The Late Pretest Problem in Randomized Control Trials of Education Interventions

Peter Z. Schochet

ACF Methods Conference, September 2012

In *Journal of Educational and Behavioral Statistics*, August 2010, Vol. 35

MATHEMATICA
Policy Research, Inc.

Pretest-Posttest Designs Are Common for Education RCTs

- ***Posttests*** **are typically the key outcome measures since the NCLB Act of 2001**
- ***Pretests*** **improve the precision of the impact estimates**
  - ***Critical*** **for clustered designs**
    - **Need 130 schools without pretests**
    - **Need 40-60 schools with pretests**

But Pretests Are Often Collected "Late"

- School staff prefer to wait until routines have been established
- Want enough signed consent forms
- Scheduling issues

Using Late Pretests Could Bias The Impact Estimates

- **Interventions are often implemented _**before**_ school starts**
  - **e.g. Teacher training**
  - **Late pretests could capture “early” impacts**

- **Pretests are sometimes collected later for controls than treatments**

The Late Pretest Problem From a Statistical Perspective

- Involves a ***variance-bias*** tradeoff
  - Late pretests can improve precision
  - But can yield biased posttest impact estimates
    - Positive early effects ⇒ Downward bias
    - Negative early effects ⇒ Upward bias

Paper Addresses Three Research Questions

1. Should late pretest data be collected?
2. What are statistical power losses with late pretests?
3. Is it preferable to obtain "true" baseline data from other sources?
   - School-level test scores from previous years

Outline of Presentation

- Quantifying the variance-bias tradeoff
- Properties of commonly-used posttest estimators with late pretests
- Simulation results
- Conclusions

Quantifying the Variance-Bias Tradeoff

- **Use the *Mean Square Error* for a posttest estimator $\hat{\gamma}$:**

$$\mathrm{MSE}(\hat{\gamma}) = \mathrm{Var}(\hat{\gamma}) + \mathrm{Bias}(\hat{\gamma})^2$$

The MSE Graphically

I

I = True Impact

(U)

U = *Unbiased* Estimator and Confidence Interval

(B I)

B = *Biased* Estimator and Confidence Interval

Key Features of the MSE

- **Smaller values are "better"**
- **Requires estimates of $\mathrm{Var}(\hat{\gamma})$ and $\mathrm{Bias}(\hat{\gamma})$**
- ***Data collection costs are not considered***

Impact Estimation Framework

Two-Level Clustered Designs

- Schools are randomly assigned
  - Treatments: $T_i = 1$
  - Controls: $T_i = 0$
- Students selected within schools
- Pretests are administered in the "fall" and posttests in the spring ("same" test)
  - Pretest values: y_{0ij}
  - Posttest values: y_{1ij}
- Data analyzed at the student level

Two Regression Models

***Posttest* Impact Model: i = Schools; j = Students**

$$(1) \quad y_{1ij} = \alpha_0 + \alpha_1 T_i + (u_{1i} + e_{1ij})$$

ρ_{01} λ_{01} **Errors Are Correlated**

***Pretest* Impact Model:**

$$(2) \quad y_{0ij} = \beta_0 + \beta_1 T_i + (u_{0i} + e_{0ij})$$

α_1 = Average Treatment Effect

β_1 = Early Average Treatment Effect

Bias and Variance Properties of Four ATE Estimators for $\hat{\gamma}_1$

I. Posttest-Only Estimator

1. Regress y_{1ij} on T_i
2. $\hat{\gamma}_{Posttest} = \bar{y}_{1T} - \bar{y}_{1C}$ is unbiased
3. Variance = MSE
4. No variance gains or biases from pretests

II. Differences-in-Differences (DID) Estimator

1. Regress $(y_{1ij} - y_{0ij})$ on T_i
2. Yields $\hat{\gamma}_{DID} = (\bar{y}_{1T} - \bar{y}_{1C}) - (\bar{y}_{0T} - \bar{y}_{0C}) \xrightarrow{p} (\alpha_1 - \beta_1)$
3. Bias = $-\beta_1$
4. But get variance gains from pretests
5. Find $Var(\hat{\gamma}_{DID}) < Var(\hat{\gamma}_{Posttest})$ if $\rho_{01}^2 \geq 0.25$

MATHEMATICA
Policy Research, Inc.

III. Analysis of Covariance (ANCOVA) Estimator

- Regress y_{1ij} on:
  - Treatment status, T_i
  - School-level pretest mean, $\bar{y}_{0i}$
  - Within-school pretest score, $y^{w}_{0ij} = y_{0ij} - \bar{y}_{0i}$
- Statistical complications
  - Pretests correlated with T_i and error terms
- Bias and variance formulas are complex

MATHEMATICA Policy Research, Inc.

IV. Unbiased ANCOVA (UANCOVA) Estimator

1. **Regress y_{1ij} on T_i and $\bar{y}_{0ia}$**
   – **$\bar{y}_{0ia}$ are aggregate test scores from prior years**

2. **$\hat{\gamma}_{UANCOVA}$ is <u>unbiased</u>**

3. **But variance gains are likely to be *less* than using the pretests**

Simulation Results

Simulation Methods

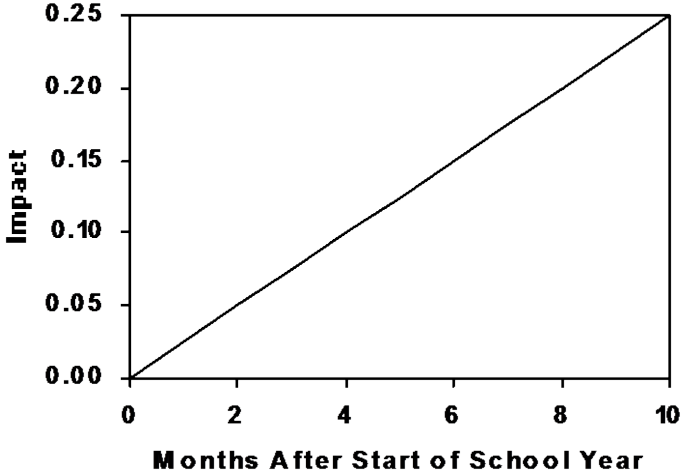
- Calculate *MSE* values for the four estimators
- Use key parameter values found in the RCT literature for achievement test scores
  - ICC values
  - Pretest-posttest correlations
  - School sample sizes

Extent of Bias Will Depend Critically on the Growth Trajectory of Impacts

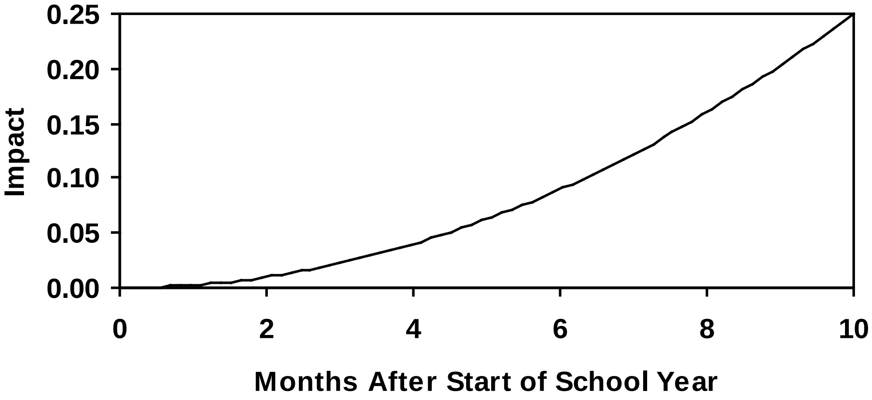
Plausible Trajectories for Test Scores:

Assume Ultimate Posttest Impact Is .25 Stds

Linear Growth



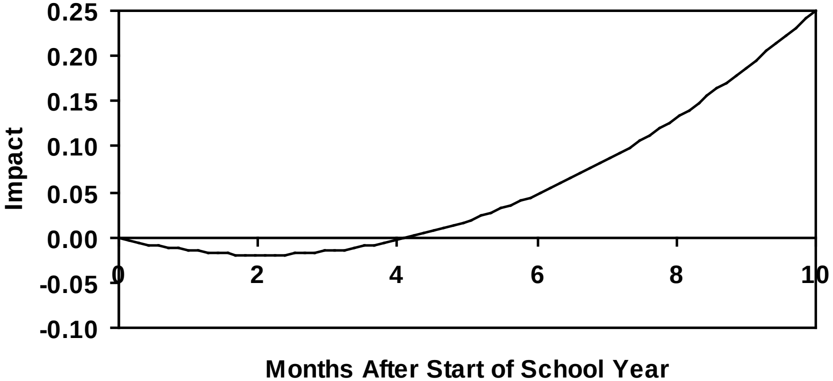
Quadratic Growth



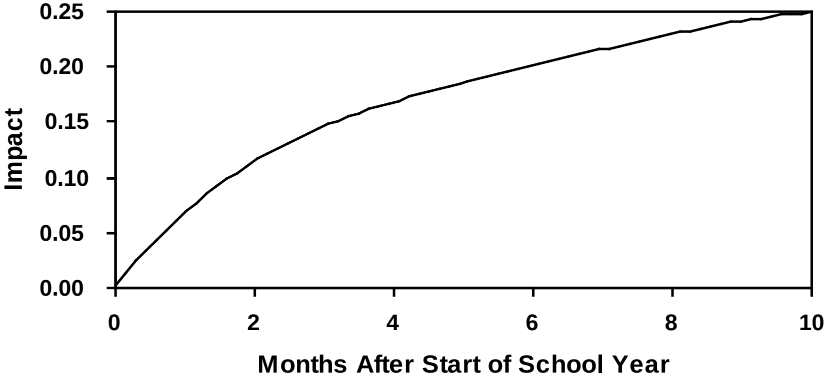
MATHEMATICA
Policy Research, Inc.

Other Plausible Trajectories

Quadratic Growth After Initial Negative Impacts



Logarithmic Growth



MATHEMATICA Policy Research, Inc.

Estimators Using Late Pretests Are Better Than Those That Do Not

- **The ANCOVA and DID estimators will typically yield smaller MSE values than the posttest-only estimator**

- **Holds if early impacts < .10 stds**

- **In graphs, holds if data are collected within:**
  - **4 months if *linear* impact growth**
  - **7 months if *quadratic* growth**
  - **2 months if *logarithmic* growth**

The ANCOVA Estimator Tends to Dominate the DID Estimator

- The ANCOVA estimator yields:
  - Less bias
  - Smaller variance
  - Smaller MSE values

Using Pretests Is Likely to Be Better Than Using *Alternative* Baselines

- **Suppose:**
  - R^2 **= .70 for the ANCOVA model**
  - R^2 **= .50 for the UANCOVA model**
- **ANCOVA is better if early impacts < .065 stds**
- **In graphs, holds if data are collected within:**
  - **2 months if *linear* impact growth**
  - **5 months if *quadratic* growth**

Key Reason We Want to Include Late Pretests

- **Even small gains in R^2 values offset biases**

Need Larger Samples With Late Pretests to Get the Same Power

Early Impact	ANCOVA (R^2 = .7) Required School Sample Sizes To Get an MDE = 0.20
0.00	40
0.02	45
0.04	51
0.06	58
0.08	67

Take Away Messages

- Late pretests are worth collecting if obtained within several months
- Not necessarily true if:
  - Impacts grow very fast
  - Alternative baselines have high R^2 values
  - Data collection costs are considered
- Need larger samples to offset power losses

Other Applications of Methods

- Other RCT areas that collect pretests
  - Early childhood
  - Health
  - Nutrition
- Propensity score matching that uses late pretests for matching