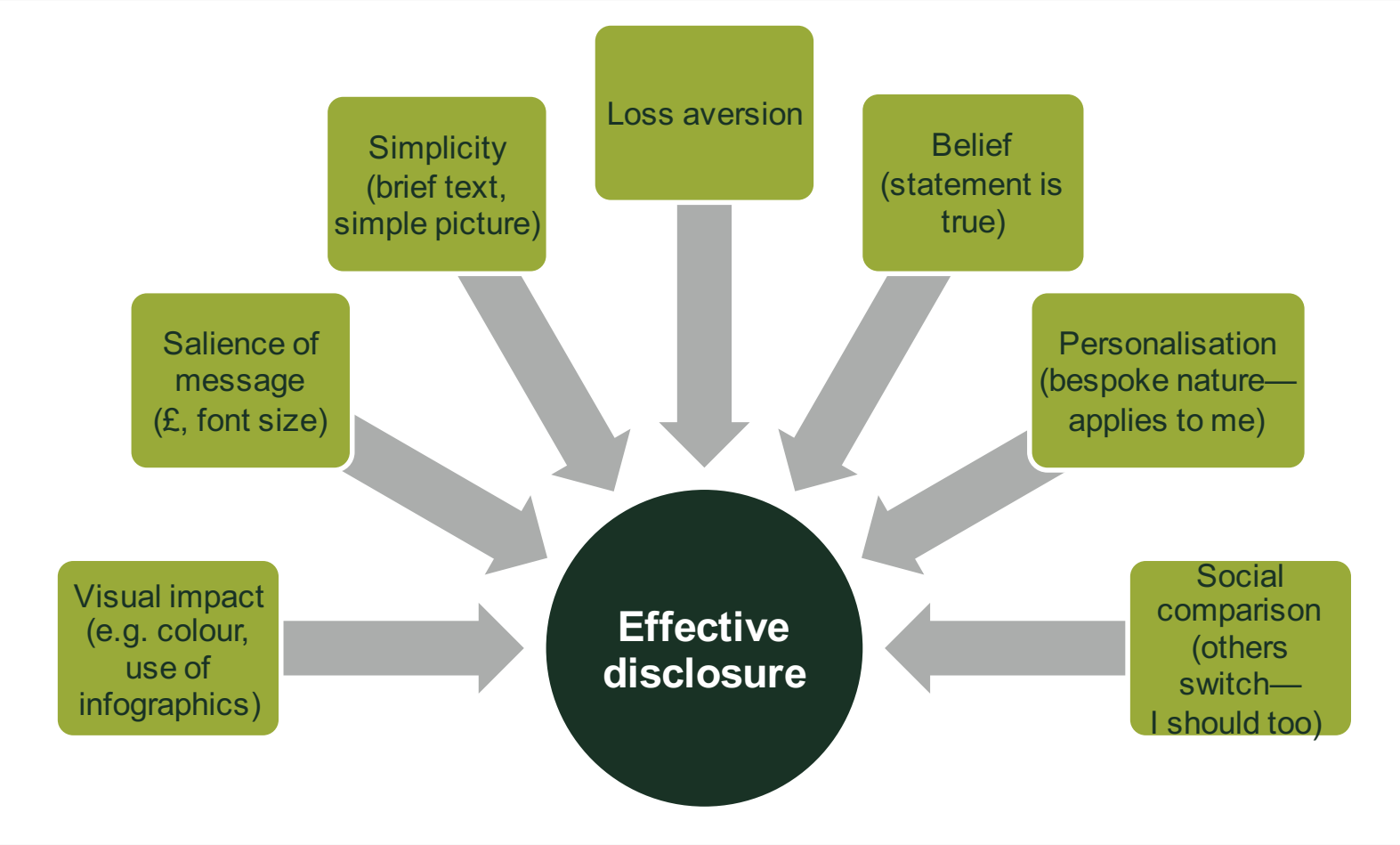
Table 2: Consistent marginal and joint distributions of two binary covariates

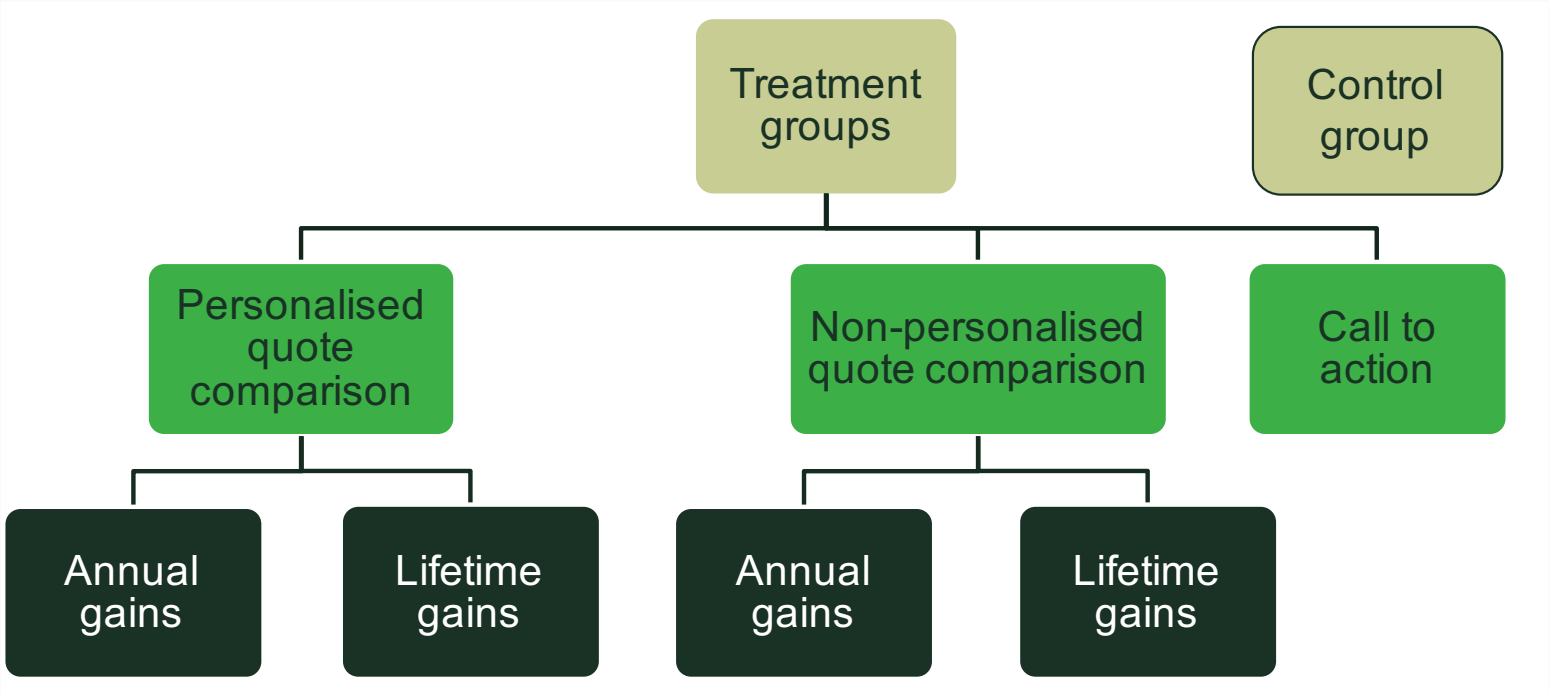
	Sex		*Party*	
	M	*F*	*Rep*	*Dem*
Control	3	3	3	3
Treatment	3	3	3	3
	Rep M	*Dem M*	*Rep F*	*Dem F*
Control	3	0	0	3
Treatment	0	3	3	0

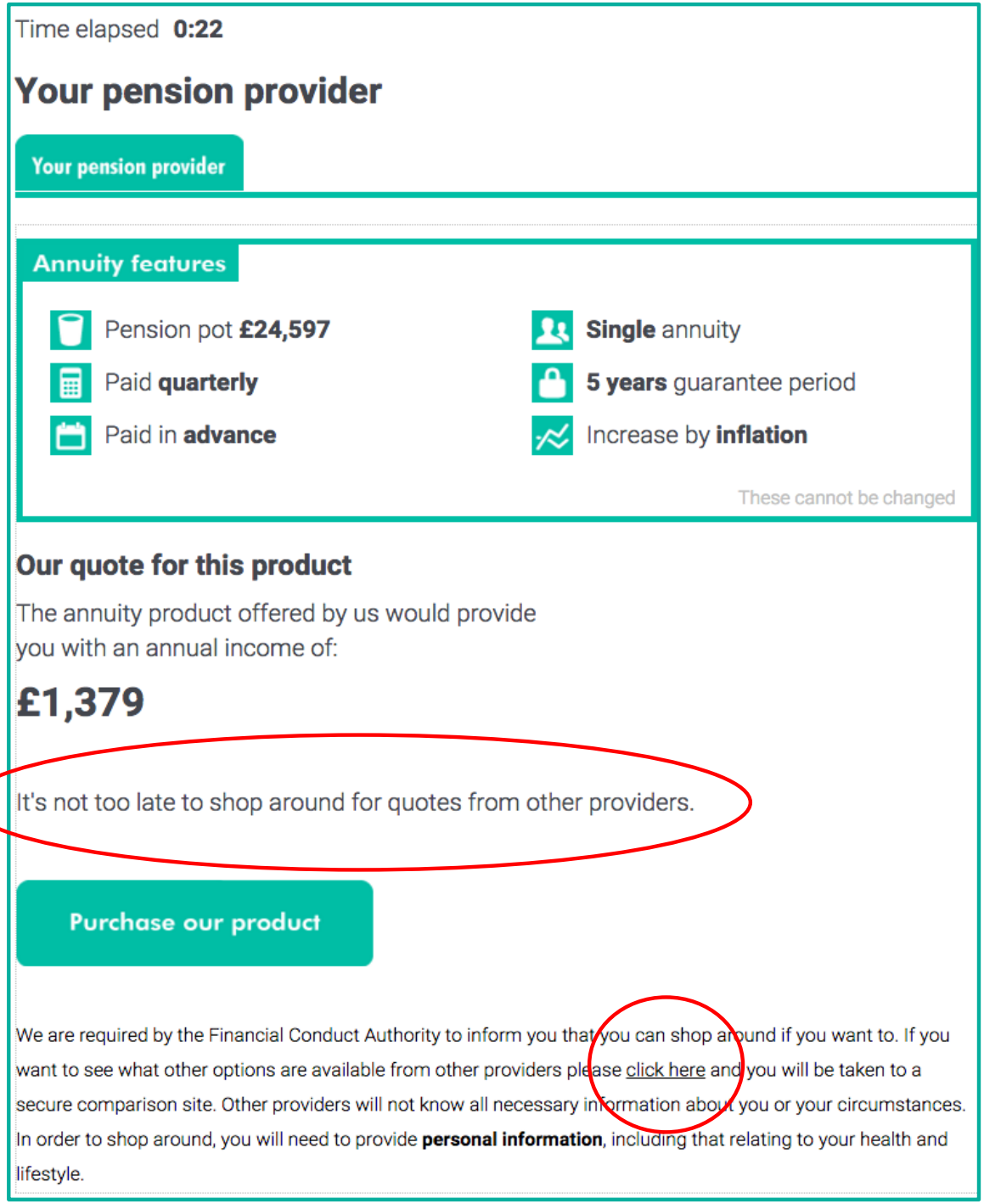
Note. The upper panel reflects the minimization approach and suggests balance between treatment conditions. However, the lower panel makes clear that results will represent isolated parts of the covariate support.

Internet Experiments @ CESS

- Online subject pool
  - Third party subject pools
  - UK CESS online subject pool
  - Santiago CESS online subject pool
- Programming
  - Qualtrics
  - Customized (FCA Example)


Loss aversion
Simplicity (brief text, simple picture)
Belief (statement is true)
Salience of message (£, font size)
Personalisation (bespoke nature—applies to me)
Visual impact (e.g. colour, use of infographics)
Effective disclosure
Social comparison (others switch—I should too)


Treatment groups
Control group
Personalised quote comparison
Non-personalised quote comparison
Call to action
Annual gains
Lifetime gains
Annual gains
Lifetime gains


Time elapsed 0:22
Your pension provider
Your pension provider
Annuity features
Pension pot £24,597
Paid quarterly
Paid in advance
Single annuity
5 years guarantee period
Increase by inflation
These cannot be changed
Our quote for this product
The annuity product offered by us would provide you with an annual income of:
£1,379
It's not too late to shop around for quotes from other providers.
Purchase our product
We are required by the Financial Conduct Authority to inform you that you can shop around if you want to. If you want to see what other options are available from other providers please click here and you will be taken to a secure comparison site. Other providers will not know all necessary information about you or your circumstances. In order to shop around, you will need to provide personal information, including that relating to your health and lifestyle.

Time elapsed **0:34**

Your pension provider

Your pension provider

Annuity features

- Pension pot **£24,597**
- Paid **quarterly**
- Paid in **advance**
- **Single** annuity
- **5 years** guarantee period
- Increase by **inflation**

These cannot be changed

Our quote for this product

The annuity product offered by us would provide you with an annual income of:

£1,379

Can you get a better income from your annuity?

Based on your key information, there are quotes available from other providers offering higher rates. If you select our product you would be losing out on **£46** a year.

Purchase our product

We are required by the Financial Conduct Authority to inform you that you can shop around if you want to. If you want to see what other options are available from other providers please click here and you will be taken to a secure comparison site. Other providers will not know all necessary information about you or your circumstances. In order to shop around, you will need to provide **personal information**, including that relating to your health and lifestyle

Time elapsed **0:27**

Your pension provider

Your pension provider

Annuity features

- Pension pot **£24,597**
- Paid **quarterly**
- Paid in **advance**
- **Single** annuity
- **5 years** guarantee period
- Increase by **inflation**

These cannot be changed

Our quote for this product

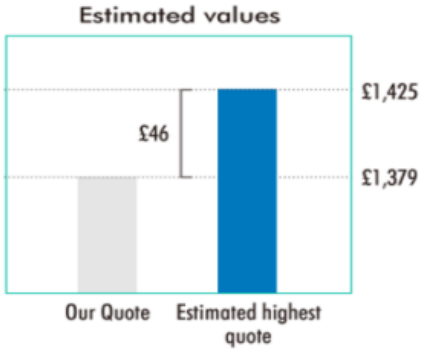
The annuity product offered by us would provide you with an annual income of:

£1,379

Can you get a better income from your annuity?

Based on your key information, we have estimated the highest annuity income you might be offered by other providers. Based on this estimate, if you select our product, you might lose out on around **£50** a year.

This estimate does not use real-time quotes from other annuity providers. As a result, the estimate may be higher or lower than the annuity quotes you would actually be offered, were you to shop around.



Purchase our product

We are required by the Financial Conduct Authority to inform you that you can shop around if you want to. If you want to see what other options are available from other providers please click here and you will be taken to a secure comparison site. Other providers will not know all necessary information about you or your circumstances. In order to shop around, you will need to provide **personal information**, including that relating to your health and

Time elapsed **0:11**

Your pension provider

Your pension provider

Annuity features

- Pension pot **£24,597**
- Paid **quarterly**
- Paid in **advance**
- **Single** annuity
- **5 years** guarantee period
- Increase by **inflation**

These cannot be changed

Our quote for this product

The annuity product offered by us would provide you with an annual income of:

£1,379

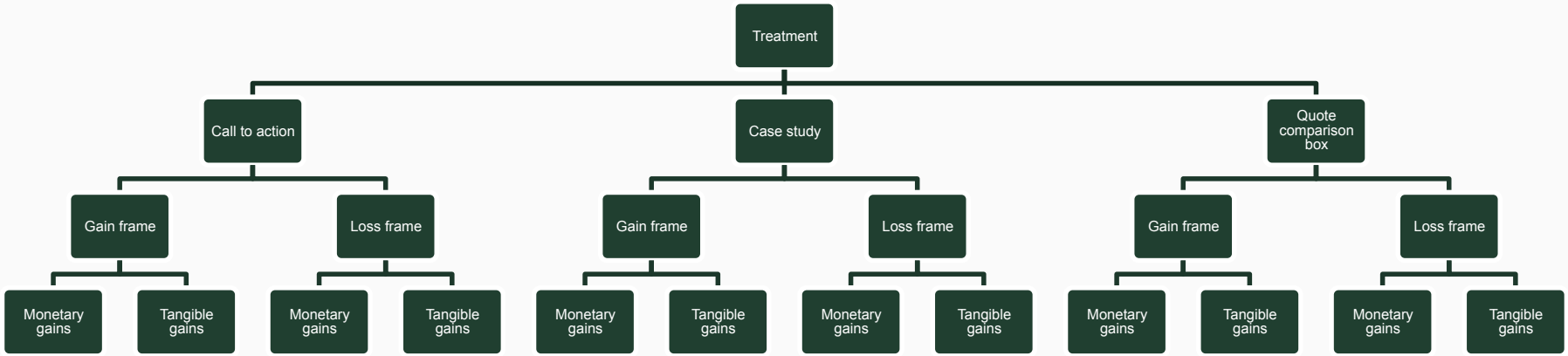
Can you get a better income from your annuity?

80% of people lose out by purchasing from their own pension provider according to a 2014 survey.

Purchase our product

We are required by the Financial Conduct Authority to inform you that you can shop around if you want to. If you want to see what other options are available from other providers please click here and you will be taken to a secure comparison site. Other providers will not know all necessary information about you or your circumstances. In order to shop around, you will need to provide **personal information**, including that relating to your health and lifestyle.

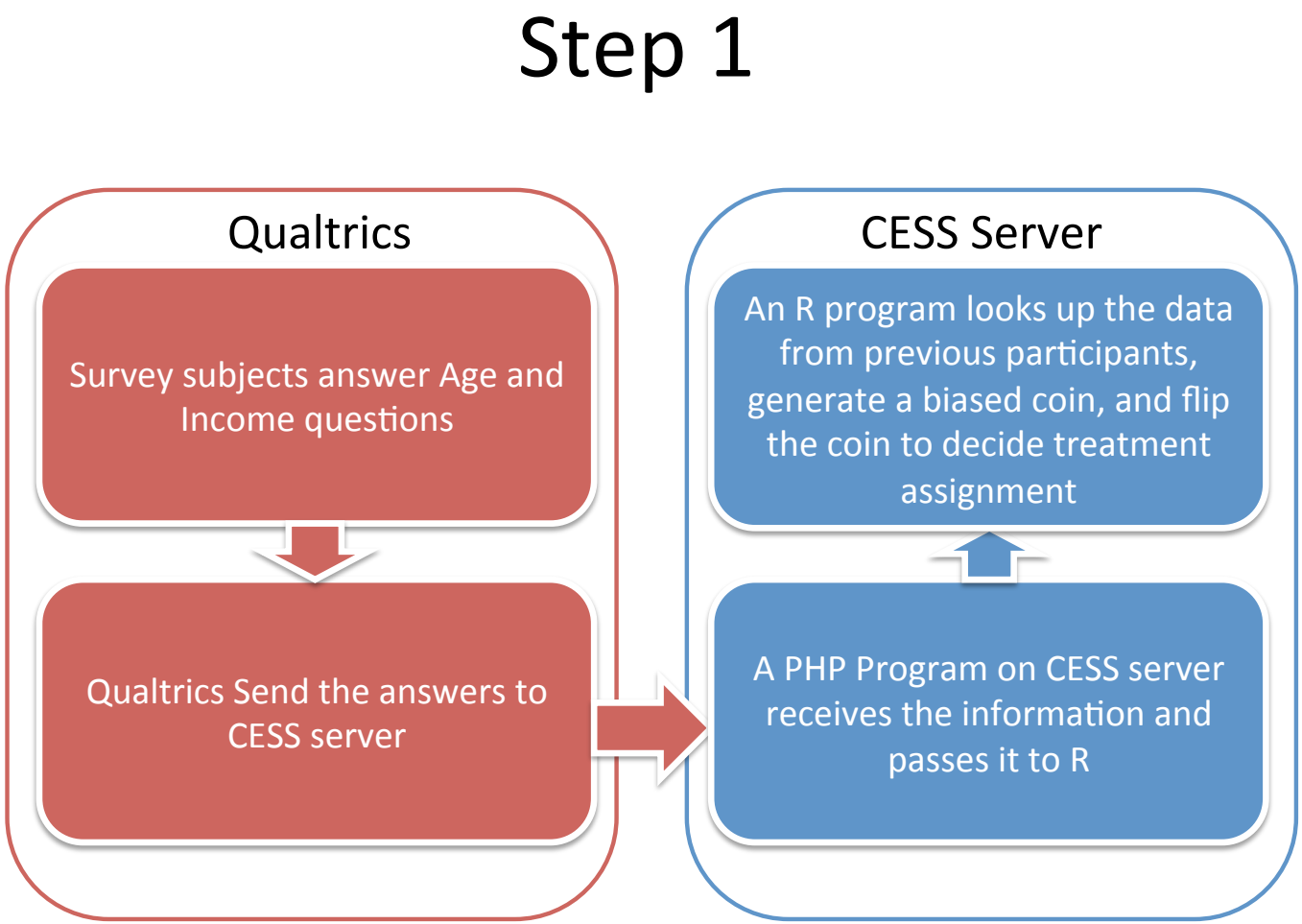
FCA: Assignment to Treatments



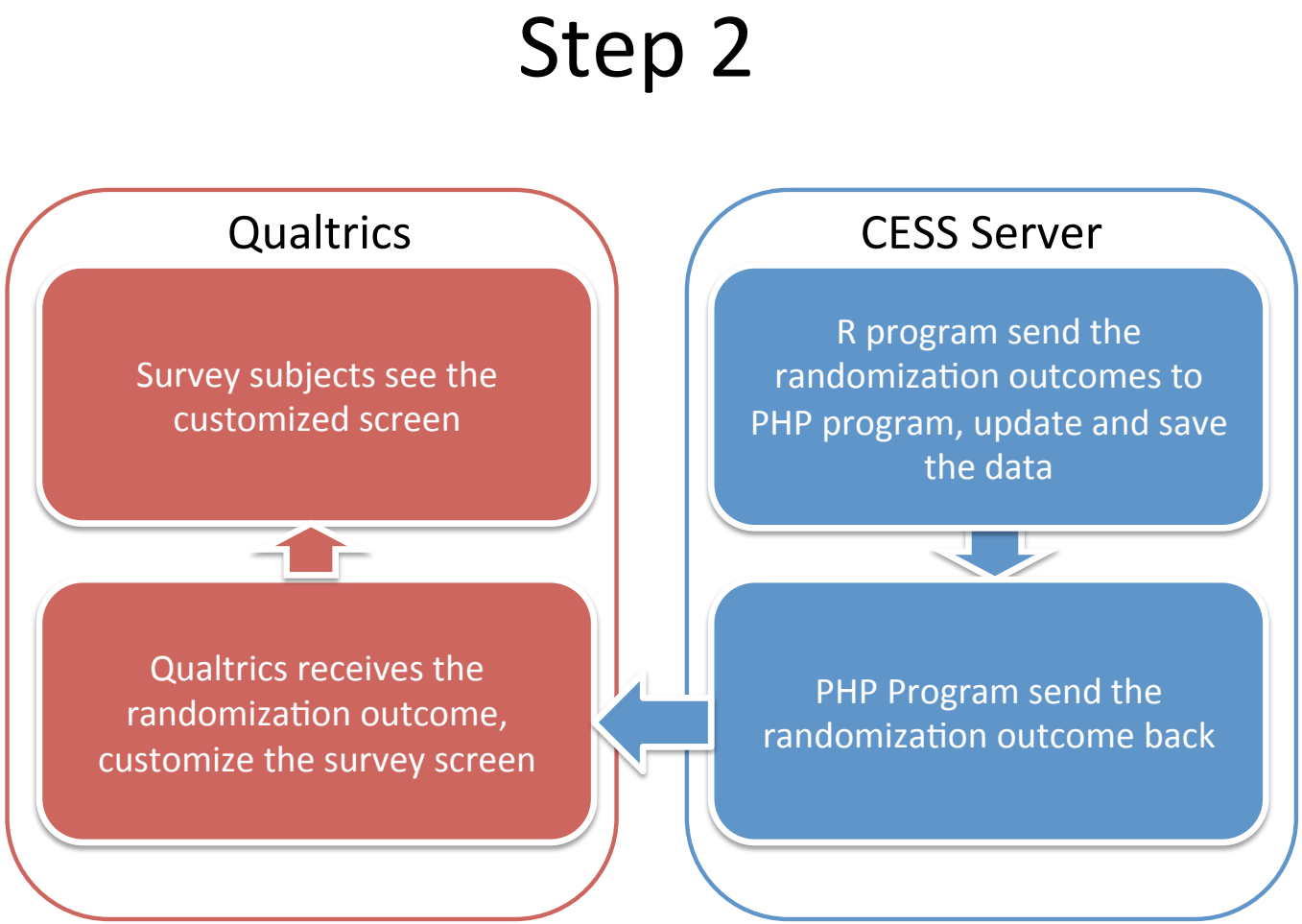
How to Assign to Treatment?

- Block random assignment (sequential online)
- Subjects are partitioned into blocks
- Complete random assignment within each block
- Example
  - Split subjects by gender
  - Select 5 subjects from the male group for the treatment
  - Select 5 subjects from the female group for the treatment

Step 1


Step 1
Qualtrics
Survey subjects answer Age and Income questions
Qualtrics Send the answers to CESS server
CESS Server
An R program looks up the data from previous participants, generate a biased coin, and flip the coin to decide treatment assignment
A PHP Program on CESS server receives the information and passes it to R

Step 2


Step 2
Qualtrics
Survey subjects see the customized screen
Qualtrics receives the randomization outcome, customize the survey screen
CESS Server
R program send the randomization outcomes to PHP program, update and save the data
PHP Program send the randomization outcome back

Conjoint Experiments

- Typical survey experiments test uni-dimensional causal effect
    - Treatment versus Control
- Typical vignette experiments: immigration, candidate, racial cue
- But what components of manipulation produce observed effect?
    - Why does immigration status matter?

Hainmueller et al

- Political Analysis 2013 Causal Inference in Conjoint Analysis
- Understanding Multidimensional Choices via State Preference Experiments
- The Hidden American Immigration Consensus: A Conjoint Analysis of Attitudes towards Immigrants
- Validating Vignette and Conjoint Survey Experiments Against Real-World Behaviour

Example: Candidates

- Respondents chose between/rank 2 candidates
  - J=2 (choices)
  - K=6 choices/evaluations
- Each candidate has a profile
  - Each profile has a set of L discretely valued attributes, or a treatment composed of L components
  - We use D to denote the total number of levels for attribute l
  - L=8 (candidate attributes, D(1)...D(6) (total number of levels for candidate's age, education, etc.), and D(7)...D(8)=2 (for military service and gender)

Please read the descriptions of the potential immigrants carefully. Then, please indicate which of the two immigrants you would personally prefer to see admitted to the United States.

	Immigrant 1	Immigrant 2
Prior Trips to the U.S.	Entered the U.S. once before on a tourist visa	Entered the U.S. once before on a tourist visa
Reason for Application	Reunite with family members already in U.S.	Reunite with family members already in U.S.
Country of Origin	Mexico	Iraq
Language Skills	During admission interview, this applicant spoke fluent English	During admission interview, this applicant spoke fluent English
Profession	Child care provider	Teacher
Job Experience	One to two years of job training and experience	Three to five years of job training and experience
Employment Plans	Does not have a contract with a U.S. employer but has done job interviews	Will look for work after arriving in the U.S.
Education Level	Equivalent to completing two years of college in the U.S.	Equivalent to completing a college degree in the U.S.
Gender	Female	Male

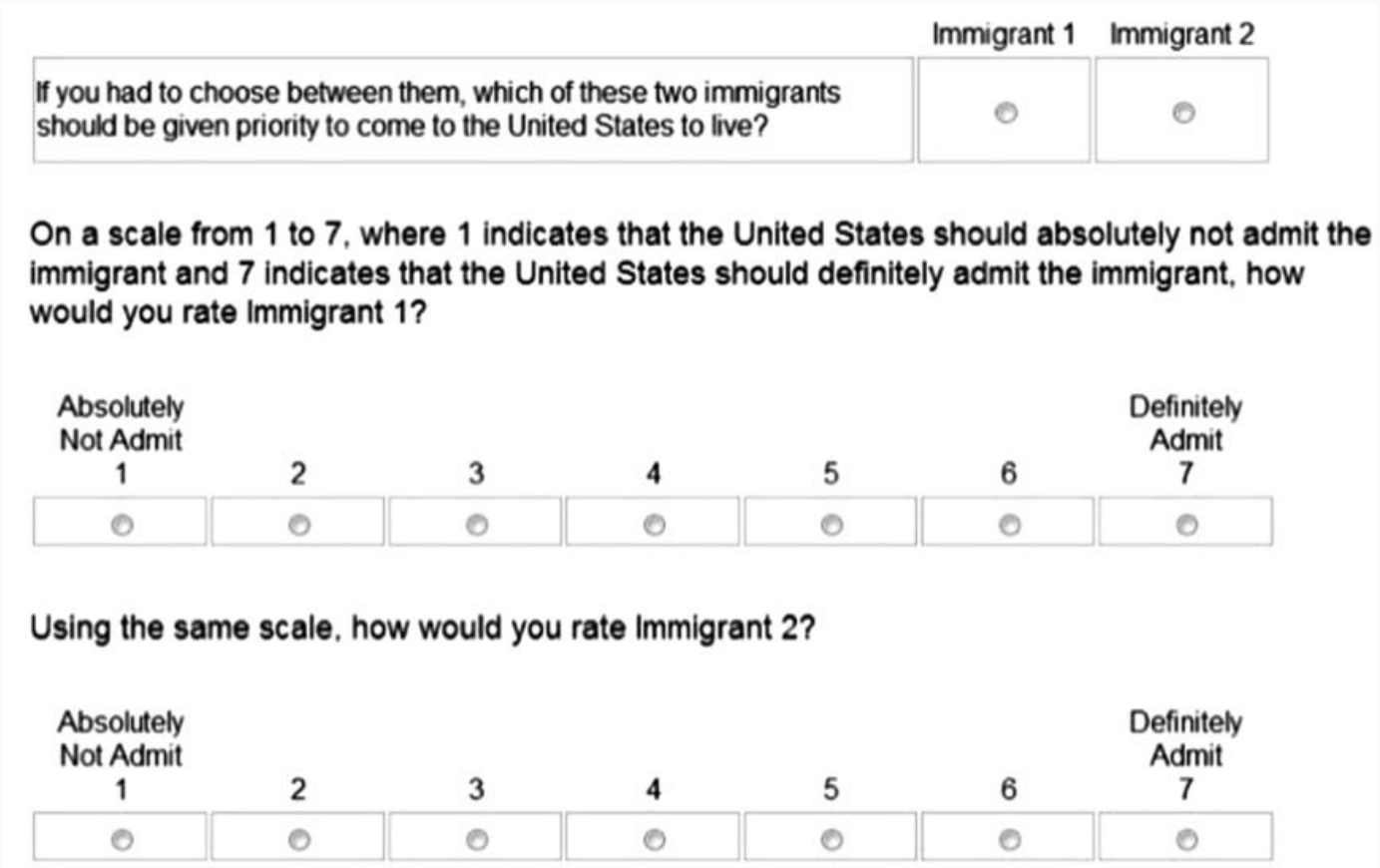


Fig. 1 Experimental design: Immigration conjoint. This figure illustrates the experimental design for the conjoint analysis that examines immigrant admission to the United States.

Respondent's JK Profiles

$$\sum_{j=1}^{J} Y_{ijk}(\bar{\mathbf{t}}) = 1$$

Stability Assumption

- Potential outcomes always take on the same value as long as all the profiles in the same choice task have identical sets of attributes

$$Y_{ijk}(\bar{T}_i) = Y_{ijk^t}(\bar{T}_i)$$

$$\text{if } \bar{T}_{ik} = \bar{T}_{ik^t}$$

$$\text{for any } j, k, k^l$$

No Profile Order Effect

$$Y_{ij}(T_{ik}) = Y_{ij'}(T'_{ik})$$

$$\text{if } T_{ijk} = T'_{ij'k}$$

$$\text{and } T_{ij'k} = T'_{ijk}$$

$$\text{for any } i, j, j', k$$

Randomisation of Profiles

$$Y_i(\mathbf{t}) \perp T_{ijkl} \text{ for any } i, j, k, l$$

- Pairwise independence between all elements of $Y_i(t)$ and T_ijkl and $0 < p(t) = p(T_{ik} = t) < 1)$

Basic Profile Effects

$$\pi(t_1, t_0) = Y_i(t_1) - Y_i(t_0)$$

Profiles	Candidate	Service	Income	Eduction
t_0	1	military	rich	college
	2	no service	poor	college
t_1	1	military	rich	college
	2	military	poor	college

Estimate Profile Effects

- Unit-level causal effects are difficult to identify
  - Involve counterfactuals and hence fundamental problem of causal inference
- Average Treatment Effects (ATE)?
  - If there are a large number of attributes with multiple levels the number of observations in each conditioning set will be virtually zero rendering estimation difficult if not impossible

Average Marginal Component Effect

$$\hat{\bar{\pi}}_1(t_1, t_0, p(\mathbf{t})) = \sum_{(t,\mathbf{t})\in\tilde{\tau}} \{\mathbb{E}[Y_{ijk}|\, T_{ijkl} = t_1,\, T_{ijk[-l]} = t,\, \mathbf{T}_{i[-j]k} = \mathbf{t}]$$
$$- \mathbb{E}[Y_{ijk}|\, T_{ijkl} = t_0,\, T_{ijk[-l]} = t,\, \mathbf{T}_{i[-j]k} = \mathbf{t}]\}$$
$$\times\, p(T_{ijk[-l]} = t, \mathbf{T}_{i[-j]k} = \mathbf{t}|(T_{ijk[-l]}, \mathbf{T}_{i[-j]k}) \in \tilde{\tau})$$

- The marginal effect of attribute l averaged over the joint distribution of the remaining attributes

Estimating AMCE

- For any attribute of interest T_{ijkl} the subclassification estimate of the AMCE can be computed simply by dividing the sample into the strata defined by T_{ijk}
- Typically the attributes on which the assignment of the attribute of interest is restricted
- Calculate the difference in the average observed choice outcomes between the treatment ($T_{ijkl} = 1$) and control ($T_{ijkl} = 0$) groups within each stratum
- Take the weighted average of these differences in means, using the known distribution of the strata as the weights

Regression Estimation

- The linear regression estimator is fully nonparametric, even though the estimation is conducted by a routine typically used for a parametric linear regression model
- Regress the outcome variable on the L sets of dummy variables
- Interaction terms for the attributes that are involved in any of the randomization restrictions used in the study
- Take the weighted average of the appropriate coefficients

Variance Estimation

- Observed choice outcomes within choice tasks strongly negatively correlated
- Both potential choice and rating outcomes within respondents are likely to be positively correlated because of unobserved respondent characteristics influencing their preferences
- Point estimates of the AMCE can be coupled with standard errors corrected for within respondent clustering
- Obtain cluster-robust standard errors for the estimated regression coefficients by using the cluster option in Stata
- Block bootstrap where respondents are resampled with replacement and uncertainty estimates are calculated based on the empirical distribution of the AMCE over the resamples

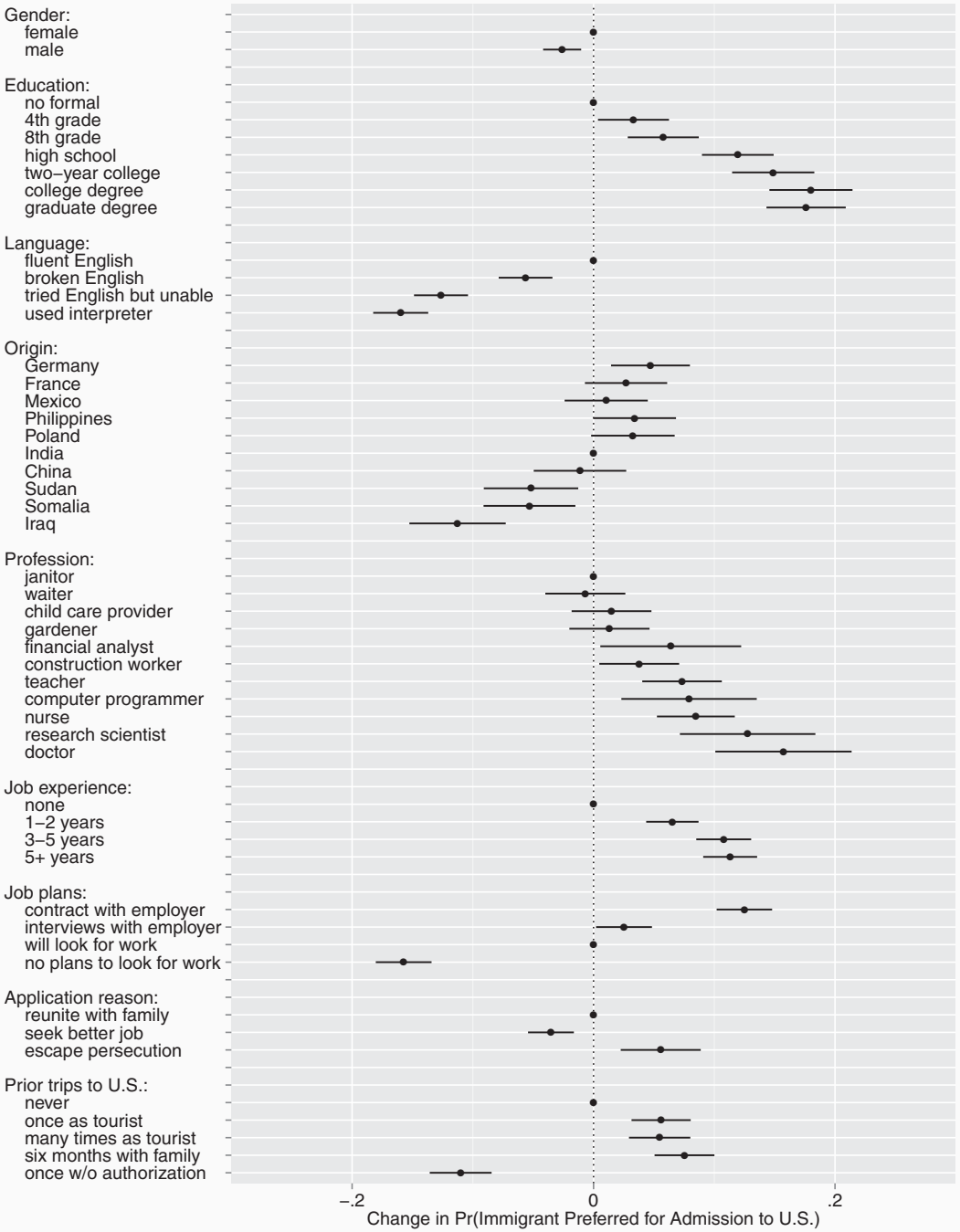
Example: Candidate Experiment

- 3,466 rated profiles - 1,733 pairings
- 311 respondents
- Design yields 186,624 possible profiles - far exceeds number of completed tasks
- Respondents rated each candidate profile on a seven-point scale, where 1 indicates that the respondent would "never support" the candidate and 7 indicates that she would "always support" the candidate
- Rescaled to 0 and 1

AMCE for candidate age levels

$$\begin{aligned}\text{rating}_{ijk} =& \beta_0 + \beta_1[\text{age}_{ijk} = 75] + \beta_2[\text{age}_{ijk} = 68] + \\ & \beta_3[\text{age}_{ijk} = 60] + \beta_4[\text{age}_{ijk} = 52] + \\ & \beta_5[\text{age}_{ijk} = 45] + \epsilon\end{aligned}$$

- The reference category is 36 years old
- βs are estimators for AMCE for ages 68, 75, etc. compared to 36


Gender:
female
male
Education:
no formal
4th grade
8th grade
high school
two−year college
college degree
graduate degree
Language:
fluent English
broken English
tried English but unable
used interpreter
Origin:
Germany
France
Mexico
Philippines
Poland
India
China
Sudan
Somalia
Iraq
Profession:
janitor
waiter
child care provider
gardener
financial analyst
construction worker
teacher
computer programmer
nurse
research scientist
doctor
Job experience:
none
1−2 years
3−5 years
5+ years
Job plans:
contract with employer
interviews with employer
will look for work
no plans to look for work
Application reason:
reunite with family
seek better job
escape persecution
Prior trips to U.S.:
never
once as tourist
many times as tourist
six months with family
once w/o authorization
−.2
0
.2
Change in Pr(Immigrant Preferred for Admission to U.S.)

CACE External Validity

- Hypothetical versus Behavioural choices
- Treatment versus Control
- Hainmueller et al 2015 PANS
- Behavioural data from Swiss referendum
- Results matched to conjoint experiment

Swiss Behavioural Data

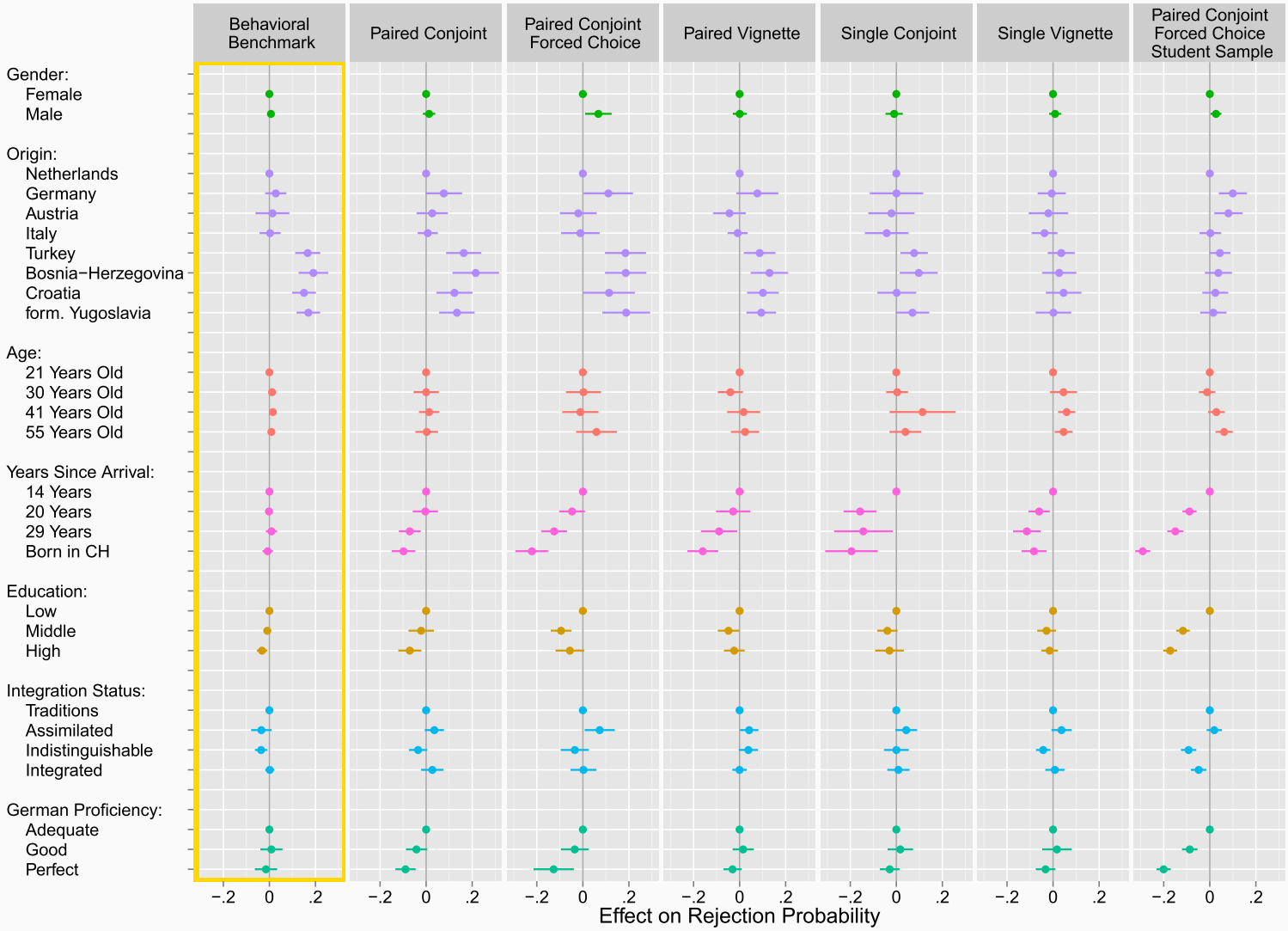
- Municipalities used referendums to vote on the naturalization applications of immigrants
- Voters received a voting leaflet with a short description of the applicant, including information about attributes, such as age, sex, education, origin, language skills, and integration status
- Voters then cast a secret ballot to accept or reject individual applicants
- These voting data yield an accurate measure of the revealed preferences of the voters what components of manipulation produce observed effect

Conjoint Experiment Data

- Respondents are presented with profiles of immigrants and then asked to decide on their application for naturalization
- List of attributes matches attributes voters saw no the voting leaflets distributed for the referendums - presented in same order as on the original leaflets
- Each respondent is randomly assigned to one of five different designs and asked to complete 10 choice tasks

Experimental Designs

Designs	Profiles
single-profile vignette	accept/reject single profile
paired-profile vignette	accept/reject two profiles
single-profile conjoint	name/value of attributes
paired-profile conjoint	accept/reject 2 applicants
paired-profile conjoint	accept/reject 1 of 2 applicants


Behavioral Benchmark
Paired Conjoint
Paired Conjoint Forced Choice
Paired Vignette
Single Conjoint
Single Vignette
Paired Conjoint Forced Choice Student Sample
Gender:
Female
Male
Origin:
Netherlands
Germany
Austria
Italy
Turkey
Bosnia−Herzegovina
Croatia
form. Yugoslavia
Age:
21 Years Old
30 Years Old
41 Years Old
55 Years Old
Years Since Arrival:
14 Years
20 Years
29 Years
Born in CH
Education:
Low
Middle
High
Integration Status:
Traditions
Assimilated
Indistinguishable
Integrated
German Proficiency:
Adequate
Good
Perfect
−.2
0
.2
Effect on Rejection Probability

Conjoint: Causal Interaction

- Egami et al 2016
- Average Marginal Interaction Effect (AIME)
- Does not depend on choice of baseline
- Non-parametrically estimated using ANOVA regression with weighted zero-sum constraints
- Reduces false discovery rate and facilitates interpretation

AME and AMIE Estimation

$$Y_i(\mathbf{t}) = \mu + \sum_{j=1}^{4} \sum_{\gamma=0}^{L_j-1} \beta_\gamma^j \mathbf{1}\{t_{ij} = \gamma\} + \sum_{j=1}^{4} \sum_{j' \neq j} \sum_{\gamma=0}^{L_j-1} \sum_{m=0}^{L_j'-1} \beta_{\gamma m}^{ij'} \mathbf{1}\{t_{ij} = \gamma, t_{ij'} = m\} + \epsilon_i(\mathbf{t})$$

$$Pr(Y_i(\mathbf{T}_i^*) > Y_i(\mathbf{T}_i^\dagger) | \mathbf{T}_i^*, \mathbf{T}_i^\dagger) = \tilde{\mu} + \sum_{j=1}^{4} \sum_{\gamma=0}^{L_j-1} \beta_\gamma^j (\mathbf{1}\{T_{ij}^* = \gamma\} - \mathbf{1}\{T_{ij}^\dagger = \gamma\})$$

$$+ \sum_{j=1}^{4} \sum_{j' \neq j} \sum_{\gamma=0}^{L_j-1} \sum_{m=0}^{L_j'-1} \beta_{\gamma m}^{ij'} (\mathbf{1}\{T_{ij}^* = \gamma, T_{ij'}^\dagger = m\} - \mathbf{1}\{T_{ij}^* = \gamma, T_{ij'}^\dagger = m\})$$

	Range	Selection Prob.
AME		
Record	0.122	1.00
Coethnicity	0.053	1.00
Platform	0.023	0.93
Degree	0.000	0.33
AMIE		
Coethnicity × Record	0.053	1.00
Record × Platform	0.030	0.92
Platform × Coethnic	0.008	0.65
Coethnicity × Degree	0.000	0.62
Platform × Degree	0.000	0.35
Record × Degree	0.000	0.09

	Factor	AME	Selection prob.
Record			
	Yes/Village	0.122	
	Yes/District	0.122	$\rangle$ 0.71
	Yes/MP	0.101	$\rangle$ 0.77
	No/Village	0.047	$\rangle$ 1.00
	No/District	0.051	$\rangle$ 0.74
	No/MP	0.047	$\rangle$ 0.74
	No/Businessman	base	$\rangle$ 1.00
Platform			
	Jobs	-0.023	
	Clinic	-0.023	$\rangle$ 0.56
	Education	base	$\rangle$ 0.94
Coethnicity		0.053	1.00
Degree		0.000	0.33

Example

$$\underbrace{\tau(\text{Coethnic, No/Business; Non-coethnic, No/MP})}_{-2.4}$$

$$= \underbrace{\psi(\text{Coethnic; Non-coethnic})}_{5.3} + \underbrace{\psi(\text{No/Business; No/MP})}_{-4.7}$$

$$+ \underbrace{\pi(\text{Coethnic, No/Business; Non-coethnic, No/MP})}_{-3.0}$$

Online Experiments: Modes & Subject Pools

MTurk: Static External Validity

- Berinksy et al 2012
- Whether estimated (average) treatment effects are accurate assessments of treatment effects for other samples
- Whether these estimates are reliable assessments of treatment effects for the same sample outside the MTurk setting

Table 1 Task title, compensation, and speed of completion for selected MTurk studies

Task title	*Date launched*	*Number of subjects*	*Pay per subject*	*Mean minutes per subject*	*Completions per day*								
					1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
Answer a survey about current affairs and your beliefs	January 5, 2010	490	$0.15	7	116	64	41	40	27	36	15	11	15
2- to 3-Min survey for political science research	March 16, 2010	500	$0.25	2	210	68	37	55	53	64	18		
4-Min survey for political science research	April 26, 2010	500	$0.40	4	298	105	79	18					
3-Min survey for political science research	April 29, 2010	200	$0.25	1	200								
3- to 4-Min survey for political science research	May 17, 2010	150	$0.45	2	150								
7- to 9-Min survey	June 24, 2010	400	$0.75	6	400								
5- to 7-Min survey	June 28, 2010	400	$0.75	5	321	79							
5- to 7-Min survey	July 3, 2010	400	$0.50	3	256	115	29						
2- to 3-Min survey	July 16, 2010	200	$0.25	3	200								

Note. The remaining subjects for the January 5, 2010, study were recruited as follows: Day 10 (11), Day 11 (10), Day 12 (17), Day 13 (22), Day 14 (29), and Day 15 (36).

		Convenience Samples			
				Adult samples (Berinsky and Kinder 2006)	
Demographics	*MTurk*	*Student samples (Kam et al. 2007)*	*Adult sample (Kam et al. 2007)*	*Experiment 1: Ann Arbor, MI*	*Experiment 2: Princeton, NJ*
Female	60.1% (2.1)	56.7% (1.3)	75.7% (4.1)	66.0%	57.1%
Age (mean years)	32.3 (0.5)	20.3 (8.2)	45.5 (.916)	42.5	45.3
Education (mean years)	14.9 (0.1)	—	5.48 (1.29)	15.1	14.9
White	83.5 (1.6)	42.5	82.2 (3.7)	81.4	72.4
Black	4.4 (0.9)			12.9	22.7
Party identification					
Democrat	40.8 (2.1)			46.1	46.5
Independent	34.1 (2.0)			20.6	17.6
Republican	16.9 (1.6)			16.3	25.8
None/other	8.2 (1.2)			17.0	10.1
N	484–551	277–1428	109	141	163

Note. Percentages except for age and education with SEs in parentheses. Adult sample from Kam et al. (2007) is for campus employee participants from their Table 1, Column 1. MTurk survey is from February/March 2010.

Table 3 Comparing MTurk sample demographics to Internet and face-to-face samples

	Internet sample		*Face-to-face samples*	
	MTurk	*ANESP*	*CPS 2008*	*ANES 2008*
Female	60.1% (2.1)	57.6% (0.9)	51.7% (0.2)	55.0% (1.3)
Education (mean years)	14.9 (0.1)	16.2 (0.1)	13.2 (0.0)	13.5 (0.1)
Age (mean years)	32.3 (0.5)	49.7 (0.3)	46.0 (0.1)	46.6 (0.5)
Mean income	$55,332 ($1,659)	$69,043 ($794)	$62,256 ($130)	$62,501 ($1,467)
Median income	$45,000	$67,500	$55,000	$55,000
Race				
White	83.5 (1.6)	83.0 (0.7)	81.2 (0.1)	79.1 (0.9)
Black	4.4 (0.9)	8.9 (0.5)	11.8 (0.1)	12.0 (0.6)
Hispanic	6.7 (1.1)	5.0 (0.4)	13.7 (0.1)	9.1 (0.5)
Marital status				
Married	39.0 (2.1)	56.8 (0.9)	55.7 (0.2)	50.1 (1.3)
Divorced	7.1 (1.1)	12.1 (0.6)	10.2 (0.1)	12.9 (0.8)
Separated	2.5 (0.7)	1.3 (0.2)	2.1 (0.1)	2.9 (0.4)
Never married	50.6 (2.1)	14.2 (0.6)	25.7 (0.2)	26.2 (1.1)
Widowed	0.7 (0.4)	4.9 (0.4)	6.3 (0.1)	7.8 (0.6)
Housing status				
Rent	52.7 (2.3)	14.3(0.1)		32 (1.2)
Own home	47.3 (2.3)	80.8 (0.8)		66.1 (1.2)
Religion				
None	41.8 (2.1)	13.1 (0.8)		26.9 (1.2)
Protestant	20.7 (1.7)	38.7 (1.4)		28.2 (1.2)
Catholic	16.5 (1.6)	22.9 (1.0)		17.5 (1.0)
Jewish	4.4 (0.9)	3.0 (0.4)		1.2 (0.3)
Other	16.5 (1.6)	22.2 (1.0)		26.2 (1.1)

Table 4 Comparing MTurk sample political and psychological measures to Internet and face-to-face samples

	Internet sample		*Face-to-face samples*	
	MTurk	*ANESP*	*CPS 2008*	*ANES 2008*
Registration and turnout				
Registered	78.8% (1.7)	92.0% (0.7)	71.0% (0.2)	78.2% (1.1)
Voter turnout 2008	70.6 (2.0)	89.8 (0.5)	63.6 (0.2)	70.4 (1.1)
Party identification (mean on 7-point scale, 7 = Strong Republican)	3.48 (0.09)	3.90 (0.05)		3.70 (0.05)
Ideology (mean on 7-point scale, 7 = Strong conservative)	3.39 (0.09)	4.30 (0.05)		4.24 (0.04)
Political Interest (mean on 5-point scale, 5 = Extremely interested)	2.43 (0.04)	2.71 (0.02)		2.93 (0.03)
Political knowledge (% correct)				
Presidential succession after Vice President	70.0 (1.3)	65.2 (2.0)		
House vote percentage needed to override a veto	81.3 (1.7)	73.6 (1.3)		
Number of terms to which an individual can be elected president	96.2 (0.8)	92.8 (0.7)		
Length of a U.S. Senate term	45.0 (2.1)	37.5 (1.3)		
Number of Senators per state	85.4 (1.5)	73.2 (1.2)		
Length of a U.S. House term	50.1 (2.1)	38.9 (1.3)		
Average	71.3	63.5		
Need for cognition (mean on 0–1 scale)	.625 (0.012)	.607 (0.006)		.559 (0.009)
Need to evaluate (mean on 0–1 scale)	.628 (0.008)	.579 (0.004)		.558 (0.005)
N	506–699	1,466–2,984	92,360	1,058–2,323

Note. Means with SEs in parentheses. CPS 2008 and ANES 2008 are weighted. Political measures are from the February/March 2010 MTurk survey ($N = 551$). Need for Cognition and Need to Evaluate are from the May 2011 MTurk survey ($N = 699$). Tests of statistical significance of differences across samples appear in the Supplementary data.

Table 6 Replication of Table 2 by Kam and Simas (2010)—risk acceptance and preference for the probabilistic outcome

	Kam and Simas (2010)			*MTurk replication*		
	(H1a) Mortality frame and risk acceptance	*(H1b) Adding controls*	*(H2) Frame × Risk acceptance*	*(H1a) Mortality frame and risk acceptance*	*(H1b) Adding controls*	*(H2) Frame × Risk acceptance*
Mortality frame in Trial 1	1.068 (0.10)	1.082 (0.10)	1.058 (0.29)	1.180 (0.10)	1.180 (0.10)	1.410 (0.31)
Risk acceptance	0.521 (0.31)	0.628 (0.32)	0.507 (0.48)	0.760 (0.29)	0.780 (0.31)	0.990 (0.42)
Female		0.105 (0.10)			−0.018 (0.11)	
Age		0.262 (0.22)			0.110 (0.31)	
Education		−0.214 (0.20)			0.025 (0.23)	
Income		0.205 (0.23)			−0.024 (0.23)	
Partisan ideology		0.038 (0.19)			0.006 (0.15)	
Risk acceptance × Mortality frame			0.023 (0.62)			−0.450 (0.58)
Intercept	−0.706 (0.155)	−0.933 (0.259)	−0.700 (0.227)	−1.060 (−0.170)	−1.100 (−0.290)	−1.190 (−0.230)
lnL	−453.185	−450.481	−453.184	−409.740	−409.662	−409.439
$p > \chi^2$	0.000	0.000	0.000	0.000	0.000	0.000
N	752	750	752	699	699	699

Note. Entries are probit coefficients with SEs in parentheses. Dependent variable is Preference for the Probabilistic Outcome (0 = deterministic outcome; 1 = probabilistic outcome). All independent variables are scaled to range from 0 to 1. MTurk survey is from May 2010. None of the differences between coefficients across studies are statistically significant (see the Supplementary data).

MTurk: Internal Validity

- The possibility that subjects violate treatment assignment by participating in a given task more than once
- Subject inattentiveness, in which case some subsets of the sample do not attend to the experimental stimuli and are effectively not treated

MTurk: Internal Validity Evidence

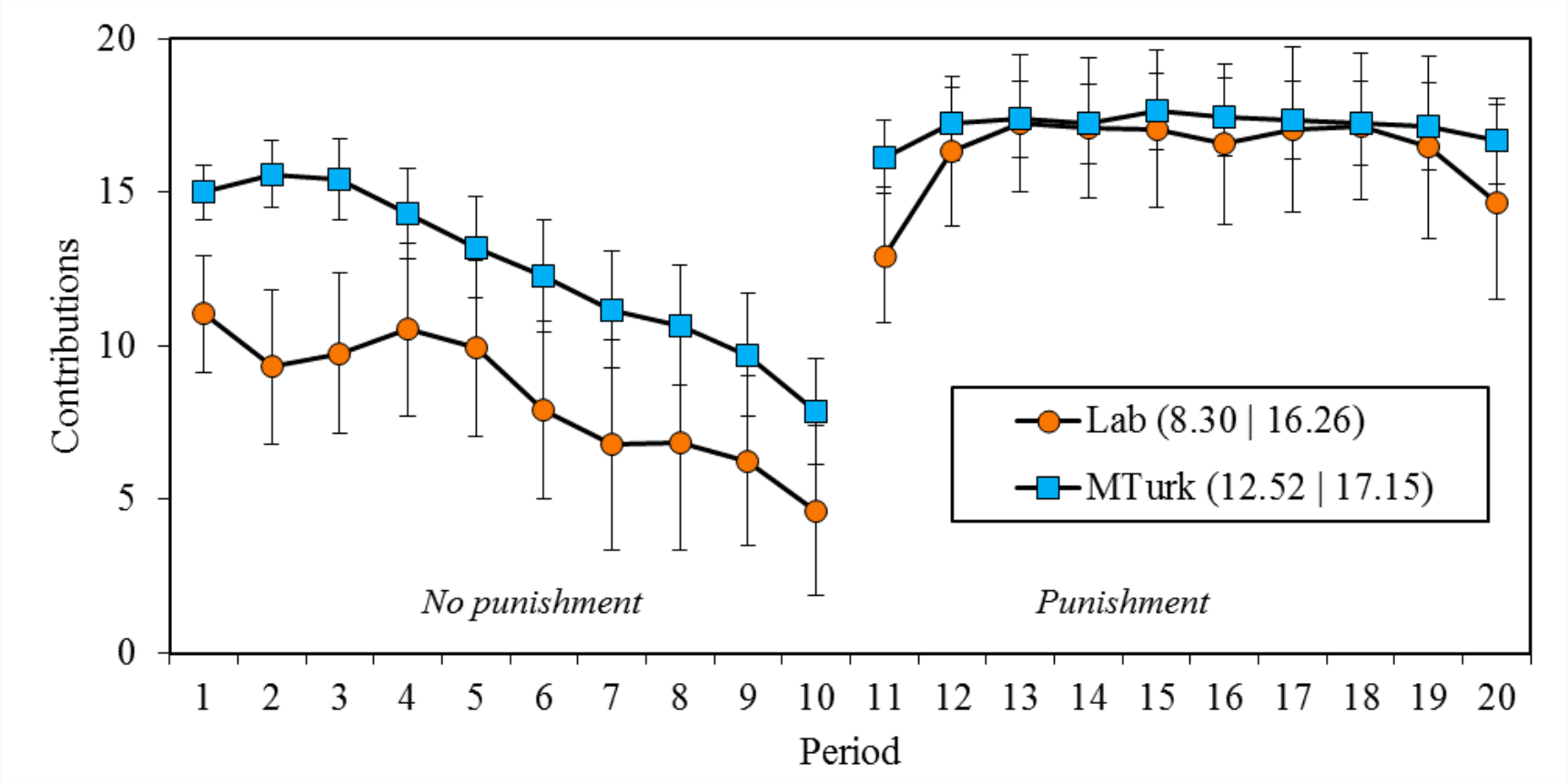
- Participating in experiments multiple times
- 7 out of 551 IP addresses (2.4%) produced 2 responses
- Attention, demand and subject motivation
- Questions to confirm attentiveness - 60% recalled political office of person described in vignette story
- 95% prior approval rating

Conclusions

- The demographic characteristics of domestic MTurk users are more representative and diverse than the corresponding student and convenience samples typically used in experimental studies
- They replicate experimental studies previously conducted using convenience and nationally representative samples, finding that the estimates of average treatment effects are similar in the MTurk and original samples
- They find that potential limitations to using MTurk to recruit subjects and conduct research, in particular, concerns about heterogeneous treatment effects, subject attentiveness, and the prevalence of habitual survey taker - are not large problems in practice

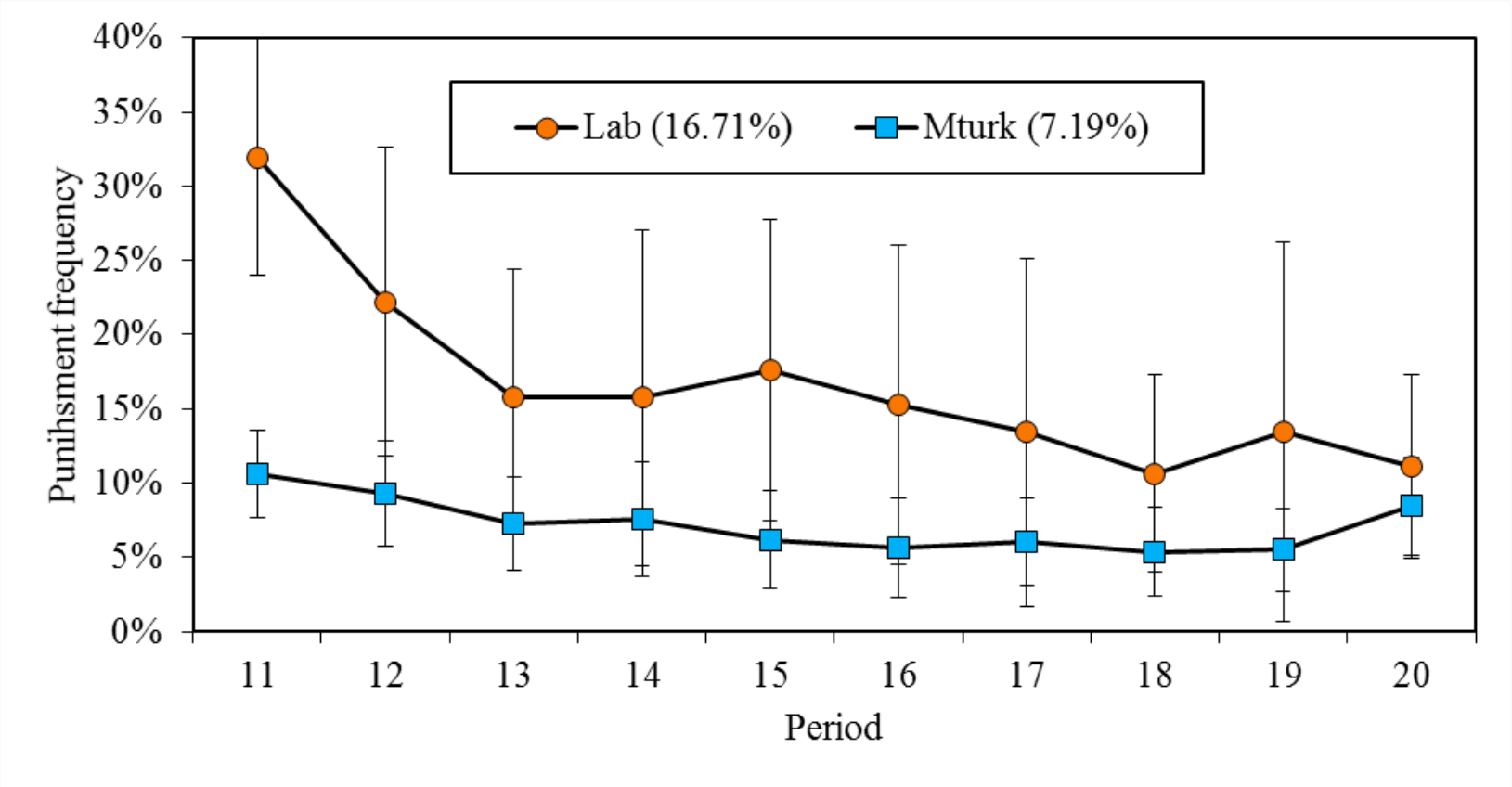
MTurk: Dynamic External Validity

- Arechar et al 2017
- Lab Experiment versus MTurk "Hot" Experiment
- Public Goods Game


20
15
10
5
0
Contributions
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Period
No punishment
Punishment
Lab (8.30 | 16.26)
MTurk (12.52 | 17.15)

	Contributions to the public good					
	No punishment			Punishment		
	Laboratory	MTurk	Pooled	Laboratory	MTurk	Pooled
Period	-0.900***	-1.074***	-1.037***	1.139	0.514*	0.682**
	(0.309)	(0.187)	(0.160)	(0.710)	(0.289)	(0.282)
Final period	-3.400	-2.292**	-2.512***	-10.203**	-4.184**	-5.795***
	(2.253)	(0.958)	(0.881)	(4.881)	(1.688)	(1.797)
MTurk			5.421***			4.193
			(1.867)			(4.904)
Constant	10.470***	17.046***	11.402***	25.980***	35.272***	29.601***
	(1.592)	(0.624)	(1.650)	(3.898)	(3.792)	(4.232)
N	720	2480	3200	720	2480	3200
F	8.75	33.66	34.45	2.19	3.12	3.75

Table 2 Cooperation dynamics. Tobit estimation with left-censoring for 'No punishment' and right-censoring for 'Punishment'. 'Period' is period number; 'Final period' is a dummy for last period; 'MTurk' is a dummy for the MTurk sample. Robust standard errors clustered on groups; * $p<0.1$, ** $p<0.05$, *** $p<0.01$.


40%
35%
30%
25%
20%
15%
10%
5%
0%
Punihsment frequency
Lab (16.71%)
Mturk (7.19%)
11
12
13
14
15
16
17
18
19
20
Period

	Participant's drop out in period t (0=no; 1=yes)				
	Pooled data			*Without punishment*	*With punishment*
	(1)	(2)	(3)	(4)	(5)
Punishment available	0.056	0.362	0.107		
	(0.598)	(0.612)	(0.611)		
Period	-0.093*	-0.118**	-0.094*	-0.265***	-0.150*
	(0.051)	(0.053)	(0.053)	(0.080)	(0.082)
First period	2.484***	2.375***	2.554***		
	(0.377)	(0.376)	(0.382)		
Earnings		-0.002	0.011		
		(0.143)	(0.143)		
Group member(s) dropped out in previous period			1.890***	3.636***	2.034***
			(0.382)	(0.394)	(0.573)
Relative average contribution				0.010	-0.082
				(0.033)	(0.053)
Relative average punishment received					-0.104
					(0.092)
Relative average punishment given					-0.204
					(0.214)
Constant	-4.064***	-3.979***	-4.220***	-3.519***	-4.282***
	(0.317)	(0.318)	(0.328)	(0.466)	(0.472)
N	8334	8327	8332	3539	3527
AIC	893.56	877.20	860.27	325.98	302.30

Dynamic Multi-modes

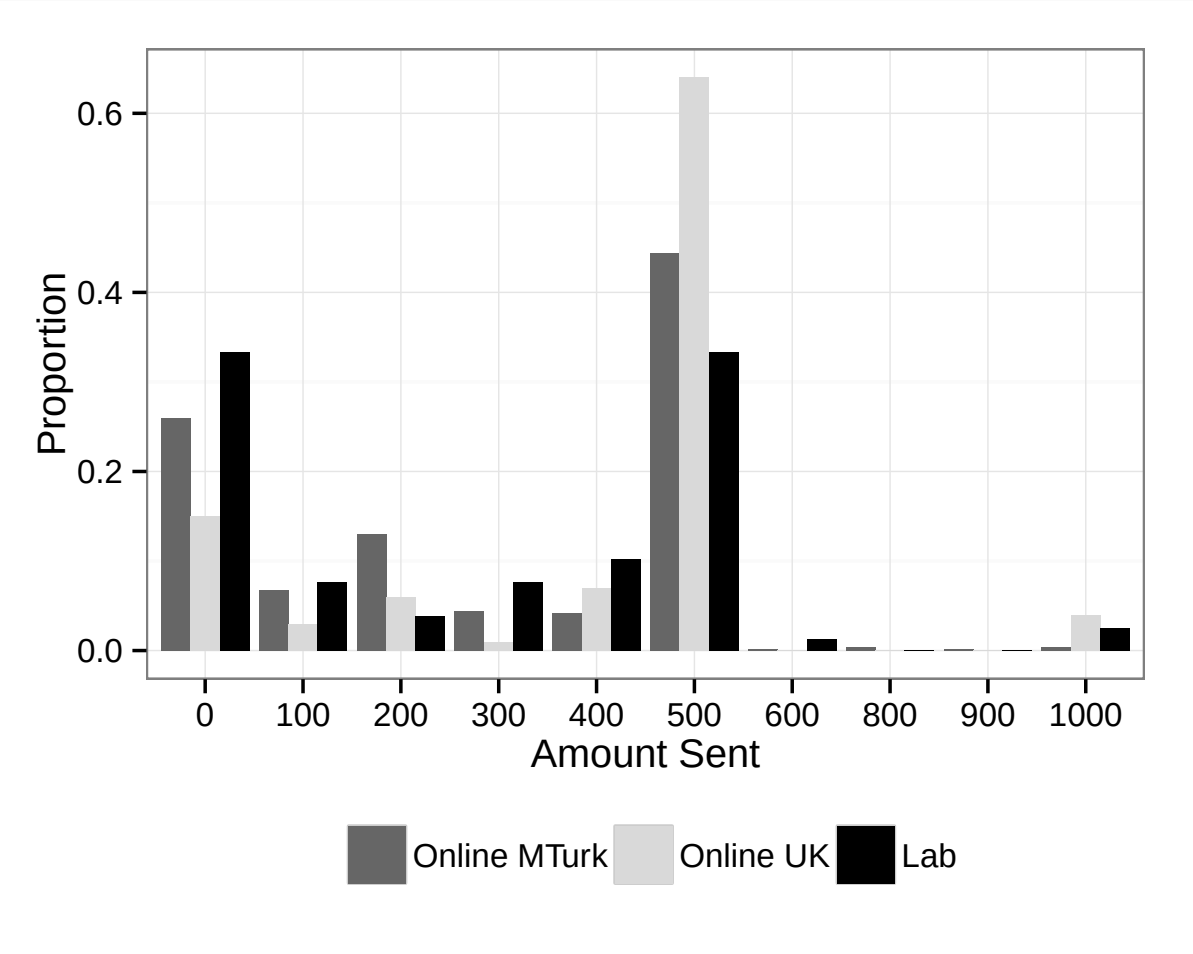
- Duch et al 2017
- Modes: Online versus Lab Experiment
- Subject Pools:
- Public Goods Game + RET

Table 1: Summary of Mode and Subject Pool Effects

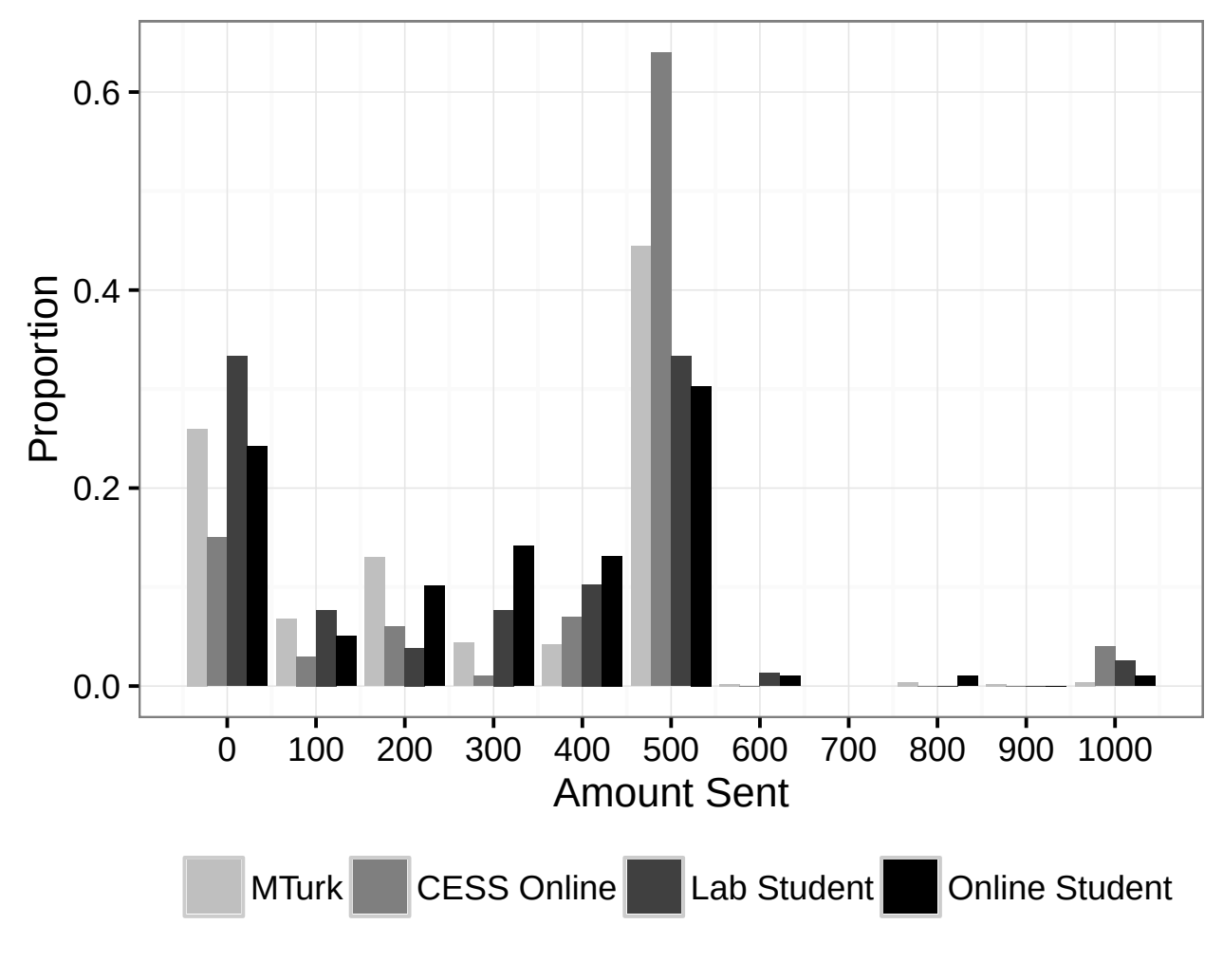
Mode	Lab Subject Pool	Online Subject Pool	Subject Pool Characteristics
Lab BD	Lab-BD	NA	NA
Online Non-synchronic	Online Lab-BD	CESS Online-BD	Estimated
Online Non-synchronic (M-Turk)	NA	MTurk-BD	NA
Lab DS	Lab-DS	NA	NA
Online Synchronic (M-Turk)	NA	MTurk-DS	NA
Online Synchronic	Online Lab-DS	NA	Estimated
Mode Effect	Estimated	Estimated	

Note Mode and Subject Pool Effects.

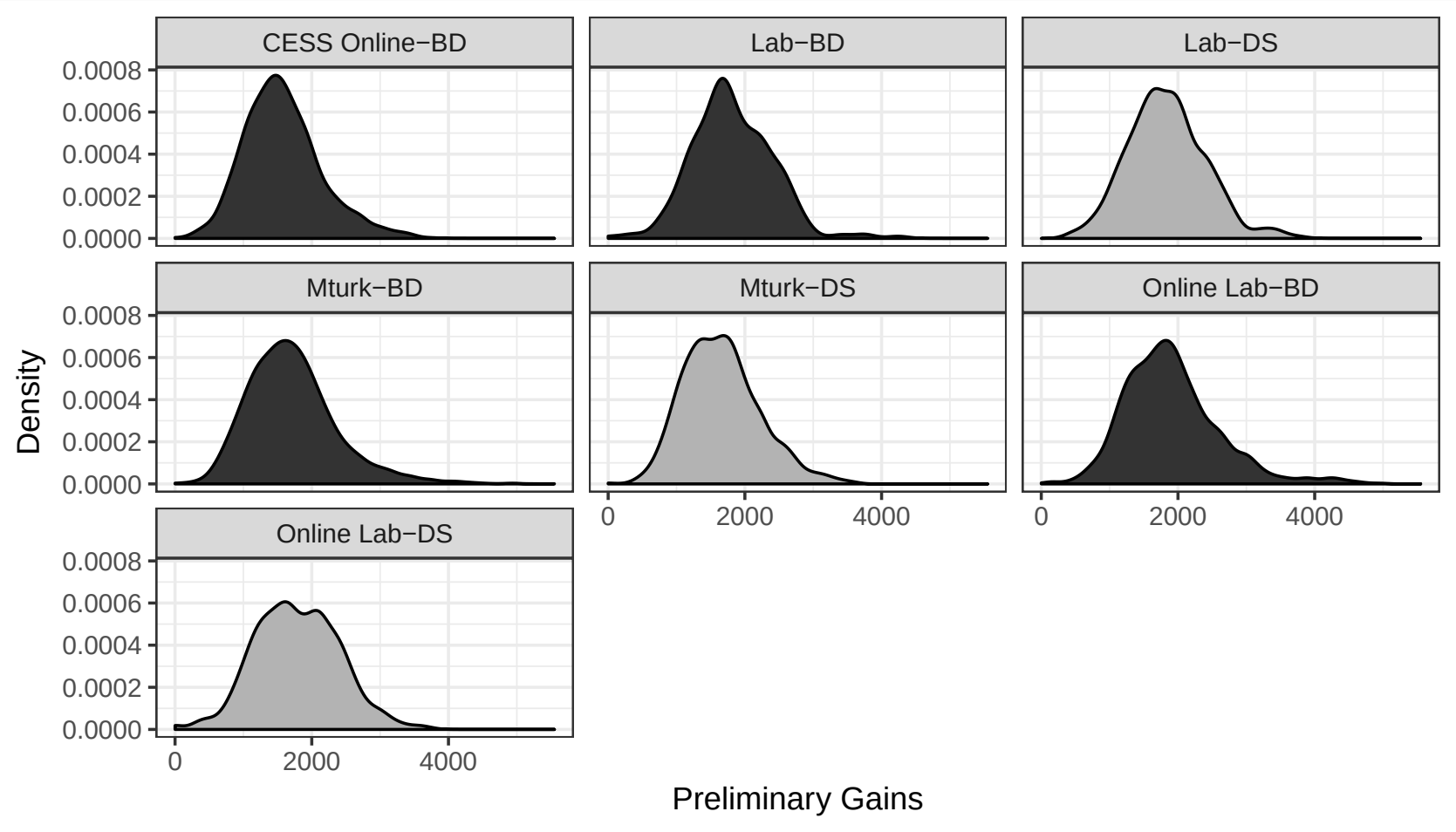
Duch & Solaz


0.6
0.4
0.2
0.0
Proportion
0
100
200
300
400
500
600
800
900
1000
Amount Sent
Online MTurk
Online UK
Lab

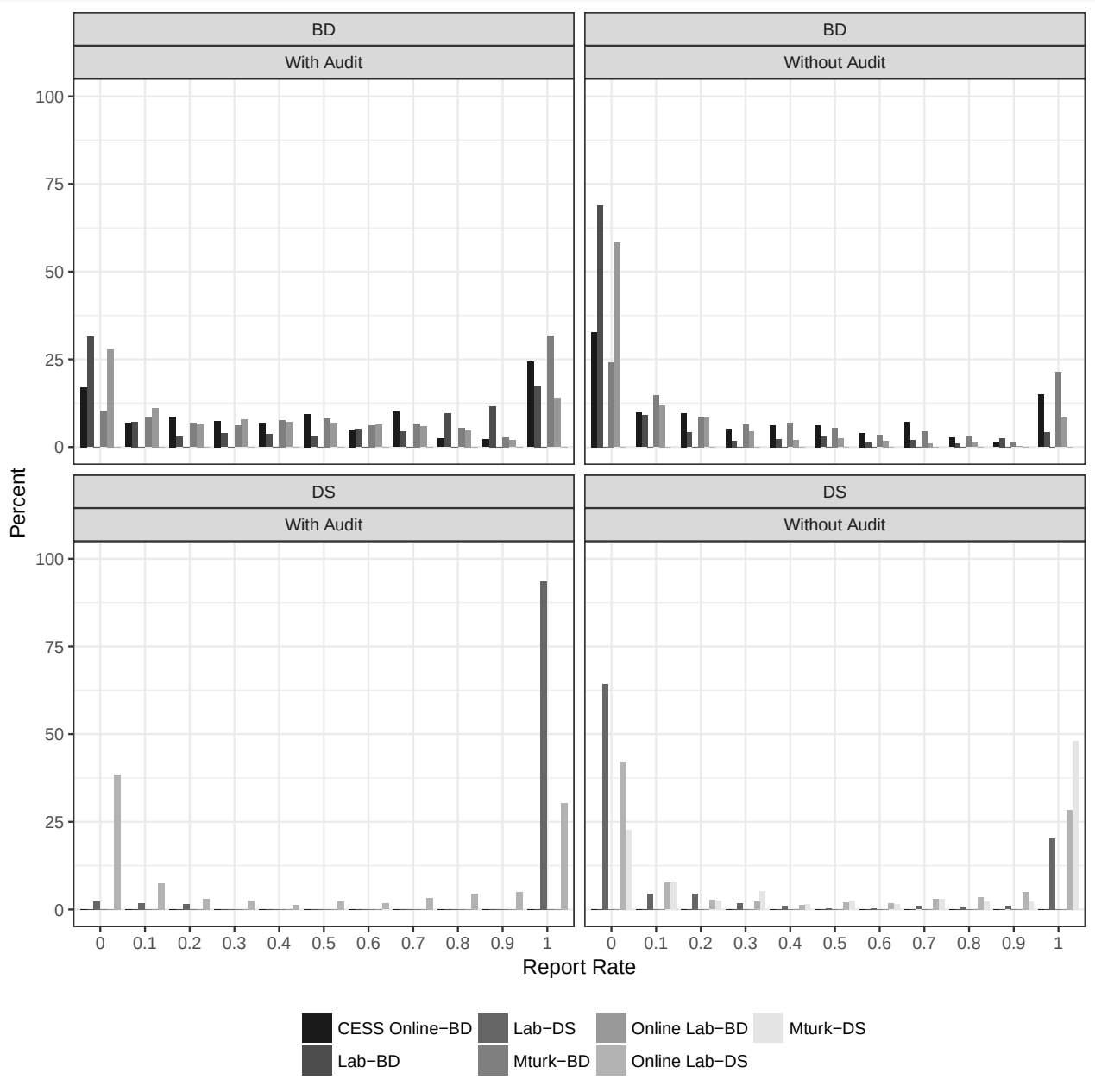
Duch & Beramindi



Real Effort Task



Report Rate


BD
With Audit
BD
Without Audit
DS
With Audit
DS
Without Audit
Percent
100
75
50
25
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Report Rate
CESS Online−BD
Lab−BD
Lab−DS
Mturk−BD
Online Lab−BD
Online Lab−DS
Mturk−DS

Multivariate

	CESS Online-BD (1)	MTurk BD (2)	Lab BD (3)	Online Lab-BD (4)	MTurk DS (5)	Lab DS (6)	Online Lab-DS (7)
# of Additions	-0.001	-0.006***	-0.023***	-0.016***	-0.013***	-0.015***	-0.011***
	(0.006)	(0.002)	(0.005)	(0.004)	(0.003)	(0.002)	(0.002)
Middle Tax Bracket	0.046	0.033*	-0.069*	0.009		0.028	
	(0.043)	(0.017)	(0.036)	(0.030		(0.021)	
Low Tax Bracket	0.083	0.062***	-0.054	0.073*	-0.062***	-0.088***	0.130***
	(0.057)	(0.021)	(0.047)	(0.037)	(0.022)	(0.021)	(0.018)
No Audit	-0.184***	-0.178***	-0.318***	-0.205***		-0.705***	-0.038**
	(0.027)	(0.011)	(0.024)	(0.020)		(0.016)	(0.018)
Dictator Game Giving	0.318***	0.311***	0.271***	0.158***	0.429***	0.323***	0.452***
	(0.066)	(0.026)	(0.050)	(0.051)	(0.044)	(0.043)	(0.040)
Integrity Score	-0.116	-0.366***	-0.220*	-0.001***	-0.010	0.115*	-0.176***
	(0.113)	(0.044)	(0.117)	(0.086)	(0.066)	(0.059)	(0.059)
Risk Preference	-0.054	-0.048*	-0.200***	-0.196***	0.107**	0.087*	0.228***
	(0.055)	(0.026)	(0.070)	(0.062)	(0.053)	(0.045)	(0.047)
Constant	0.402***	0.637***	0.916***	0.609***	0.601***	0.944***	0.508***
	(0.113)	(0.039)	(0.115)	(0.084)	(0.045)	(0.051)	(0.042)
Observations	725	4,460	728	968	1,632	1,440	2,249
R^2	0.116	0.135	0.287	0.218	0.075	0.577	0.113
Adjusted R^2	0.107	0.134	0.281	0.212	0.072	0.575	0.111