

Understanding TIF Evaluation Estimates: Effect Size and Power

John Keltz

CECR Evaluation
Conference



Introduction

- Effect Size
 - Explanation
 - TIF evaluation
- Power Analysis
 - Explanation
 - Simulation
 - Discussion

Effect Size

- For ease of understanding and comparing results across studies, estimates can be reported in effect sizes.

$$\text{Effect size} = \frac{\text{Treatment group mean} - \text{Control group mean}}{\text{Standard deviation}}$$

- Effect size informs us of the effects of a treatment relative to the population.
- Education studies often report effect sizes on test scores.

Effect Size

- For ease of understanding and comparing results across studies, estimates can be reported in effect sizes.

$$\text{Effect size} = \frac{\text{Treatment group mean} - \text{Control group mean}}{\text{Standard deviation}}$$

- Effect size informs us of the effects of a treatment relative to the population.
- Education studies often report effect sizes on test scores.

Effect Size

- If the value added outcome measure was standardized, then coefficient on the TIF effect is a Value Added effect size.
 - This informs us how the TIF program has improved a school relative to the distribution of school value added.
 - However, we might want to transform the TIF effect to an effect standardized by student test scores.

$$\text{TIF effectsize}_{\text{VA}} * \frac{\text{Value added standard deviation}}{\text{Test score standard deviation}} = \text{TIF effectsize}_{\text{test score}}$$

Effect Size: In context

An effect size of .25 would move a student from the 50th percentile of students to the 60th percentile. An effect size of 1 would move the student to the 84th percentile.

- Krueger (1999) found that the Tennessee STAR class size reduction resulted in an effect size of about .2.
- Milwaukee public school students receive an average effect size of .2 to .25 from a year of school.
- Coe (2002) reports recent studies that show effect sizes of .3 to .6 for a year of school in England.

Effect Size

- **Discussion:** what effect sizes might we expect or want to be able to detect for a successful TIF program?

Power Analysis

- Researchers designing experiments must ensure that the experiment will be capable of detecting expected effect sizes.
 - Standard error of the result estimates will depend on the parameters of the experiment.
 - Researchers select parameters to ensure the standard error will be small enough to detect reasonable effects.

Power Analysis

- Power Analysis considers both types of statistical error:
 - Type I error: the statistical test accepts an effect that doesn't exist
 - Type II error: the statistical test rejects an effect that actually does exist
 - Traditionally, power analysis is used to design an experiment that limits type I error to 5% and type II error to 20%.
- Power is the probability that a statistic accepts an effect that actually does exist

Power Analysis

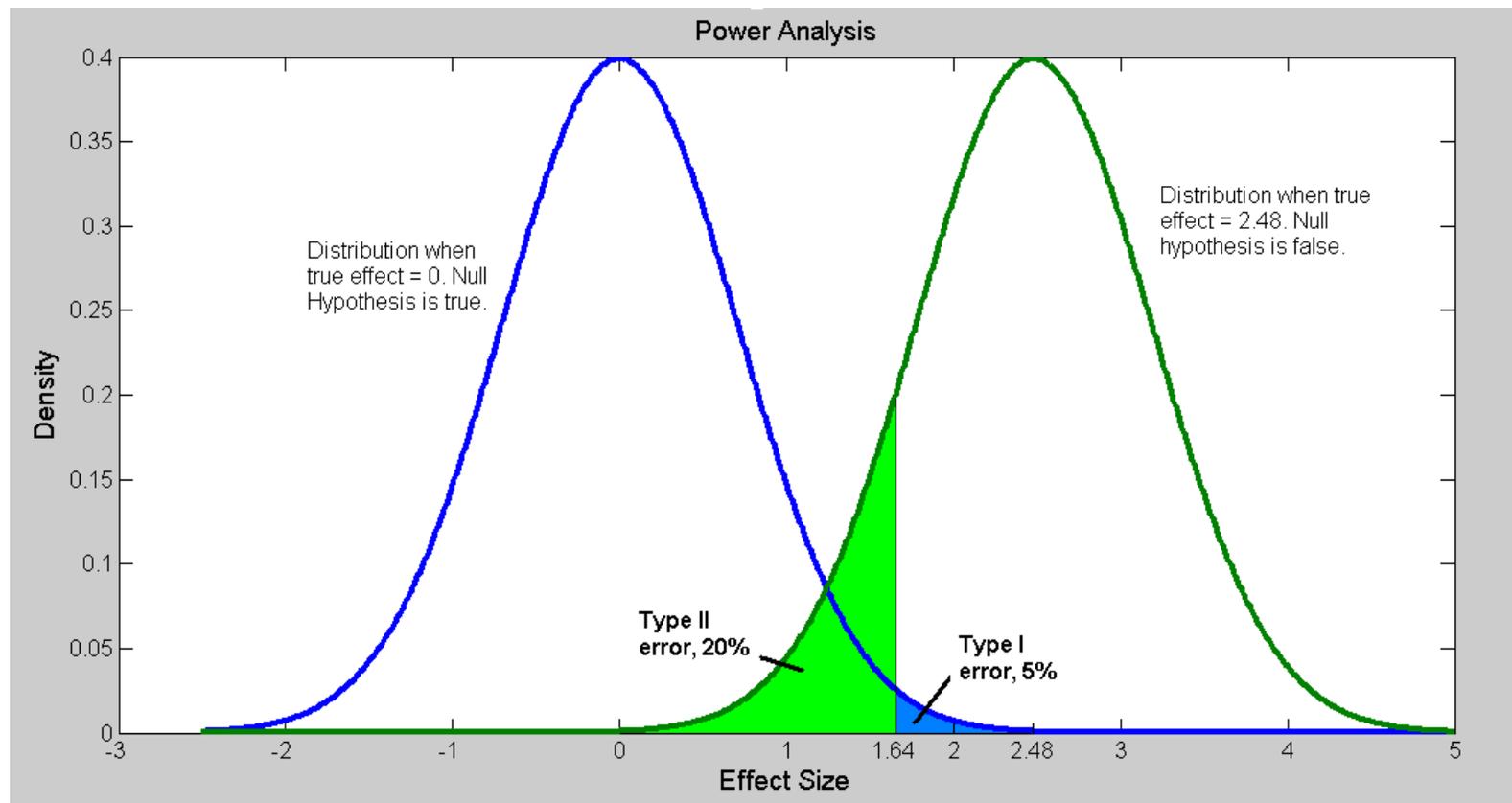
- If a one-tailed test is used, the experiment will be appropriately powered to detect an effect 2.48 times greater than the standard error of the estimate.

5% type I implies 1.64

20% type II implies .84

$$1.64 + .84 = 2.48$$

Power Analysis



Power Analysis

- We can then use power analysis to predict the detectable effect size for both experimental and quasi-experimental models.
 - Power can be predicted by using variance information in a mathematical formula.
 - In the absence of power formulas for other evaluation models, power can be predicted with real or simulated data.

Power Analysis: Simulation

- Analysis was performed using VA data from elementary schools in a large public school district.
 - District contained 120 elementary schools, and an average of 45 students per grade per school.
 - Value added results had a standard deviation of 9 test points for an exam with a standard deviation of 50.
 - Standard error on the grade level value added was an average of 4.4. The error on the school level value added was an average of 3.1.

Power Analysis: Simulation

- Two years of VA data were used; TIF indicators were randomly assigned to half of the schools in the second year.
 - Designed to simulate a differences in differences model of the TIF program.

$$\text{TIF effect} = (\text{School_VA}_{1,1} - \text{School_VA}_{1,0}) - (\text{School_VA}_{0,1} - \text{School_VA}_{0,0})$$

where subscripts represent: treatment school, year
treatment occurs at 1,1

- Expectation of TIF coefficient in the simulation is zero, but the simulation estimation error should be the same as an actual evaluation.

Power Analysis: Simulation Model

- School level differences in differences model

$$VA_{s,t} = \alpha + \beta_1 TIF_{s,t} + \beta_2 Y_t + \beta_3 S_s + \varepsilon_{s,t}$$

Where,

$VA_{s,t}$ = value added (school s, year t)

$TIF_{s,t}$ = indicator for TIF participation

Y_t = year

S_s = school

$\varepsilon_{s,t}$ = error

Regression is weighted by the inverse of value added standard errors

Power Analysis: Simulation Results

Power Estimates for dif-in-dif TIF evaluation, by number of schools

| Treatment Schools | Control Schools | | | |
|-------------------|-----------------|----|----|----|
| | 5 | 10 | 20 | 40 |
| 5 | 0.297 | | | |
| 10 | | | | |
| 20 | | | | |
| 40 | | | | |

Standard Errors in student effect size units

Power Analysis: Results

Power Estimates for dif-in-dif TIF evaluation, by number of schools

| Treatment Schools | Control Schools | | | |
|-------------------|-----------------|-------|-------|-------|
| | 5 | 10 | 20 | 40 |
| 5 | 0.297 | | | |
| 10 | | 0.171 | | |
| 20 | | | 0.108 | |
| 40 | | | | 0.078 |

Standard Errors in student effect size units

Power Analysis: Results

Power Estimates for dif-in-dif TIF evaluation, by number of schools

| Treatment Schools | Control Schools | | | |
|-------------------|-----------------|-------|-------|-------|
| | 5 | 10 | 20 | 40 |
| 5 | 0.297 | 0.239 | 0.189 | 0.164 |
| 10 | | 0.171 | | |
| 20 | | | 0.108 | |
| 40 | | | | 0.078 |

Standard Errors in student effect size units

Power Analysis: Results

Power Estimates for dif-in-dif TIF evaluation, by number of schools

| Treatment Schools | Control Schools | | | |
|-------------------|-----------------|-------|-------|-------|
| | 5 | 10 | 20 | 40 |
| 5 | 0.297 | 0.239 | 0.189 | 0.164 |
| 10 | 0.217 | 0.171 | | |
| 20 | 0.188 | | 0.108 | |
| 40 | 0.162 | | | 0.078 |

Standard Errors in student effect size units

Power Analysis: Results

Power Estimates for dif-in-dif TIF evaluation, by number of schools

| Treatment Schools | Control Schools | | | |
|-------------------|-----------------|-------|-------|-------|
| | 5 | 10 | 20 | 40 |
| 5 | 0.297 | 0.239 | 0.189 | 0.164 |
| 10 | 0.217 | 0.171 | 0.147 | 0.124 |
| 20 | 0.188 | 0.144 | 0.108 | 0.095 |
| 40 | 0.162 | 0.120 | 0.094 | 0.078 |

Standard Errors in student effect size units

Power Analysis: Discussion

- **Discussion:** An evaluation with 20 treatment schools and 20 control schools was powered to detect an effect of .108. What else can be done to improve power?

Power Analysis: Discussion

- **Discussion:** An evaluation with 20 treatment schools and 20 control schools was is powered to detect an effect of .108. What else can be done to improve power?
 - Additional years
 - Remove baseline year through matching or improved program assignment
 - Predictive school level covariates?

Power Analysis: Additional Topics

- Effect of additional years
- District level effects
- Varying school sizes
- Different value-added estimators

TIF Evaluation Examples and Discussion

John Keltz and Peter Witham

CECR Evaluation Conference

Introduction

- Example A: Small district
 - Discussion: district effect
 - Discussion: matching variables
- Example B: Group of small districts
 - Discussion: Outcome variable level
 - Discussion: Whether to estimate new outcomes
- Example C: Large District
 - Discussion: Controlling for selection

Example A: Small District

| District Features | |
|-----------------------------|--|
| # of Districts | 1 |
| # of Schools in District | 10 |
| # of Schools in TIF Program | 10 |
| School Selection Criteria | <i>All Schools Qualify for TIF Eligibility Criteria</i> |
| Years of TIF Program | <i>1 Year</i> |
| Student Outcome Data | <i>State Value Added at School Level (2 Years of VA Results)</i> |

Questions for Evaluation

- What are the best options for a control group?
- What are the threats to validity?
- What evaluation model should be used?
- Will the evaluation detect reasonable effect sizes?

Questions for Evaluation

- What is the best option for a control group?
 - A) A district of similar size and demographics
 - B) Nearby/similar districts
 - C) All districts in the state
- What are the threats to validity?
 - District effect is collinear with TIF status.
- What evaluation model should be used?
 - Difference in difference
 - Matched districts

Questions for Evaluation

- Will the evaluation detect reasonable effect sizes?
 - The district effect is very limiting. If we view the program as taking place at the district level we only have one observation of the treatment.
 - **Discussion:** Can we justify ignoring the district effect and evaluate at the school level? Are there any other ways to increase power?

Questions for Evaluation

- **Discussion:** What is “similar” when selecting schools or districts?
 - Demographics
 - Pretests
 - Past Value Added
 - Teacher tenure
 - School size
 - Urban/rural similarities

Example B: Group of small districts

| District Features | |
|-----------------------------|---|
| # of Districts | 7 |
| # of Schools in Districts | 25 |
| # of Schools in TIF Program | 25 |
| School Selection Criteria | <i>All eligible schools selected</i> |
| Years of TIF Program | <i>1 Year</i> |
| Student Outcome Data | <i>Achievement test data (3 Years of Results)</i> |

Questions for Evaluation

- What is the best option for a control group?
- What are the threats to validity?
- What evaluation model should be used?
- How should the evaluation differ when value added scores are not available?

Questions for Evaluation

- What is the best option for a control group?
 - A) Districts of similar size and demographics
 - B) Nearby districts
 - C) All districts in the state
- What are the threats to validity?
 - Selection of districts into TIF program.
- What evaluation model should be used?
 - Difference in difference
 - Matched districts

Questions for Evaluation

- How should the evaluation differ when value added scores are not available?
 - Evaluator can find school or district effects for treatment and control groups and then regress the effects on the right hand side variables.
 - Some evaluators may choose to do this even when value added is available.

Modeling: Value added as an outcome variable

- **Discussion:** Should we use value added measurements at the school level, grade level, or class level?
 - Program implementation is at the school level
 - Estimation often occurs at grade level
 - Does this depend on incentive structure?
 - Do we have reason to believe TIF would have different effects on different grades?

Modeling: Value added as an outcome variable

- Note: If outcome variable is smaller than the school level, and there is more than one observation per school, per year, standard errors may be incorrect.
 - Error will be correlated for the within school observations.
 - Corrections include the Moulton correction or a software package that does this for you, such as the cluster command in STATA.

Modeling: Value added as an outcome variable

- **Discussion:** When might the evaluator want to compute a new outcome variable instead of using TIF Value Added numbers?
 - Different control groups
 - Could use different exams
 - Do we want to avoid post-estimation steps to value-added, such as shrinkage?

Example C: Large District

| District Features | |
|-----------------------------|---|
| # of Districts | 1 |
| # of Schools in District | 150 |
| # of Schools in TIF Program | 20 |
| School Selection Criteria | <i>100 Eligible schools, 40 volunteers, 20 volunteers selected based on "readiness"</i> |
| Years of TIF Program | <i>1 Year</i> |
| Student Outcome Data | <i>District Value Added at School Level (2 Years of VA Results)</i> |

Questions for Evaluation

- What is the best option for a control group?
- What are the threats to validity?
- What evaluation model should be used?
- Will the evaluation detect reasonable effect sizes?

Questions for Evaluation

- What is the best option for a control group?
 - A) Other volunteers
 - B) All eligible schools
 - C) All schools in district
- What are the threats to validity?
 - Unobservables associated with both volunteering and “readiness”. Could press district for quantitative measure of readiness.
- What evaluation model should be used?
 - Two year time series
- Will the evaluation detect reasonable effect sizes?

Questions for Evaluation

- **Discussion:** How can we control for selection into the program?

Other Assignment Options

- How else could the district have assigned students?
 - Random assignment of 40 volunteers.
 - Tradeoff of readiness argument versus effective evaluation.
 - Selection on observables of 40 volunteers.
 - District could quantify “readiness”, allowing for regression discontinuity.
 - Less power than random assignment case.