RCTs: Advanced Topics

NYU Wagner

Rajeev Dehejia

Fisher randomization test

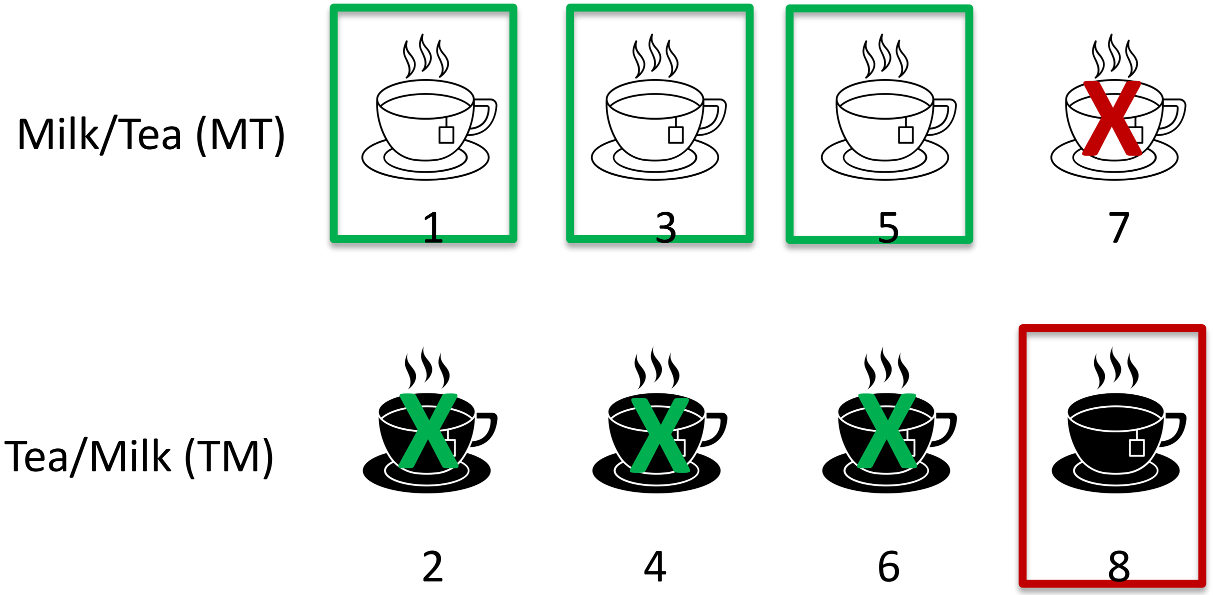
- In this model, the **only source of uncertainty is from randomizing the treatment**. This implies a simple, powerful, exact form of hypothesis testing.
- Example: Dr. Muriel Bristol* drinks a cup of tea and she claims she can tell whether milk was poured first or tea.
- In a randomized trial, where there were 4 cups of each type, she is asked to identify “4 of a kind” (that is, she is allowed to taste all 8, but is asked to choose the 4 that have been prepared the same way) **she gets 3 of the 4 selections right. How likely is it that she has no skill?**

**The famous tea example came from a real-life encounter between Sir Ronald Fisher and Dr. Muriel Bristol, a phyc ist (studied algae), who claimed she could tell the method in which tea was prepared.*

The Tea Experiment

***Fisher**: "Please identify the four cups in this randomized group of eight cups in which the milk was poured before the tea"*

***Bristol**: "The milk was poured before the tea in cups 1, 3, 5, and 8."*



Milk/Tea (MT)

1 3 5 7

Tea/Milk (TM)

2 4 6 8

Research question: How likely is it that Dr. Bristol chose the number of correct selections due to chance alone? (How likely is it the H_0 is true?)

Testing in small samples: Fisher’s exact test

- Test of differences in means with large *N*:

$$H_0 : E[Y_1] = E[Y_0], \qquad H_1 : E[Y_1] \neq E[Y_0]$$

- Fisher’s Exact Test with a small *N*:

$$H_0 : Y_1 = Y_0\ , \qquad H_1 : Y_1 \neq Y_0 \qquad \text{(sharp null)}$$

- Let Ω be the set of all possible randomization realizations.
- We only observe the outcomes, Y_i, for one realization of the experiment. We calculate $\hat{\alpha} = \bar{Y}_1 - \bar{Y}_0$.
- Under the sharp null hypothesis we can calculate the value that the difference of means would have taken under any other realization, $\hat{\alpha}(\omega)$, for $\omega \in \Omega$.

Fisher randomization test

- H_0: No treatment effect (she has no ability, so she picks at random)
- Then what is the chance of picking this grouping **by chance**?
- Hypergeometric distribution
- Pr($\underline{>}$3 from Milk/Tea group, $\underline{<}$1 Tea/Milk)

= Pr(3MT,1TM)+Pr(4MT,0TM)=17/70

How many ways are there to choose 3 from the MT group and 1 from the TM group?

$$\frac{\boxed{\binom{4}{3}\binom{4}{1}}}{\boxed{\binom{8}{4}}} = \frac{\frac{4!}{3!1!}\ \frac{4!}{1!3!}}{\frac{8!}{4!4!}} = \frac{16}{70} \quad + \quad \frac{\binom{4}{4}\binom{4}{0}}{\binom{8}{4}} = \frac{\frac{4!}{4!0!}\ \frac{0!}{4!0!}}{\frac{8!}{4!4!}} = \frac{1}{70}$$

How many ways are there to choose 4 cups to put the tea in first?

$$P\,(\text{“}proportion\ of\ correct\ selections\text{”})$$

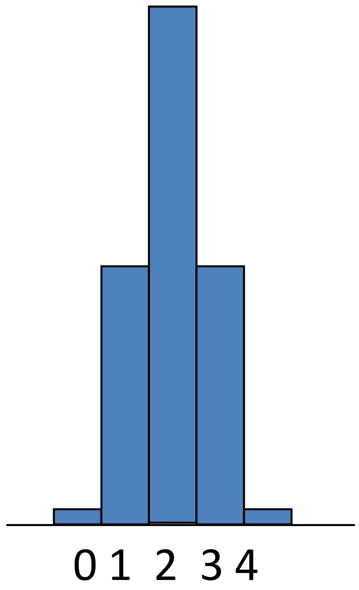
$$= \frac{'number\ of\ ways\ to\ select\ 3MT{:}\,1\,TM'}{'number\ of\ ways\ to\ guess'}$$

$$P\,(\text{“}all\ correct\text{”}) = \frac{1}{'number\ of\ ways\ to\ guess'}$$

$$= \frac{1}{\binom{8}{4}} = \frac{1}{70}$$

Fisher randomization test or permutation test

- Full distribution looks like this →
- More generally you can simulate the distribution under the null when the exact distribution is not available.



Simulating Fisher’s test

Unit	Outcome under treatment	Outcome under control	$H_{0:}\ \tau=0$
1	$\boxed{Y_{11}}$	$\cancel{Y_{01}}$	$Y_{11}=Y_{01}$
2	$\cancel{Y_{12}}$	$\boxed{Y_{02}}$	$Y_{12}=Y_{02}$
3	$\boxed{Y_{13}}$	$\cancel{Y_{03}}$	$Y_{13}=Y_{03}$
4	$\boxed{Y_{14}}$	$\cancel{Y_{04}}$	$Y_{14}=Y_{04}$
N	$\cancel{Y_{1N}}$	$\boxed{Y_{0N}}$	$Y_{1N}=Y_{0N}$
			Full data is observed

Simulating Fisher's test

- For any given assignment to treatment and observed difference in mean, null hypothesis is no effect. **H_0: any difference in mean is due to chance, not the treatment**

- If so, we observe the full data, and can simulate the full distribution of mean effects: randomly reassign units to treatment and control; **$Y_{1i}=Y_{0i}$ under the null, so we can compute new mean difference; simulate many times**. This is the key concept undergirding randomization as the gold standard of "overcoming" the fundamental problem

Another example

Suppose that we assign 4 individuals out of 8 to the treatment: $\hat{\alpha} = \bar{Y}_1 - \bar{Y}_0$

Y_i	12	4	6	10	6	0	1	1	$\hat{\alpha}$
T_i	1	1	1	1	0	0	0	0	= 6
									(ω)
$\omega = 1$	1	1	1	1	0	0	0	0	6
$\omega = 2$	1	1	1	0	1	0	0	0	4
$\omega = 3$	1	1	1	0	0	1	0	0	1
$\omega = 4$	1	1	1	0	0	0	1	0	1.5
				. . .					
$\omega = 70$	0	0	0	0	1	1	1	1	-6

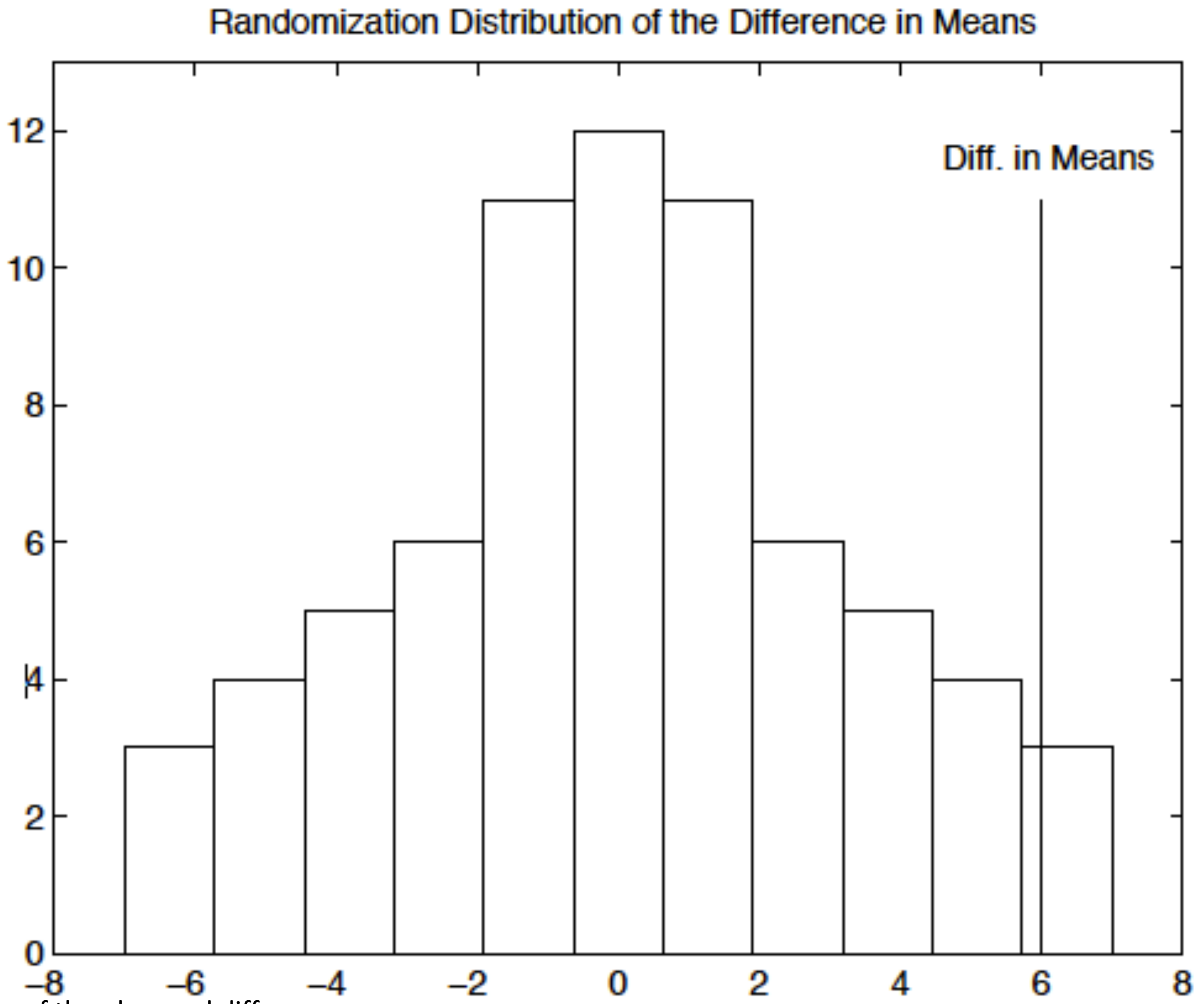
- The randomization distribution of $\hat{\alpha}$ (under the sharp null hypothesis) is

$$\Pr\left(\hat{\alpha} \le z\right) = \frac{1}{70} \sum_{\omega\in\Omega} 1\left\{\hat{\alpha}\left(\omega\right) \le z\right\}$$

•Now, find $\bar{z} = \inf\left\{z : \mathrm{P}\Big|\left(|\hat{\alpha}| > z\right) \le 0.05\right\}$

- Reject the null hypothesis $H_0 : Y_{1i} - Y_{0i} = 0$ for all i, against the alternative hypothesis $H_0 : Y_{1i} - Y_{0i} \neq 0$ for some i, at the 5% significance level if $\left|\hat{\alpha}\right| > \bar{z}$

Testing in small samples: Fisher’s exact test



Probability of the observed difference in means given all the potential differences in means (with randomization)

$$\Pr(|\hat{\alpha}(\omega)| \geq 6) = 0.0857$$

Experimental design: relative sample sizes for fixed *N*

Suppose that you have *N* experimental subjects and you have to decide how many will be in the treatment group and how many in the control group. We know that:

$$\bar{Y}_1 - \bar{Y}_0 \sim \left(\mu_1 - \mu_0, \frac{\sigma_1^2}{N_1} + \frac{\sigma_0^2}{N_0} \right).$$

Problem

Choose N_1 and N_0, such that $N_1 + N_0 = N$, to minimize the variance of the estimator of the average treatment effect.

The variance of $\bar{Y}_1 - \bar{Y}_0$ *is:*

$$\operatorname{var}\left(\bar{Y}_1 - \bar{Y}_0\right) = \frac{\sigma_1^2}{pN} + \frac{\sigma_0^2}{(1-p)N}$$

where $p = N_1/N$ is the proportion of treated in the sample.

Experimental design: relative sample sizes for fixed *N*

Find the value *p** that minimizes $\text{var}\left(\bar{Y}_1-\bar{Y}_0\right)$:

$$-\frac{\sigma_1^{2}}{p^{*2}N}+\frac{\sigma_0^{2}}{(1-p^*)^{2}N}=0.$$

Therefore:

$$\frac{1-p^*}{p^*}=\frac{\sigma_0}{\sigma_1},$$

and

$$p^*=\frac{\sigma_1}{\sigma_1+\sigma_0}=\frac{1}{1+\sigma_0/\sigma_1}$$

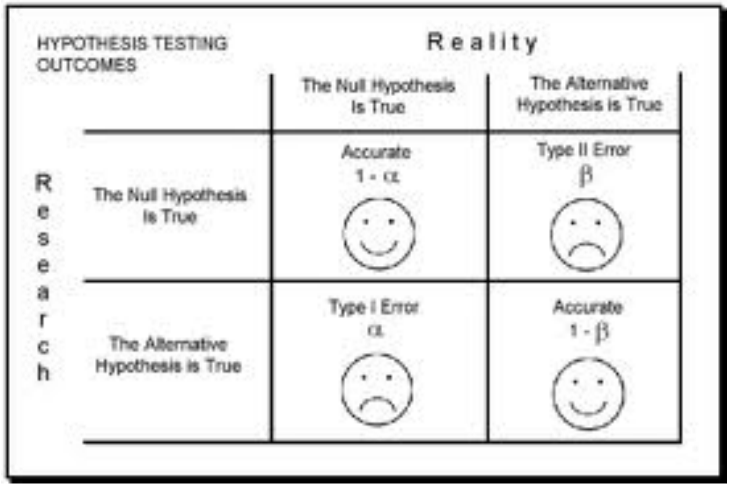
A “rule of thumb” for the case $\sigma_1 \approx \sigma_0$ is $p^* = 0.5$

For practical reasons it is sometimes better to choose unequal sample sizes (even if $\sigma_1 \approx \sigma_0$)

Experimental design: power calculations to choose *N*

- Recall that for a statistical test:
  - **Type I error**: Rejecting the null if the null is true
  - **Type II error**: Not rejecting the null if the null is false
- Size of a test is the probability of a type I error. Usually 0.05
- Power of a test is one minus the probability of type II error, i.e. the probability of rejecting the null if the null is false. Usually 0.8

Determine power of statistical test



HYPOTHESIS TESTING OUTCOMES		Reality	
		The Null Hypothesis Is True	The Alternative Hypothesis is True
Research	The Null Hypothesis Is True	Accurate $1 - \alpha$ ☺	Type II Error β ☹
	The Alternative Hypothesis is True	Type I Error α ☹	Accurate $1 - \beta$ ☺

- Type I Error
  - Common choice 5% level of significance or Confidence = 1-α=0.95
  - I.e., you are willing to accept a 5% chance of false positive.
- Type II Error:
  - Common choice β=0.2 or Power=1-β=0.8.
  - I.e., you are willing to accept 20% chance of saying there is no effect when there is, or a test that has 80% power (correctly categorizes a program as having no effect when that is true).
  - As you decrease Type I Error, increase Type II Error
- Seriousness of Type I vs Type II Error
  - Type I Error commonly seen as 4x as serious as Type II Error
    - 5/100 vs 20/100

Experimental design: what affects power?

- The precision (i.e., 1/variance) of *within* sample measures
  - Inherent variability of the outcome we are interested in.
  - **Statistical power increases with the sample size** in standard sampling.
- The size of the difference we are trying to detect.
  - Precision to $1000 is a lot if we are trying to detect a 10k difference, little if we are hoping to detect a $500 difference.
  - MDES " δ “: Minimum detectable effect size; you can calculate this based on alpha, beta, and N, but the interpretation is conceptual/relative to the outcome you’re evaluation; is the MDES ”big enough” to be meaningful?
- But when is a sample “large enough”?
- We want to find *N* such that we will be able to detect an average treatment effect of size δ or larger with high probability.

Power and size of a test

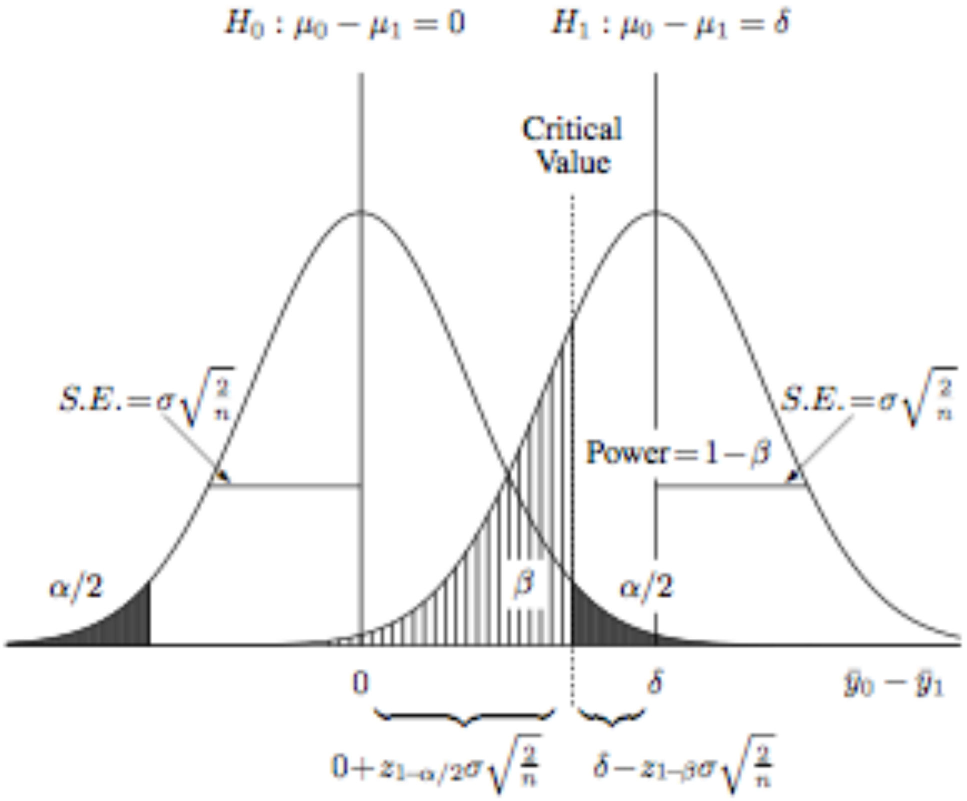

$H_0 : \mu_0 - \mu_1 = 0$
$H_1 : \mu_0 - \mu_1 = \delta$
Critical Value
$S.E. = \sigma\sqrt{\frac{2}{n}}$
$S.E. = \sigma\sqrt{\frac{2}{n}}$
$\text{Power} = 1 - \beta$
$\alpha/2$
β
$\alpha/2$
0
δ
$\bar{y}_0 - \bar{y}_1$
$0 + z_{1-\alpha/2}\sigma\sqrt{\frac{2}{n}}$
$\delta - z_{1-\beta}\sigma\sqrt{\frac{2}{n}}$

Fig. 2.1 Sampling model for two independent sample case. Two-sided alternative, equal variances under null and alternative hypotheses.

Simplest case of power analysis

- Key question: how big do the treatment and control effects have to be?
- Recall standard error of a mean:

$$\frac{\sum_i (x_i - \bar{x})^2}{(n-1)n}$$

- Declines with n, which you get to choose. Suppose you want to test a hypothesis at a scale of (*p*)*s*.

$$\frac{p \cdot s}{{}^{s}\!/\!_{\sqrt{n}}} > t_{\alpha/2}$$

$$n > \left(t_{\alpha/2} / p\right)^2$$

Then need sample size of 200+ for significance at 1 %.

Power calculations

- More generally power depends on effect size, sample size, the significance criterion, and the amount of variation in the data.

Power calculations with equal and known variances

Suppose that $Y_1 \sim (\mu_1, \sigma^2)$, $Y_0 \sim (\mu_0, \sigma^2)$

Assume also that $p = 0.5$, so $N_0 = N_1 = N/2$.

Let effect size $\delta = \mu_1 - \mu_0$.

Then, for the t-statistic of equality of means:

$$\frac{\bar{Y}_1 - \bar{Y}_0 - \delta}{\sqrt{\dfrac{\sigma_1^2}{N_1} + \dfrac{\sigma_0^2}{N_0}}} = \frac{\bar{Y}_1 - \bar{Y}_0 - \delta}{\sqrt{\dfrac{2\sigma^2}{N} + \dfrac{2\sigma^2}{N}}} = \frac{\bar{Y}_1 - \bar{Y}_0 - \delta}{2\sigma / \sqrt{N}} \sim N\left(0,1\right),$$

if the sample is large enough. Therefore:

$$t = \frac{\bar{Y}_1 - \bar{Y}_0}{\sqrt{\dfrac{\sigma_1^2}{N_1} + \dfrac{\sigma_0^2}{N_0}}} \sim \left(\frac{\delta\sqrt{N}}{2\sigma}, 1\right)$$

Power calculations with equal and known variances

The probability of rejecting the null $\mu_1 - \mu_0 = 0$ is:

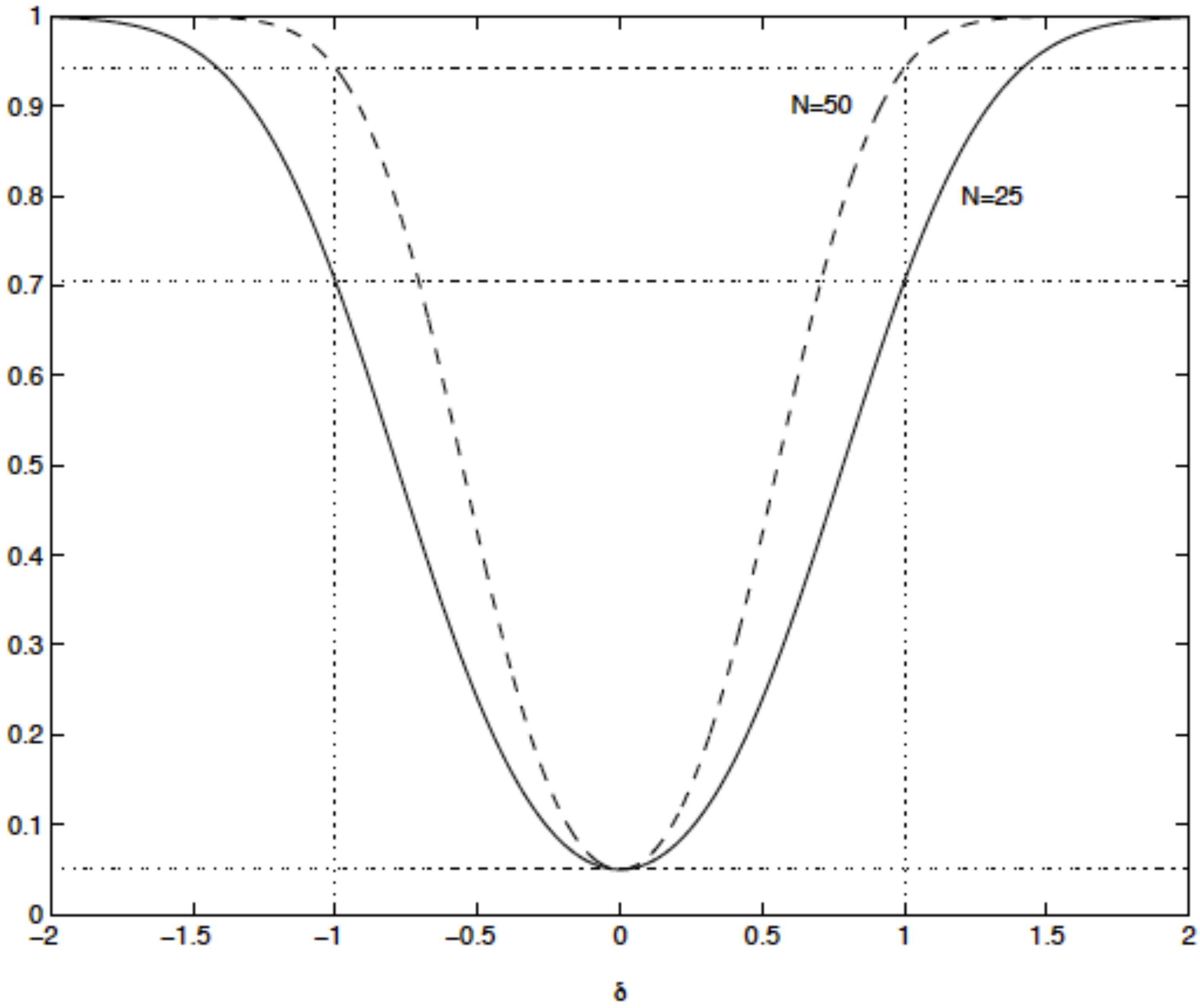
$$\Pr\left(|\mathrm{t}| > 1.96\right) \quad = \quad \Pr\left(t < -1.96\right) + \Pr\left(t > 1.96\right)$$

$$= \quad \Pr\left(t - \frac{\delta\sqrt{n}}{2\sigma} < -1.96 - \frac{\delta\sqrt{n}}{2\sigma}\right)$$

$$+ \quad \Pr\left(t - \frac{\delta\sqrt{n}}{2\sigma} > 1.96 - \frac{\delta\sqrt{n}}{2\sigma}\right)$$

$$= \quad \Phi\left(-1.96 - \frac{\delta\sqrt{n}}{2\sigma}\right) + \left(1 - \Phi\left(1.96 - \frac{\delta\sqrt{n}}{2\sigma}\right)\right)$$

Power functions for N = 25, N = 50, and σ^2 = 1



General formula for the power function ($p \neq 0.5$, $\sigma_0^2 \neq \sigma_1^2$)

Pr (reject $\mu_1 - \mu_0 = 0 \mid \mu_1 - \mu_0 = \delta$)

$$= \Phi\left(-1.96 - \delta \Big/ \sqrt{\frac{\sigma_1^2}{pN} + \frac{\sigma_0^2}{(1-p)N}}\right)$$

$$+ \left(1 - \Phi\left(1.96 - \delta \Big/ \sqrt{\frac{\sigma_1^2}{pN} + \frac{\sigma_0^2}{(1-p)N}}\right)\right).$$

To choose N we need to specify:

1. δ: minimum detectable magnitude of treatment effect
2. Power value (usually 0.80 or higher)
3. σ_1^2 and σ_0^2 (usually $\sigma_1^2 = \sigma_0^2$) (e.g., using "benchmarked" or previous measures, might use historical data or similar studies)
4. p: proportion of observations in the treatment group if $\sigma_1 = \sigma_0$, then the power is maximized by $p = 0.5$

Example: power study for GS

- We were running classroom experiment in India on a psychological intervention. Needed to know how many classrooms to include in the experiment.

- We found a paper that used a math test as an outcome. In this paper the standard deviation of the test score was 1. The approximate treatment effect was 0.5 of a standard deviation for a scaled effect size of 0.5.

- *For example: If the avg. classroom score is 75 with a 5 percentage point (pp) standard deviation, we would want a sample size large enough to be able to detect a 2.5 pp treatment effect.*

Quick calculation

- Simplest formula suggests

Per the assumption and proof we discussed, 1.96 is the t statistic for a sample size large enough that we can assume a normal distribution of means

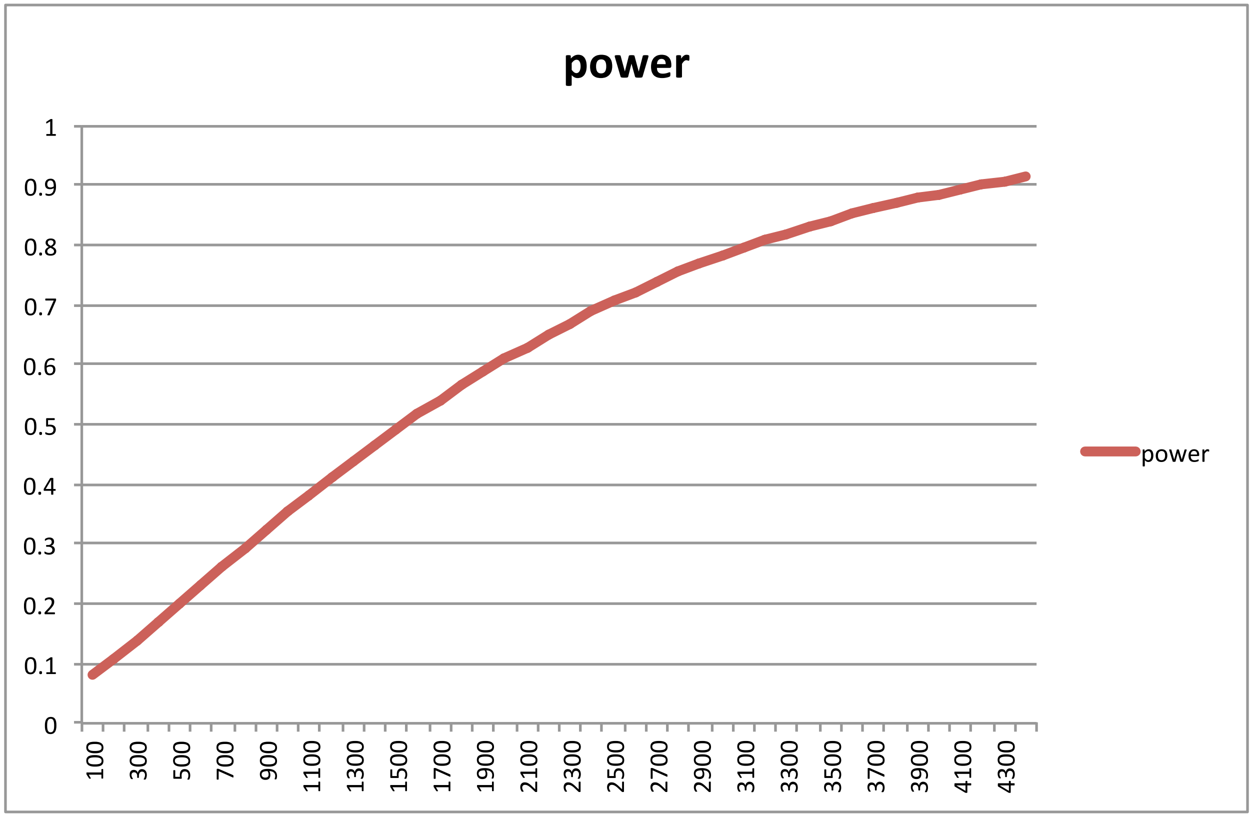
$$n > \left(t_{\alpha/2} / p\right)^2 = \boxed{(1.96\,/}\,0.5)^2 \cong 16$$

or 16 classrooms would suffice. But suppose we want to be more conservative, and suppose we want to observe an effect size as small as 0.1 of standard deviation. Using the example from the last slide, this would be a 0.5 pp effect size.

$$n > \left(t_{\alpha/2} / p\right)^2 = \left(1.96 / 0.1\right)^2 \cong 400$$

But this can be quite inaccurate

- sd=1, delta=0.1, alpha=0.95



Dealing with clustering

- **Clustering reduces power,** so you need a larger sample to achieve the same MDES as a non-clustered experiment
- Now suppose that students are organized in classrooms. No such thing as 4000 individual students. You get them in rooms of 30 each. How many classrooms do you need?
- Question is how correlated is what is going on inside classroom? If each kid is independent, answer is just 4000/30.
- But if each child in a classroom is correlated, you need even more observations.

Dealing with clustering

- To solve this you need a significantly more complicated formula (or Stata).
- But if the intraclass correlation (ICC) is 0.1 of a standard deviation, then it turns out you need about 547 classrooms with 30 students each!
- On the other hand, if effect size is 0.5, then 23 classrooms is enough.

Stata basics: Sample size calcs w/ cluster RCTs

Mean of control group, desired mean of treatment group if MDES = 0.1

```
sampsi 1 1.1,sd1(1) sd2(1) // What sample size do I need with a 0.1 MDES, where the sd of the T and C groups are both 1?
sampclus,obs(30) rho(0.1) // How many clusters do I need with 30 observations per cluster and in ICC of 0.1?
```

```
sampsi 1 1.5,sd1(1) sd2(1)
sampclus,obs(30) rho(0.1)
```

Other useful power commands to review (in the recitation)

```
power oneprop
power onemean
power twoprop
power twomeans
```

Threats to the validity of randomized experiments

- Internal validity: can we estimate treatment effect for our particular sample?
  - Fails when there are systematic differences between treated and controls (other than the treatment itself) that affect the outcome and that we cannot control for

- External validity: can we extrapolate our estimates to other populations?
  - Fails when the treatment effect is different outside the evaluation environment

Most common threats to internal validity

- Failure of randomization
- Non-compliance with experimental protocol
- Attrition

Most common threats to external validity

- Non-representative sample
- Non-representative program
  - The treatment differs in actual implementations
  - Scale effects
  - Actual implementations are not randomized (nor full scale)

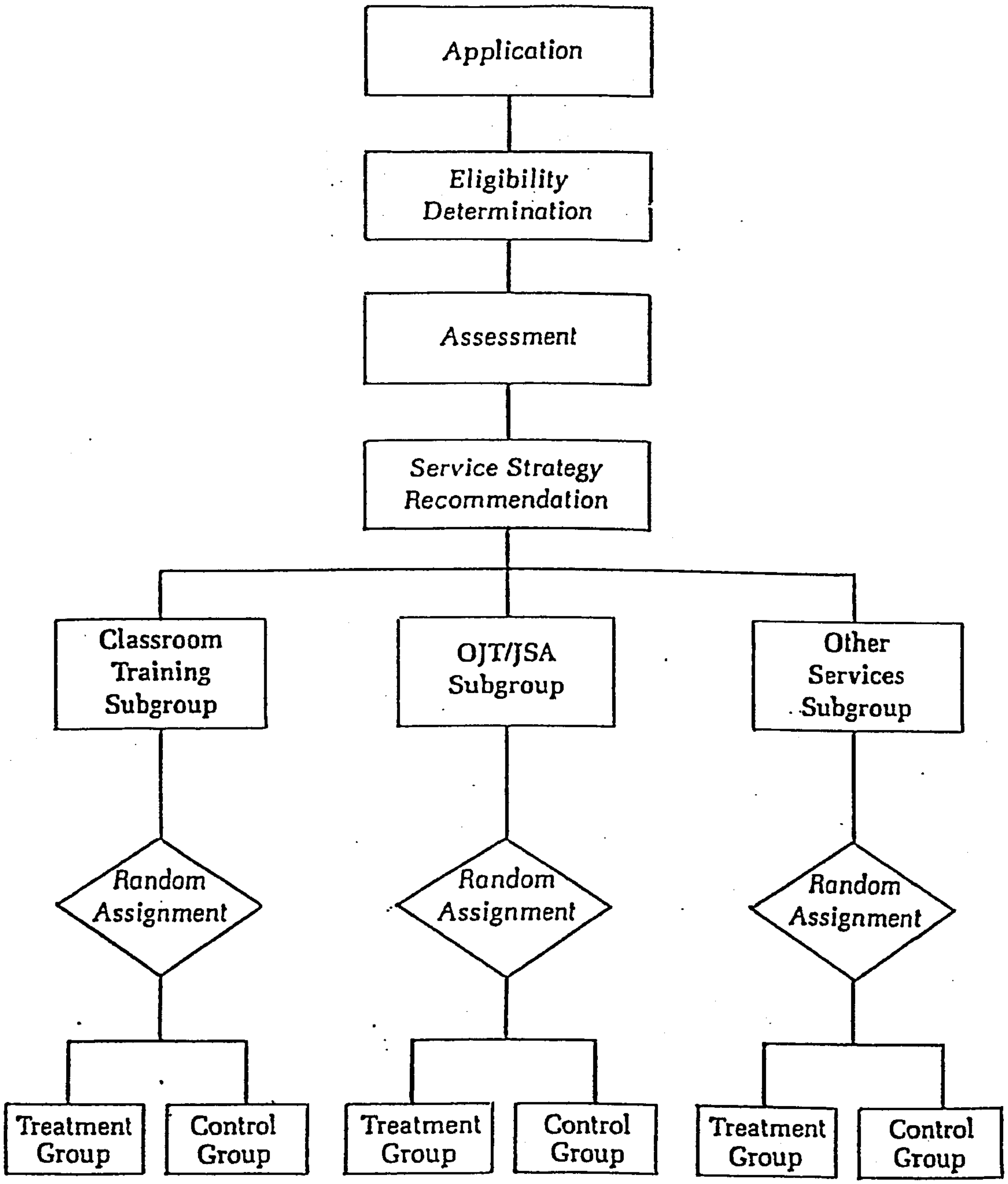
Example: Job Training Partnership Act (JTPA)

- Largest randomized training evaluation ever undertaken in the U.S.; started in 1983 at 649 sites throughout the country
- Sample: Disadvantaged persons in the labor market (previously unemployed or low earnings)
- T: Assignment to one of three general service strategies
  - classroom training in occupational skills
  - on-the-job training and/or job search assistance
  - other services (e.g. probationary employment)
- Y: earnings 30 months following assignment
- X: characteristics measured before assignment (age, gender, previous earnings, race, etc.)

Exhibit 5 Impacts on Total 30-Month Earnings: Assignees and Enrollees, by Target Group

	Mean earnings		*Impact per assignee*		
	Treatment group (1)	*Control group (2)*	*In dollars (3)*	*As a percent of (2)*	*Impact per enrollee in dollars*
Adult women	$ 13,417	$ 12,241	$ 1,176***	9.6%	$ 1,837***
Adult men	19,474	18,496	978*	5.3	1,599*
Female youths	10,241	10,106	135	1.3	210
Male youth non-arrestees	15,786	16,375	-589	-3.6	-868
Male youth arrestees					
Using survey data	14,633	18,842	-4,209**	-22.3	-6,804**
Using scaled UI data	14,148	14,152	-4	0.0	-6

Sources: Estimates based on First and Second Follow-up Survey responses and earnings data from state unemployment insurance (UI) agencies.
Sample sizes: adult women, 6,102; adult men, 5,102; female youths, 2,657; male youth non-arrestees, 1,704; male youth arrestees, 416.
* Statistically significant at the .10 level, ** at the .05 level, *** at the .01 level (two-tailed test).


Application
Eligibility Determination
Assessment
Service Strategy Recommendation
Classroom Training Subgroup
OJT/JSA Subgroup
Other Services Subgroup
Random Assignment
Random Assignment
Random Assignment
Treatment Group
Control Group
Treatment Group
Control Group
Treatment Group
Control Group

Means and Standard Deviations

	Entire	Assignment		Difference
	Sample	Treatment	Control	(t-stat.)
A. Men				
Number of observations	5,102	3,399	1,703	
Treatment				
Training	.42	.62	.01	.61
	[.49]	[.48]	[.11]	(70.34)
Outcome variable				
30 month earnings	19,147	19,520	18,404	1,116
	[19,540]	[19,912]	[18,760]	(1.96)
Baseline Characteristics				
Age	32.91	32.85	33.04	-.19
	[9.46]	[9.46]	[9.45]	(-.67)
High school or GED	.69	.69	.69	-.00
	[.45]	[.45]	[.45]	(-.12)
Married	.35	.36	.34	.02
	[.47]	[.47]	[.46]	(1.64)
Black	.25	.25	.25	.00
	[.44]	[.44]	[.44]	(.04)
Hispanic	.10	.10	.09	.01
	[.30]	[.30]	[.29]	(.70)
Worked less than 13	.40	.40	.40	.00
weeks in past year	[.47]	[.47]	[.47]	(.56)

B. Women				
Number of observations	6,102	4,088	2,014	
Treatment				
Training	.45	.66	.02	.64
	[.50]	[.47]	[.13]	(80.24)
Outcome variable				
30 month earnings	13,029	13,439	12,197	1,242
	[13,415]	[13,614]	[12,964]	(3.46)
Baseline Characteristics				
Age	33.33	33.33	33.35	-.02
	[9.78]	[9.77]	[9.81]	(-.09)
High school or GED	.72	.73	.70	.03
	[.43]	[.43]	[.44]	(2.01)
Married	.22	.22	.21	.01
	[.40]	[.40]	[.39]	(1.55)
Black	.26	.27	.26	.01
	[.44]	[.44]	[.44]	(.95)
Hispanic	.12	.12	.12	-.00
	[.32]	[.32]	[.33]	(-.89)
Worked less than 13 weeks in past year	.52	.52	.52	-.00
	[.47]	[.47]	[.47]	(-.08)
AFDC	.31	.30	.31	-.01
	[.46]	[.46]	[.46]	(-1.03)

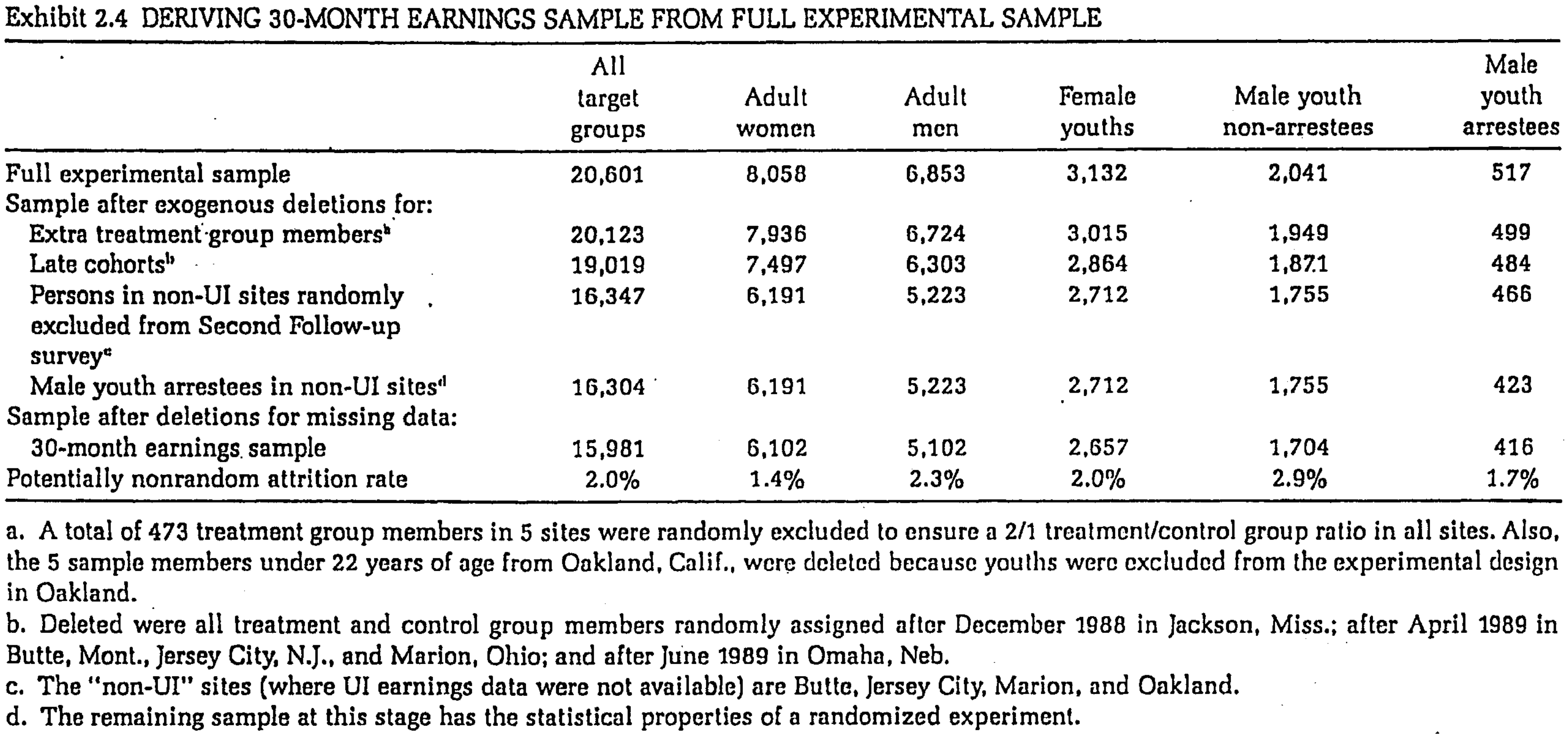


Exhibit 2.4 DERIVING 30-MONTH EARNINGS SAMPLE FROM FULL EXPERIMENTAL SAMPLE

	All target groups	Adult women	Adult men	Female youths	Male youth non-arrestees	Male youth arrestees
Full experimental sample	20,601	8,058	6,853	3,132	2,041	517
Sample after exogenous deletions for:						
Extra treatment group members[a]	20,123	7,936	6,724	3,015	1,949	499
Late cohorts[b]	19,019	7,497	6,303	2,864	1,871	484
Persons in non-UI sites randomly excluded from Second Follow-up survey[c]	16,347	6,191	5,223	2,712	1,755	466
Male youth arrestees in non-UI sites[d]	16,304	6,191	5,223	2,712	1,755	423
Sample after deletions for missing data:						
30-month earnings sample	15,981	6,102	5,102	2,657	1,704	416
Potentially nonrandom attrition rate	2.0%	1.4%	2.3%	2.0%	2.9%	1.7%

a. A total of 473 treatment group members in 5 sites were randomly excluded to ensure a 2/1 treatment/control group ratio in all sites. Also, the 5 sample members under 22 years of age from Oakland, Calif., were deleted because youths were excluded from the experimental design in Oakland.

b. Deleted were all treatment and control group members randomly assigned after December 1988 in Jackson, Miss.; after April 1989 in Butte, Mont., Jersey City, N.J., and Marion, Ohio; and after June 1989 in Omaha, Neb.

c. The "non-UI" sites (where UI earnings data were not available) are Butte, Jersey City, Marion, and Oakland.

d. The remaining sample at this stage has the statistical properties of a randomized experiment.

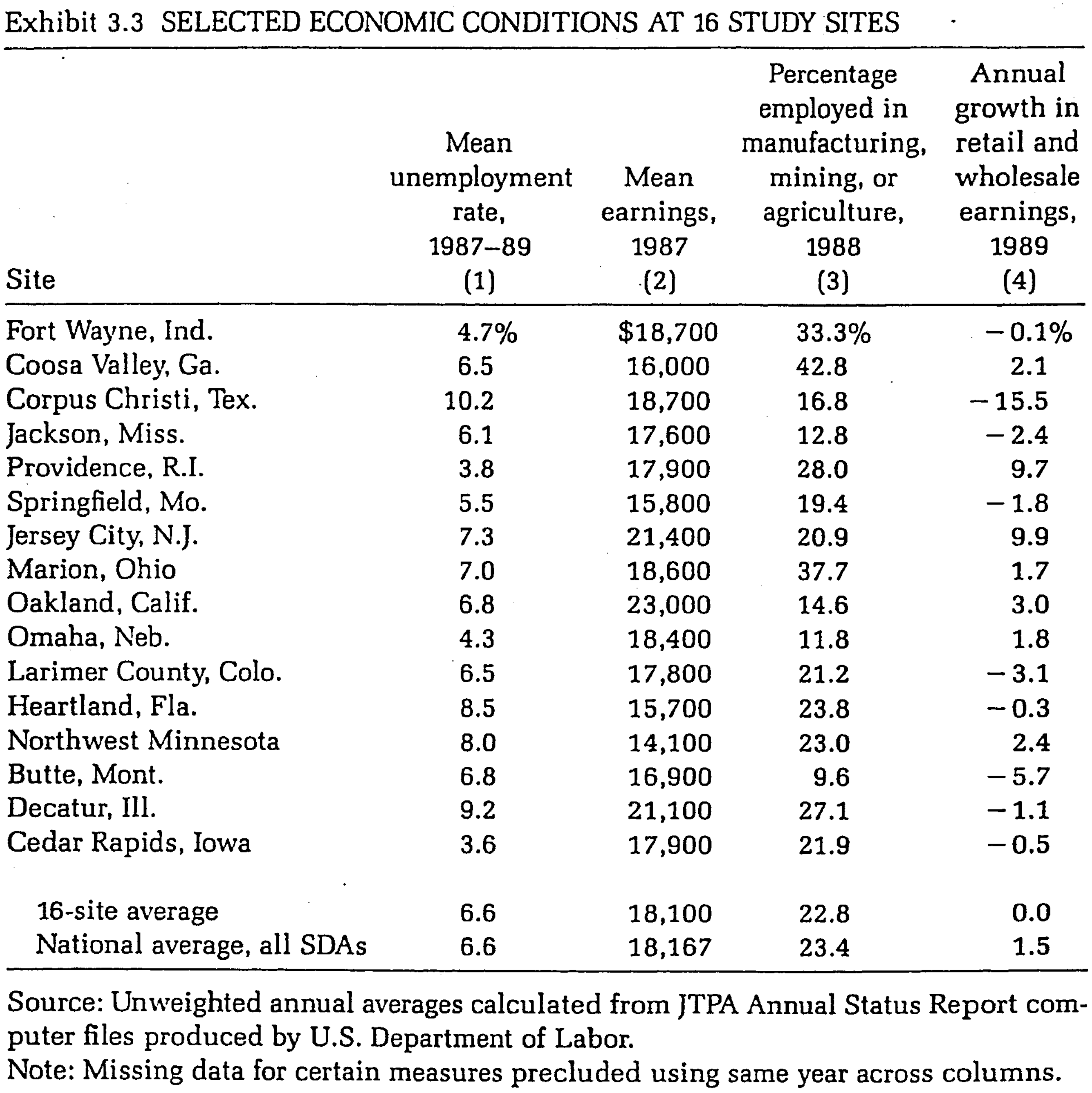


Exhibit 3.3 SELECTED ECONOMIC CONDITIONS AT 16 STUDY SITES

Site	Mean unemployment rate, 1987–89 (1)	Mean earnings, 1987 (2)	Percentage employed in manufacturing, mining, or agriculture, 1988 (3)	Annual growth in retail and wholesale earnings, 1989 (4)
Fort Wayne, Ind.	4.7%	$18,700	33.3%	−0.1%
Coosa Valley, Ga.	6.5	16,000	42.8	2.1
Corpus Christi, Tex.	10.2	18,700	16.8	−15.5
Jackson, Miss.	6.1	17,600	12.8	−2.4
Providence, R.I.	3.8	17,900	28.0	9.7
Springfield, Mo.	5.5	15,800	19.4	−1.8
Jersey City, N.J.	7.3	21,400	20.9	9.9
Marion, Ohio	7.0	18,600	37.7	1.7
Oakland, Calif.	6.8	23,000	14.6	3.0
Omaha, Neb.	4.3	18,400	11.8	1.8
Larimer County, Colo.	6.5	17,800	21.2	−3.1
Heartland, Fla.	8.5	15,700	23.8	−0.3
Northwest Minnesota	8.0	14,100	23.0	2.4
Butte, Mont.	6.8	16,900	9.6	−5.7
Decatur, Ill.	9.2	21,100	27.1	−1.1
Cedar Rapids, Iowa	3.6	17,900	21.9	−0.5
16-site average	6.6	18,100	22.8	0.0
National average, all SDAs	6.6	18,167	23.4	1.5

Source: Unweighted annual averages calculated from JTPA Annual Status Report computer files produced by U.S. Department of Labor.
Note: Missing data for certain measures precluded using same year across columns.

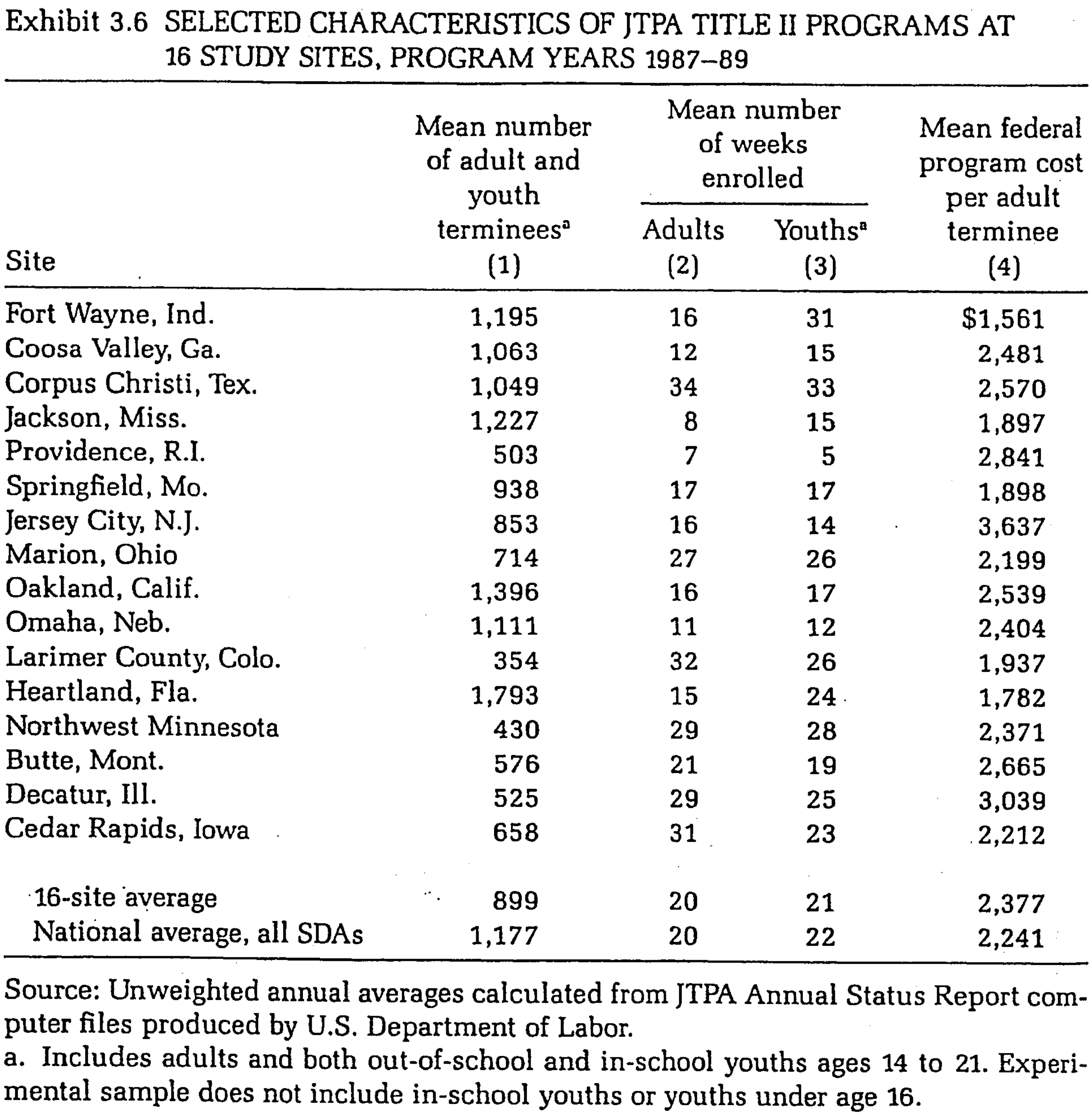


Exhibit 3.6 SELECTED CHARACTERISTICS OF JTPA TITLE II PROGRAMS AT 16 STUDY SITES, PROGRAM YEARS 1987–89

Site	Mean number of adult and youth terminees[a]	Mean number of weeks enrolled		Mean federal program cost per adult terminee
		Adults	Youths[a]	
	(1)	(2)	(3)	(4)
Fort Wayne, Ind.	1,195	16	31	$1,561
Coosa Valley, Ga.	1,063	12	15	2,481
Corpus Christi, Tex.	1,049	34	33	2,570
Jackson, Miss.	1,227	8	15	1,897
Providence, R.I.	503	7	5	2,841
Springfield, Mo.	938	17	17	1,898
Jersey City, N.J.	853	16	14	3,637
Marion, Ohio	714	27	26	2,199
Oakland, Calif.	1,396	16	17	2,539
Omaha, Neb.	1,111	11	12	2,404
Larimer County, Colo.	354	32	26	1,937
Heartland, Fla.	1,793	15	24	1,782
Northwest Minnesota	430	29	28	2,371
Butte, Mont.	576	21	19	2,665
Decatur, Ill.	525	29	25	3,039
Cedar Rapids, Iowa	658	31	23	2,212
16-site average	899	20	21	2,377
National average, all SDAs	1,177	20	22	2,241

Source: Unweighted annual averages calculated from JTPA Annual Status Report computer files produced by U.S. Department of Labor.

a. Includes adults and both out-of-school and in-school youths ages 14 to 21. Experimental sample does not include in-school youths or youths under age 16.

A final word about policy outcome

After the results of the National JTPA study were released, in 1994, funding for JTPA training for the youth were drastically cut:

Spending on JTPA Programs

Year	Youth Training Grants	Adult Training Grants
1993	677	1015
1994	609	988
1995	127	996
1996	127	850
1997	127	895

Revisiting SUTVA

- Recall our tacit assumption throughout: Stable Unit Treatment Value Assumption.
- Essentially the treatment from one unit does not interfere with the treatment from another.
- But in many real settings this can’t be true: spillovers and externalities.
- Experiments can help us to deal with this.

Example: Miguel and Kremer

- Idea:
  - Effect of health on schooling (but we assume that schooling → income)
  - Externalities
    - Important for policy
    - Intellectual curiosity
    - Recent obsession in applied economics (should we care?)

Setting

- Busia, Kenya – very poor
- 1.3 billion people worldwide have hookworms and roundworms
- 200 million have schistosomiasis
- Transmitted through contact with fecal matter (hookworm,roundworm) or infected freshwater (schistosomiasis)
- Low cost treatment

Setting (continued)

- Treatment at the school level – 75 schools and 30000 students
- 3 groups:
  - In 1998 – group 1 treated
  - In 1999 – group 1 & 2 treated
  - In 2001 – all are treated
- Reason? – Equity? Is this a good trick to randomize?

Baseline comparison

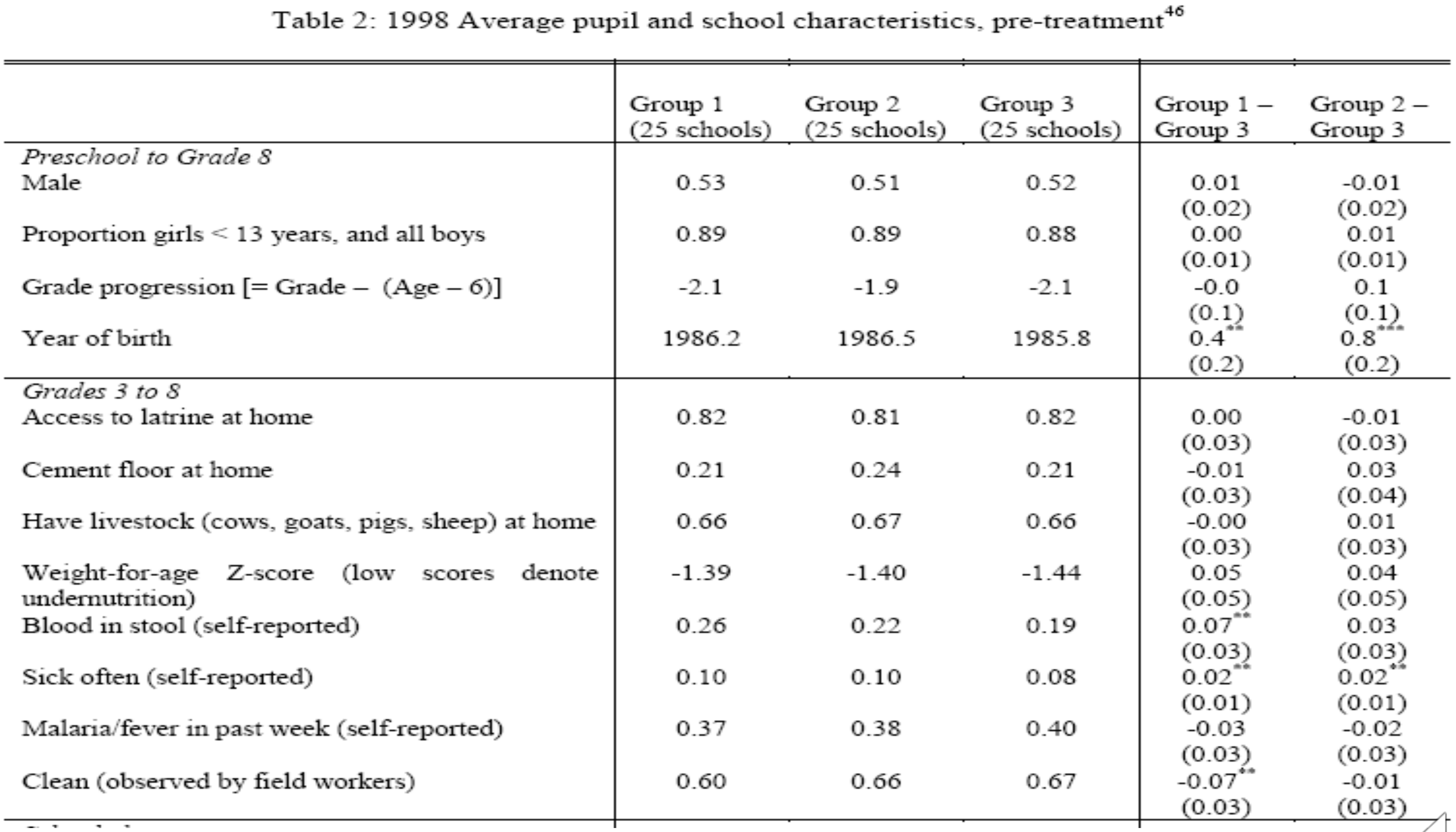


Table 2: 1998 Average pupil and school characteristics, pre-treatment[46]

	Group 1 (25 schools)	Group 2 (25 schools)	Group 3 (25 schools)	Group 1 – Group 3	Group 2 – Group 3
Preschool to Grade 8					
Male	0.53	0.51	0.52	0.01 (0.02)	-0.01 (0.02)
Proportion girls < 13 years, and all boys	0.89	0.89	0.88	0.00 (0.01)	0.01 (0.01)
Grade progression [= Grade – (Age – 6)]	-2.1	-1.9	-2.1	-0.0 (0.1)	0.1 (0.1)
Year of birth	1986.2	1986.5	1985.8	0.4** (0.2)	0.8*** (0.2)
Grades 3 to 8					
Access to latrine at home	0.82	0.81	0.82	0.00 (0.03)	-0.01 (0.03)
Cement floor at home	0.21	0.24	0.21	-0.01 (0.03)	0.03 (0.04)
Have livestock (cows, goats, pigs, sheep) at home	0.66	0.67	0.66	-0.00 (0.03)	0.01 (0.03)
Weight-for-age Z-score (low scores denote undernutrition)	-1.39	-1.40	-1.44	0.05 (0.05)	0.04 (0.05)
Blood in stool (self-reported)	0.26	0.22	0.19	0.07** (0.03)	0.03 (0.03)
Sick often (self-reported)	0.10	0.10	0.08	0.02** (0.01)	0.02** (0.01)
Malaria/fever in past week (self-reported)	0.37	0.38	0.40	-0.03 (0.03)	-0.02 (0.03)
Clean (observed by field workers)	0.60	0.66	0.67	-0.07** (0.03)	-0.01 (0.03)

These kids are sick

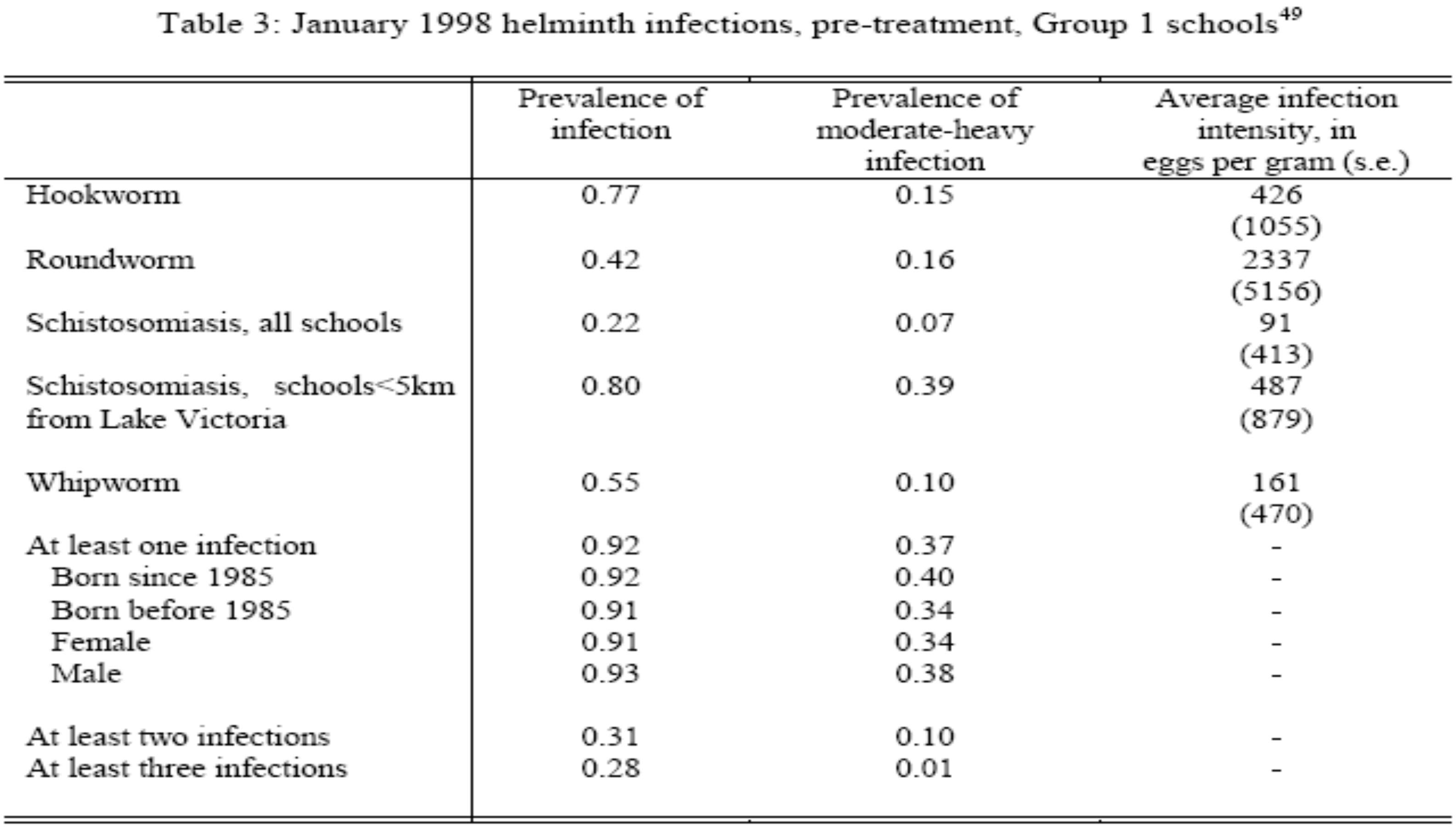


Table 3: January 1998 helminth infections, pre-treatment, Group 1 schools[49]

	Prevalence of infection	Prevalence of moderate-heavy infection	Average infection intensity, in eggs per gram (s.e.)
Hookworm	0.77	0.15	426 (1055)
Roundworm	0.42	0.16	2337 (5156)
Schistosomiasis, all schools	0.22	0.07	91 (413)
Schistosomiasis, schools<5km from Lake Victoria	0.80	0.39	487 (879)
Whipworm	0.55	0.10	161 (470)
At least one infection	0.92	0.37	-
Born since 1985	0.92	0.40	-
Born before 1985	0.91	0.34	-
Female	0.91	0.34	-
Male	0.93	0.38	-
At least two infections	0.31	0.10	-
At least three infections	0.28	0.01	-

Understanding treatment

- Apart from worm treatment:
  - Program provides education (is this bad?/interpretation)
  - Ethical problems (consent)
  - Other worries (expectations)

Understanding compliance

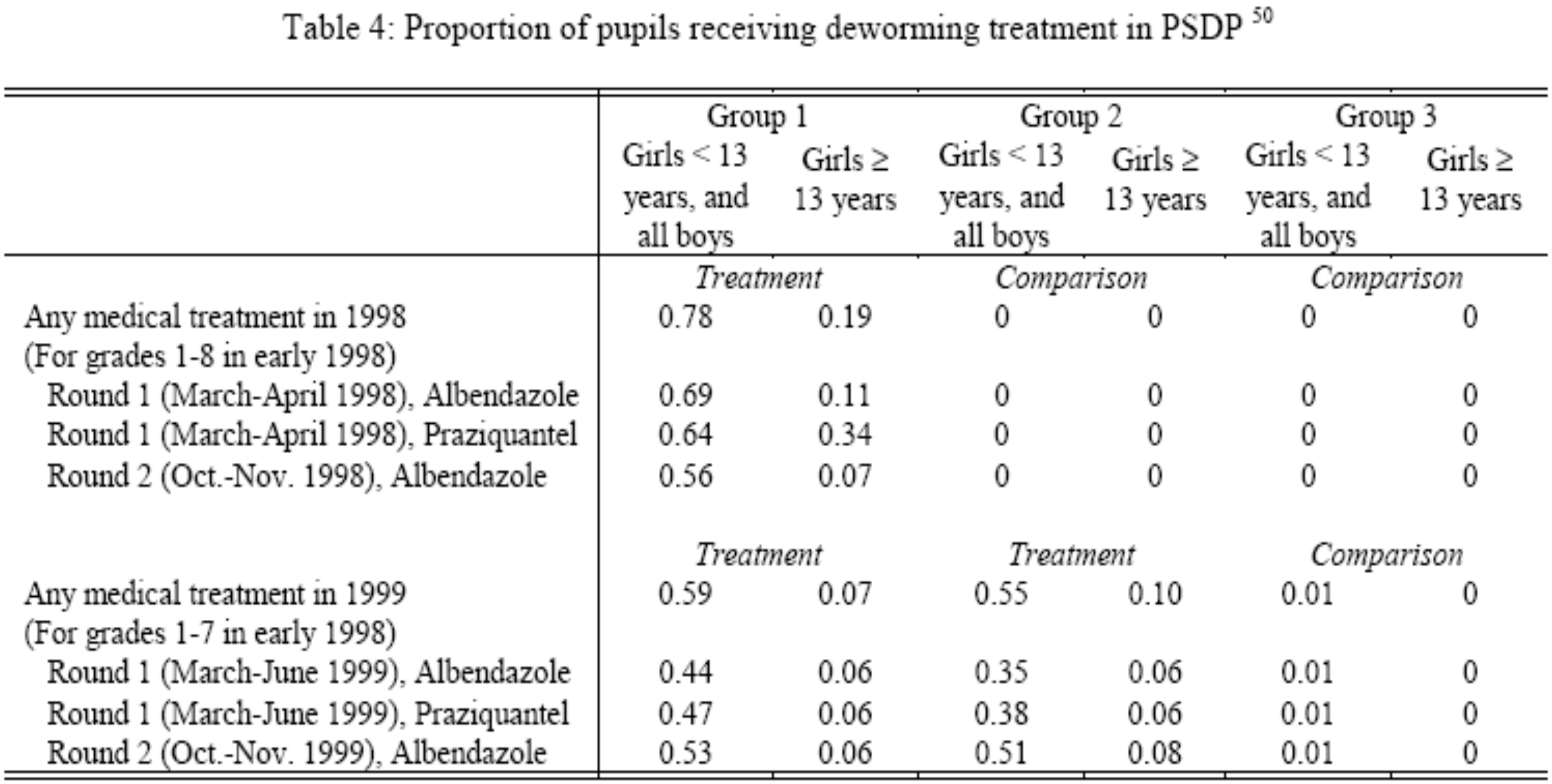


Table 4: Proportion of pupils receiving deworming treatment in PSDP [50]

	Group 1		Group 2		Group 3	
	Girls < 13 years, and all boys	Girls ≥ 13 years	Girls < 13 years, and all boys	Girls ≥ 13 years	Girls < 13 years, and all boys	Girls ≥ 13 years
	Treatment		*Comparison*		*Comparison*	
Any medical treatment in 1998	0.78	0.19	0	0	0	0
(For grades 1-8 in early 1998)						
Round 1 (March-April 1998), Albendazole	0.69	0.11	0	0	0	0
Round 1 (March-April 1998), Praziquantel	0.64	0.34	0	0	0	0
Round 2 (Oct.-Nov. 1998), Albendazole	0.56	0.07	0	0	0	0
	Treatment		*Treatment*		*Comparison*	
Any medical treatment in 1999	0.59	0.07	0.55	0.10	0.01	0
(For grades 1-7 in early 1998)						
Round 1 (March-June 1999), Albendazole	0.44	0.06	0.35	0.06	0.01	0
Round 1 (March-June 1999), Praziquantel	0.47	0.06	0.38	0.06	0.01	0
Round 2 (Oct.-Nov. 1999), Albendazole	0.53	0.06	0.51	0.08	0.01	0

Empirical strategy (1): direct effects

$$Y_{ijt} = a + \beta_1 \cdot T_{1it} + \beta_2 \cdot T_{2it} + X_{ijt}'\delta + u_i + e_{ijt}$$

- *i*-school, *j*-student, *t*-year
- T_1, T_2- indicate first and second year of deworming
- The beta’s are the interesting coefficients.

Empirical strategy (1): direct effects

$$Y_{ijt} = a + \beta_1 \cdot T_{1it} + \beta_2 \cdot T_{2it} + X_{ijt}'\delta + u_i + e_{ijt}$$

```
reg health_outcome year1_deworm year2_deworm
age sex mom_educ dad_educ i.school
```

Empirical strategy (2): externalities

$$Y_{ijt} = a + \beta_1 \cdot T_{1it} + \beta_2 \cdot T_{2it} + X_{ijt}{}'\delta$$

$$+ \sum_d (\gamma_d \cdot N_{dit}{}^{T}) + \sum_d (\phi_d \cdot N_{dit}) + u_i + e_{ijt}$$

- *N*= # of pupils in other schools within *X* kilometers (spillover effect)
- N_t = # of pupils in other schools within *X* kilometers that were treated (contamination effect)
- Key strategy: randomization not only of your treatment assignment, but of your neighbors.
- Implies density of your neighbor’s treatment status is also randomly assigned.

Direct effects: first stage

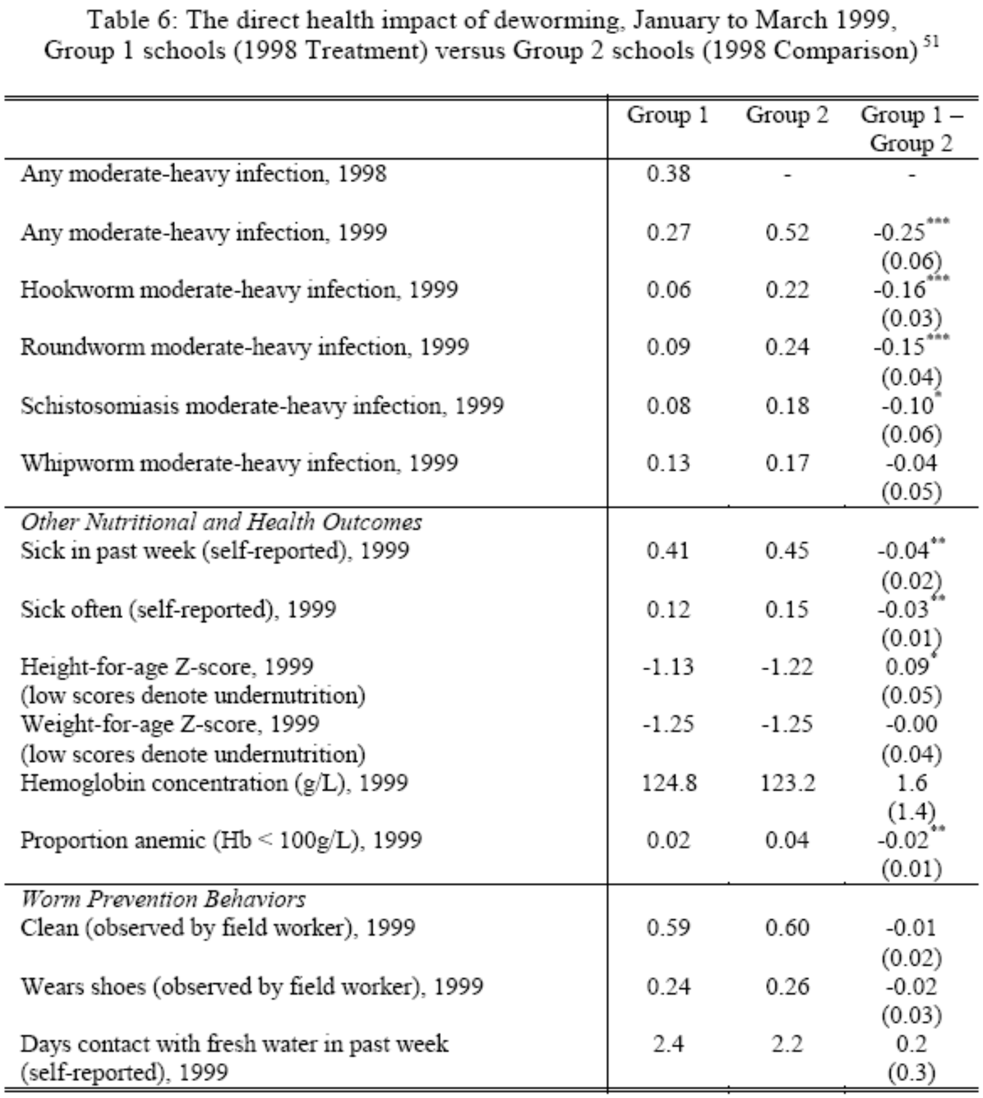


Table 6: The direct health impact of deworming, January to March 1999,
Group 1 schools (1998 Treatment) versus Group 2 schools (1998 Comparison) [51]

	Group 1	Group 2	Group 1 – Group 2
Any moderate-heavy infection, 1998	0.38	-	-
Any moderate-heavy infection, 1999	0.27	0.52	-0.25^{***} (0.06)
Hookworm moderate-heavy infection, 1999	0.06	0.22	-0.16^{***} (0.03)
Roundworm moderate-heavy infection, 1999	0.09	0.24	-0.15^{***} (0.04)
Schistosomiasis moderate-heavy infection, 1999	0.08	0.18	-0.10^{*} (0.06)
Whipworm moderate-heavy infection, 1999	0.13	0.17	-0.04 (0.05)
Other Nutritional and Health Outcomes			
Sick in past week (self-reported), 1999	0.41	0.45	-0.04^{**} (0.02)
Sick often (self-reported), 1999	0.12	0.15	-0.03^{**} (0.01)
Height-for-age Z-score, 1999 (low scores denote undernutrition)	-1.13	-1.22	0.09^{*} (0.05)
Weight-for-age Z-score, 1999 (low scores denote undernutrition)	-1.25	-1.25	-0.00 (0.04)
Hemoglobin concentration (g/L), 1999	124.8	123.2	1.6 (1.4)
Proportion anemic (Hb < 100g/L), 1999	0.02	0.04	-0.02^{**} (0.01)
Worm Prevention Behaviors			
Clean (observed by field worker), 1999	0.59	0.60	-0.01 (0.02)
Wears shoes (observed by field worker), 1999	0.24	0.26	-0.02 (0.03)
Days contact with fresh water in past week (self-reported), 1999	2.4	2.2	0.2 (0.3)

Across school externality

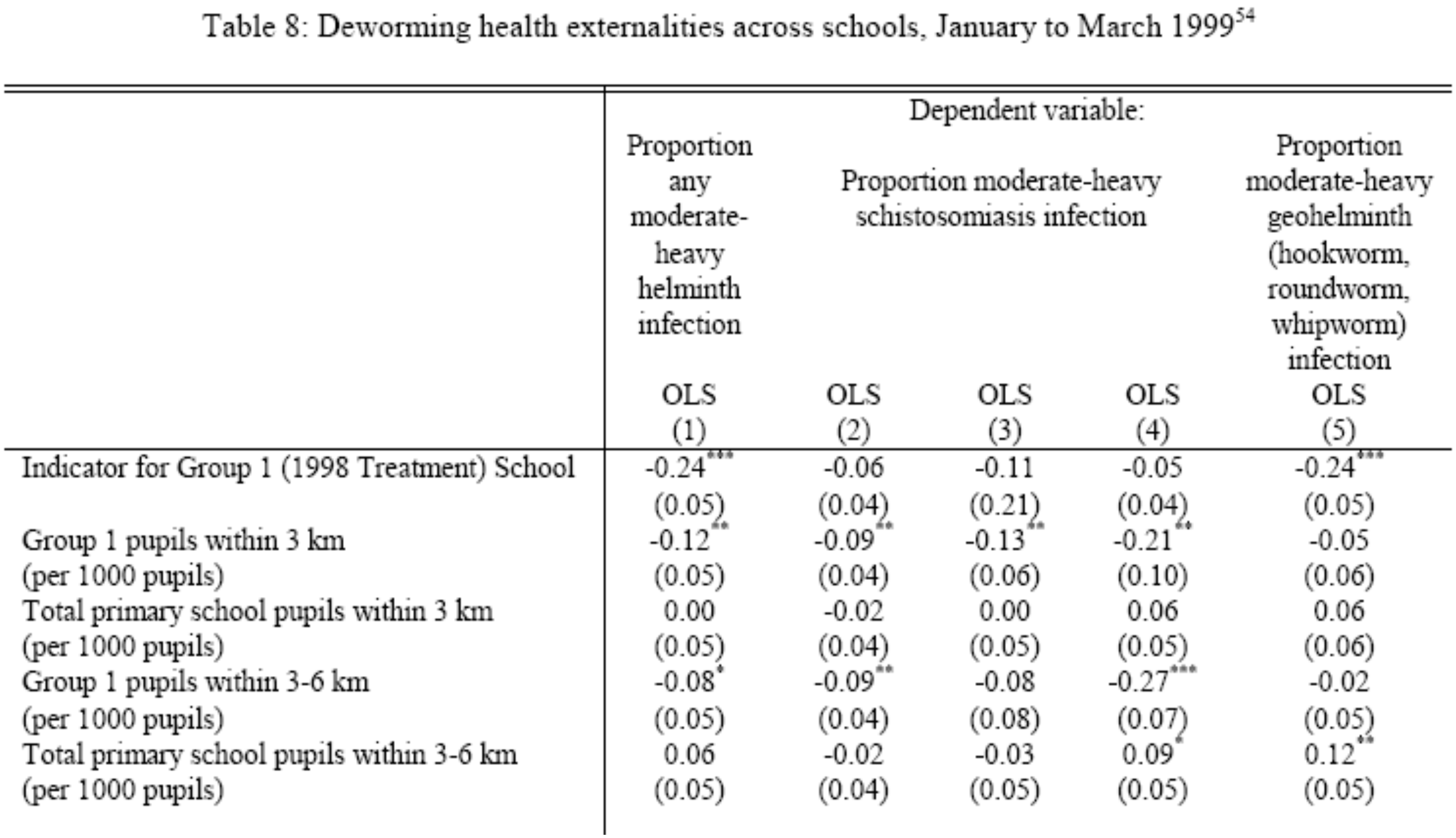
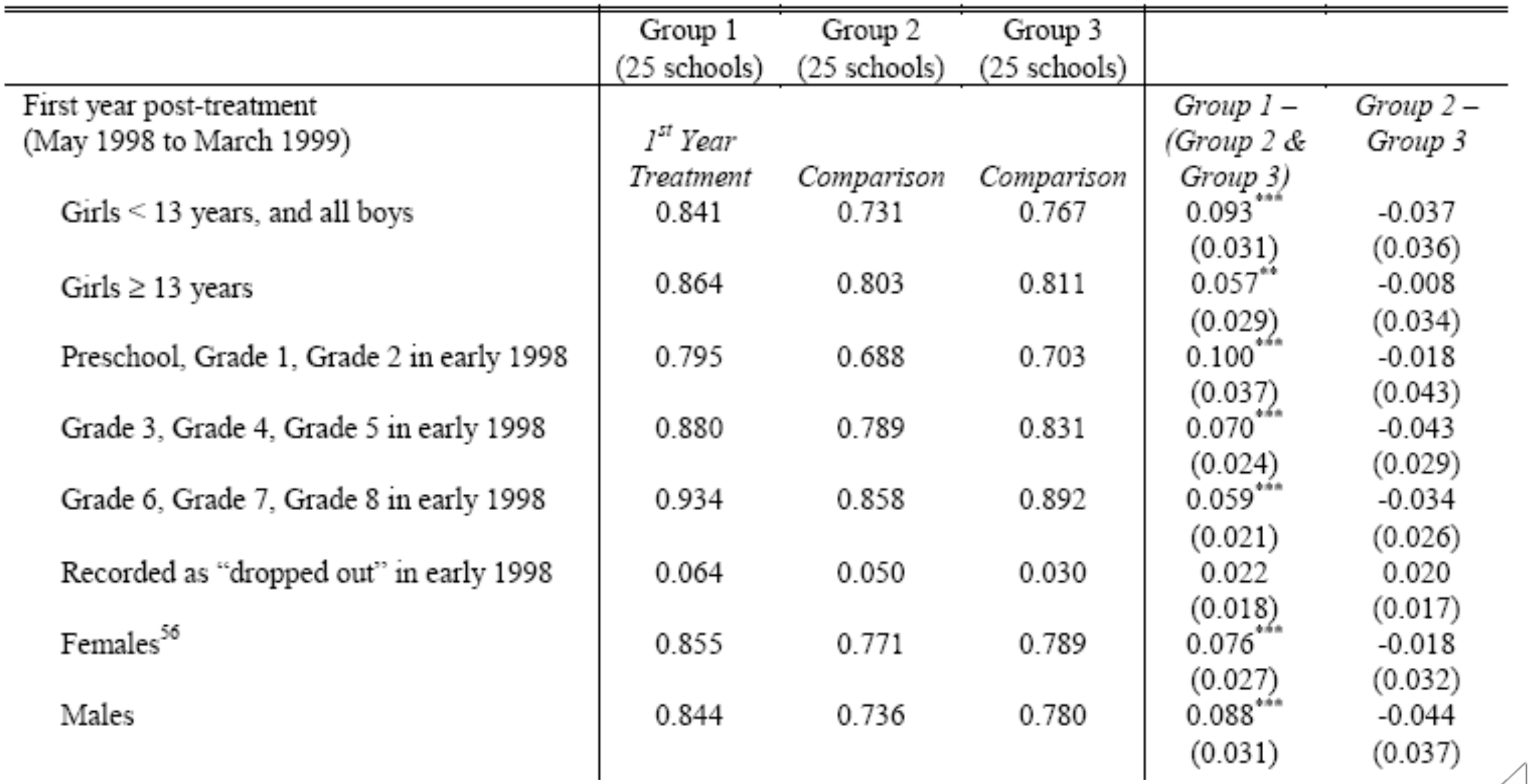


Table 8: Deworming health externalities across schools, January to March 1999[54]

	Dependent variable:				
	Proportion any moderate-heavy helminth infection	Proportion moderate-heavy schistosomiasis infection			Proportion moderate-heavy geohelminth (hookworm, roundworm, whipworm) infection
	OLS	OLS	OLS	OLS	OLS
	(1)	(2)	(3)	(4)	(5)
Indicator for Group 1 (1998 Treatment) School	-0.24***	-0.06	-0.11	-0.05	-0.24***
	(0.05)	(0.04)	(0.21)	(0.04)	(0.05)
Group 1 pupils within 3 km	-0.12**	-0.09**	-0.13**	-0.21**	-0.05
(per 1000 pupils)	(0.05)	(0.04)	(0.06)	(0.10)	(0.06)
Total primary school pupils within 3 km	0.00	-0.02	0.00	0.06	0.06
(per 1000 pupils)	(0.05)	(0.04)	(0.05)	(0.05)	(0.06)
Group 1 pupils within 3-6 km	-0.08*	-0.09**	-0.08	-0.27***	-0.02
(per 1000 pupils)	(0.05)	(0.04)	(0.08)	(0.07)	(0.05)
Total primary school pupils within 3-6 km	0.06	-0.02	-0.03	0.09*	0.12**
(per 1000 pupils)	(0.05)	(0.04)	(0.05)	(0.05)	(0.05)

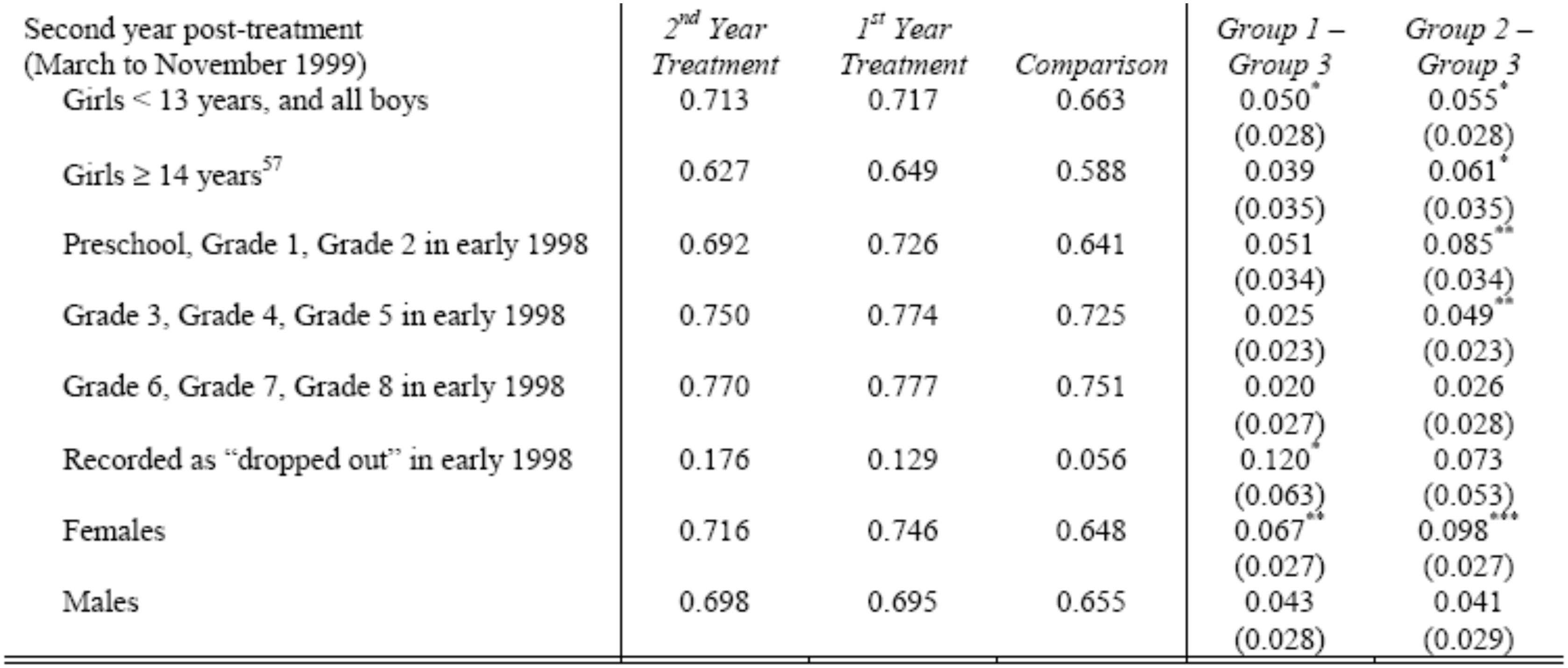
Second stage: schooling

Table 9: School participation, school-level data[55]



	Group 1 (25 schools)	Group 2 (25 schools)	Group 3 (25 schools)		
First year post-treatment (May 1998 to March 1999)	*1st Year Treatment*	*Comparison*	*Comparison*	*Group 1 – (Group 2 & Group 3)*	*Group 2 – Group 3*
Girls < 13 years, and all boys	0.841	0.731	0.767	0.093^{***} (0.031)	-0.037 (0.036)
Girls ≥ 13 years	0.864	0.803	0.811	0.057^{**} (0.029)	-0.008 (0.034)
Preschool, Grade 1, Grade 2 in early 1998	0.795	0.688	0.703	0.100^{***} (0.037)	-0.018 (0.043)
Grade 3, Grade 4, Grade 5 in early 1998	0.880	0.789	0.831	0.070^{***} (0.024)	-0.043 (0.029)
Grade 6, Grade 7, Grade 8 in early 1998	0.934	0.858	0.892	0.059^{***} (0.021)	-0.034 (0.026)
Recorded as "dropped out" in early 1998	0.064	0.050	0.030	0.022 (0.018)	0.020 (0.017)
Females[56]	0.855	0.771	0.789	0.076^{***} (0.027)	-0.018 (0.032)
Males	0.844	0.736	0.780	0.088^{***} (0.031)	-0.044 (0.037)

Second stage: schooling



Second year post-treatment	*2nd Year*	*1st Year*		*Group 1 –*	*Group 2 –*
(March to November 1999)	*Treatment*	*Treatment*	*Comparison*	*Group 3*	*Group 3*
Girls < 13 years, and all boys	0.713	0.717	0.663	0.050*	0.055*
				(0.028)	(0.028)
Girls ≥ 14 years[57]	0.627	0.649	0.588	0.039	0.061*
				(0.035)	(0.035)
Preschool, Grade 1, Grade 2 in early 1998	0.692	0.726	0.641	0.051	0.085**
				(0.034)	(0.034)
Grade 3, Grade 4, Grade 5 in early 1998	0.750	0.774	0.725	0.025	0.049**
				(0.023)	(0.023)
Grade 6, Grade 7, Grade 8 in early 1998	0.770	0.777	0.751	0.020	0.026
				(0.027)	(0.028)
Recorded as “dropped out” in early 1998	0.176	0.129	0.056	0.120*	0.073
				(0.063)	(0.053)
Females	0.716	0.746	0.648	0.067**	0.098***
				(0.027)	(0.027)
Males	0.698	0.695	0.655	0.043	0.041
				(0.028)	(0.029)

School externalities (across)

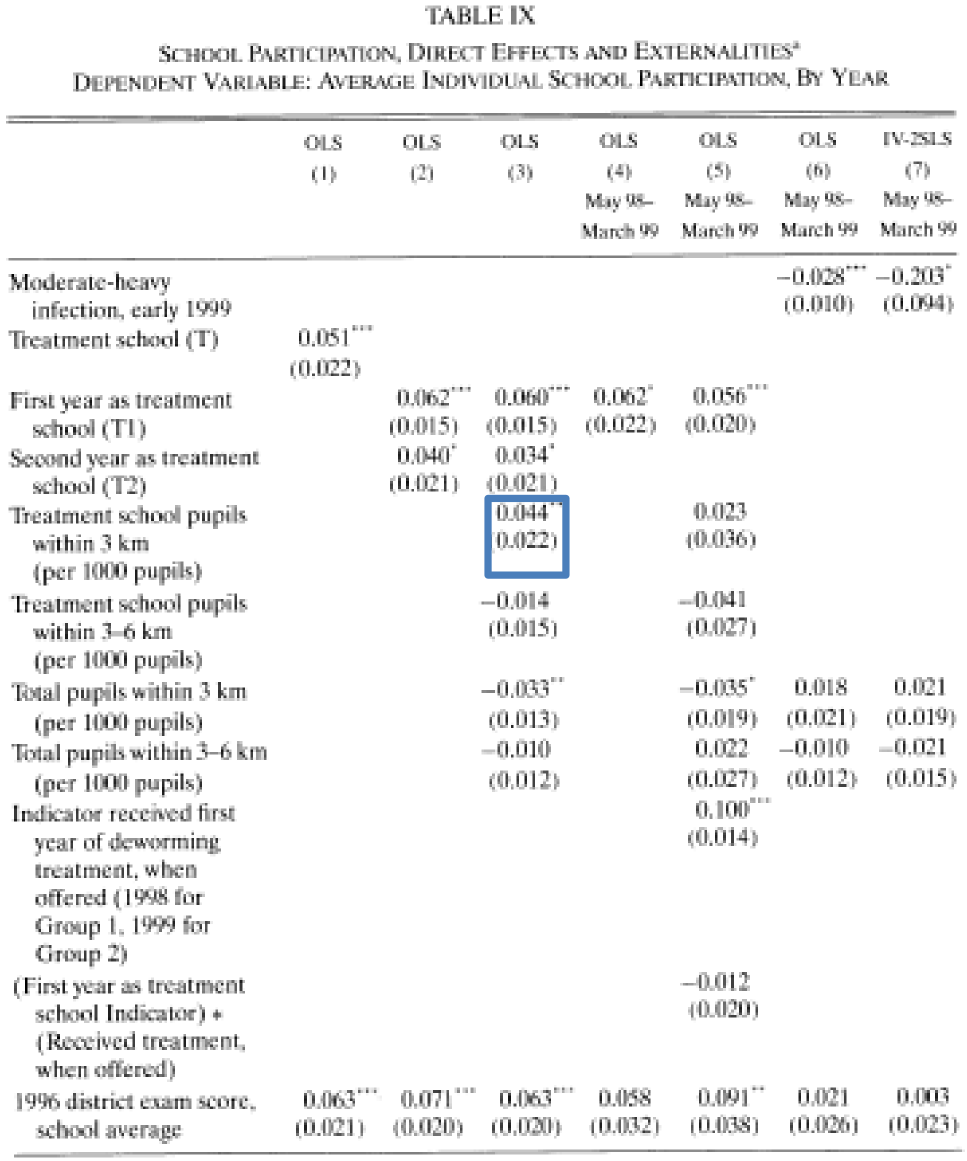


TABLE IX

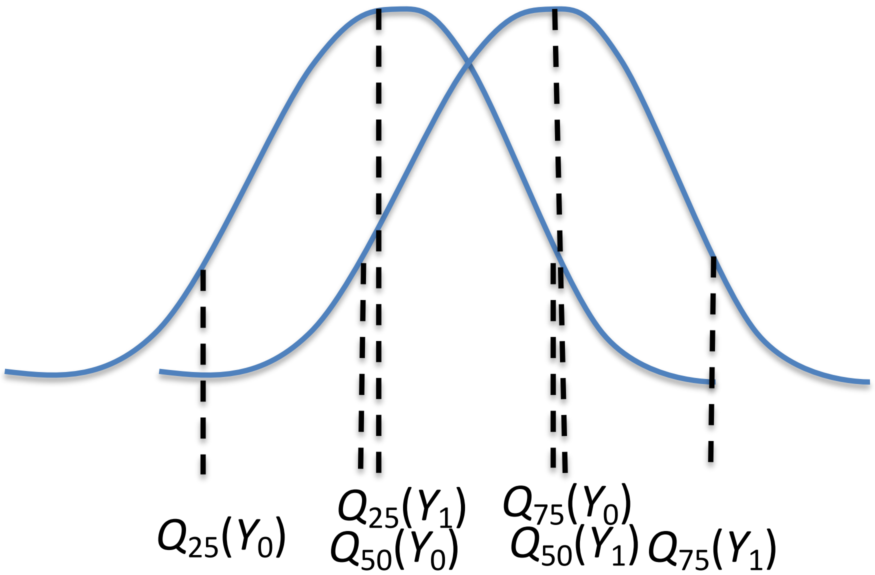
SCHOOL PARTICIPATION, DIRECT EFFECTS AND EXTERNALITIES[a]
DEPENDENT VARIABLE: AVERAGE INDIVIDUAL SCHOOL PARTICIPATION, BY YEAR

	OLS (1)	OLS (2)	OLS (3)	OLS (4) May 98– March 99	OLS (5) May 98– March 99	OLS (6) May 98– March 99	IV-2SLS (7) May 98– March 99
Moderate-heavy infection, early 1999						-0.028^{***} (0.010)	-0.203^{*} (0.094)
Treatment school (T)	0.051^{***} (0.022)						
First year as treatment school (T1)		0.062^{***} (0.015)	0.060^{***} (0.015)	0.062^{*} (0.022)	0.056^{***} (0.020)		
Second year as treatment school (T2)		0.040^{*} (0.021)	0.034^{*} (0.021)				
Treatment school pupils within 3 km (per 1000 pupils)			0.044^{**} (0.022)		0.023 (0.036)		
Treatment school pupils within 3–6 km (per 1000 pupils)			−0.014 (0.015)		−0.041 (0.027)		
Total pupils within 3 km (per 1000 pupils)			-0.033^{**} (0.013)		-0.035^{*} (0.019)	0.018 (0.021)	0.021 (0.019)
Total pupils within 3–6 km (per 1000 pupils)			−0.010 (0.012)		0.022 (0.027)	−0.010 (0.012)	−0.021 (0.015)
Indicator received first year of deworming treatment, when offered (1998 for Group 1, 1999 for Group 2)					0.100^{***} (0.014)		
(First year as treatment school Indicator) * (Received treatment, when offered)					−0.012 (0.020)		
1996 district exam score, school average	0.063^{***} (0.021)	0.071^{***} (0.020)	0.063^{***} (0.020)	0.058 (0.032)	0.091^{**} (0.038)	0.021 (0.026)	0.003 (0.023)

Quantile treatment effects

- Recall that randomized experiments allow us to compute treatment effects not only for means but also for other features of the distribution.
- Practically relevant are quantile treatment effects.
- E.g., we might be interested whether at the 75th percentile there was a positive treatment effect.

Quantile treatment effects



- $Q_{50}(Y_1)$-$Q_{50}(Y_0)$
- $Q_{25}(Y_1)$-$Q_{25}(Y_0)$
- $Q_{75}(Y_1)$-$Q_{75}(Y_0)$

QTE vs Q(TE)

- What we are identifying is the quantile treatment effect, not the treatment effect of the quantiles.
- I.e., just the difference between the n-th treatment and control percentiles, not the n-th percentile of the treatment-control difference.

Example: Bitler, Gelbach, Hoynes

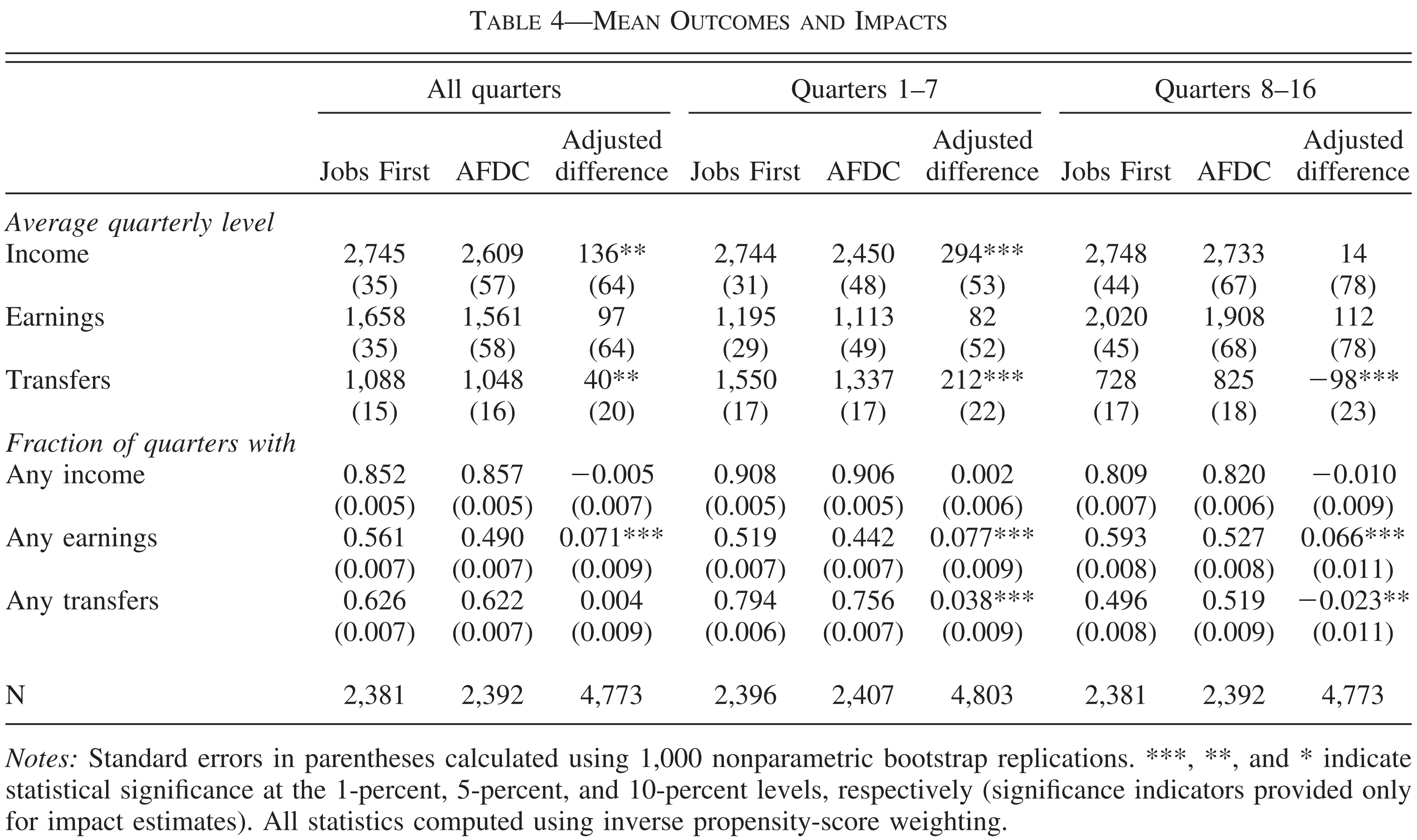


Table 4—Mean Outcomes and Impacts

	All quarters			Quarters 1–7			Quarters 8–16		
	Jobs First	AFDC	Adjusted difference	Jobs First	AFDC	Adjusted difference	Jobs First	AFDC	Adjusted difference
Average quarterly level									
Income	2,745	2,609	136**	2,744	2,450	294***	2,748	2,733	14
	(35)	(57)	(64)	(31)	(48)	(53)	(44)	(67)	(78)
Earnings	1,658	1,561	97	1,195	1,113	82	2,020	1,908	112
	(35)	(58)	(64)	(29)	(49)	(52)	(45)	(68)	(78)
Transfers	1,088	1,048	40**	1,550	1,337	212***	728	825	−98***
	(15)	(16)	(20)	(17)	(17)	(22)	(17)	(18)	(23)
Fraction of quarters with									
Any income	0.852	0.857	−0.005	0.908	0.906	0.002	0.809	0.820	−0.010
	(0.005)	(0.005)	(0.007)	(0.005)	(0.005)	(0.006)	(0.007)	(0.006)	(0.009)
Any earnings	0.561	0.490	0.071***	0.519	0.442	0.077***	0.593	0.527	0.066***
	(0.007)	(0.007)	(0.009)	(0.007)	(0.007)	(0.009)	(0.008)	(0.008)	(0.011)
Any transfers	0.626	0.622	0.004	0.794	0.756	0.038***	0.496	0.519	−0.023**
	(0.007)	(0.007)	(0.009)	(0.006)	(0.007)	(0.009)	(0.008)	(0.009)	(0.011)
N	2,381	2,392	4,773	2,396	2,407	4,803	2,381	2,392	4,773

Notes: Standard errors in parentheses calculated using 1,000 nonparametric bootstrap replications. ***, **, and * indicate statistical significance at the 1-percent, 5-percent, and 10-percent levels, respectively (significance indicators provided only for impact estimates). All statistics computed using inverse propensity-score weighting.

Example (con't)

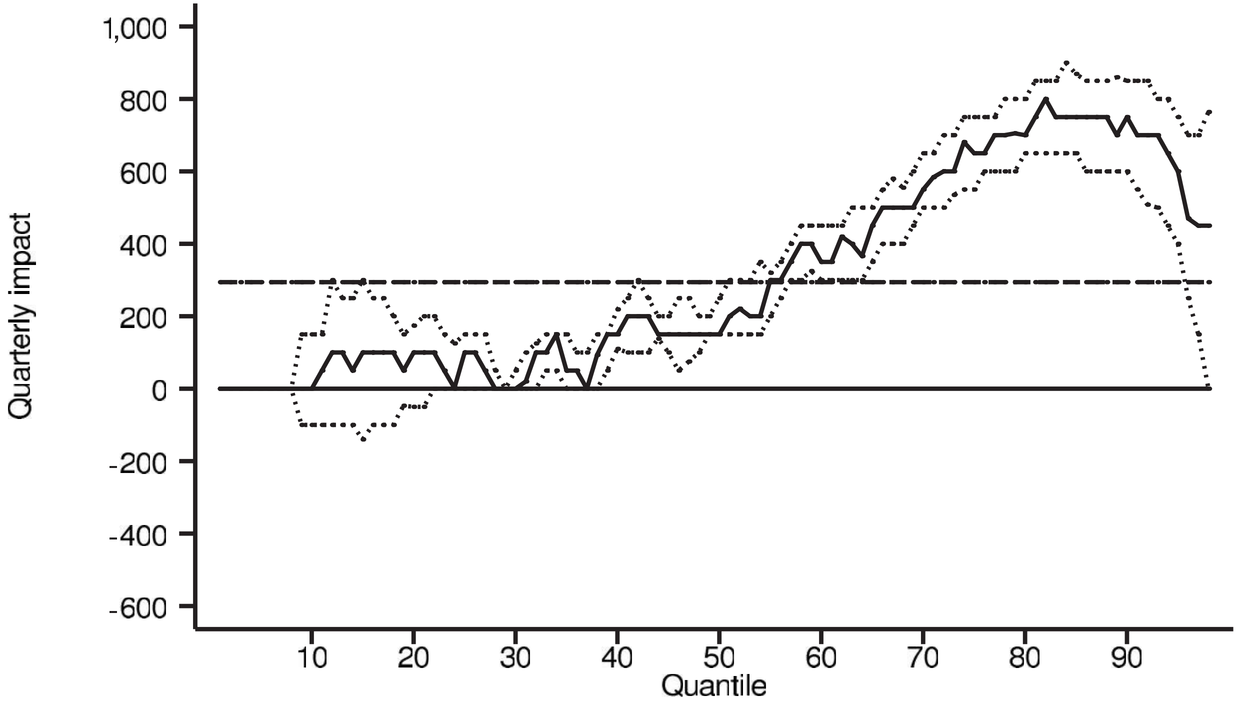


Figure 7. Quantile Treatment Effects on the Distribution of Income, Quarters 1–7

Notes: Solid line is QTE; dotted lines provide bootstrapped 90-percent confidence intervals; dashed line is mean impact; all statistics computed using inverse propensity-score weighting. See text for more details.