Detecting and Quantifying Variation In Effects of Program Assignment (ITT)

Howard Bloom
Stephen Raudenbush
Michael Weiss
Kristin Porter

Presented to the Workshop on “Learning about and from Variation in Program Impacts?” at Stanford University on July 18, 2016. The presentation is based on research funded by the Spencer Foundation and the William T. Grant Foundation.

This Session

Goal: To *illustrate* and *integrate* key concepts

Topics

- Defining variation in program effects
- Detecting and quantifying this variation

Empirical Examples

- A secondary analysis of three MDRC work/welfare studies
  (59 sites with 1,176 individuals randomized per site, on average)
- A secondary analysis of the National Head Start Impact Study
  (198 sites with 19 individuals randomized per site, on average)

Reference

- Bloom, H.S., S.W. Raudenbush, M.J. Weiss and K. Porter (conditional acceptance) *Journal of Research on Educational Effectiveness*.

Part I

Defining Individual Variation in Program Effects

Distribution of Individual Program Effects

Individual potential outcomes

$Y_i(1) =$ outcome with treatment

$Y_i(0) =$ outcome without treatment

Individual program effect

$$B_i = Y_i(1) - Y_i(0)$$

Population mean program effect

$$E(B) = E(Y(1) - Y(0))$$

Population program effect variance

$$Var(B) = Var(Y(1) - Y(0))$$

Population program effect distribution = ????

Distribution of Individual Program Effects

(continued)

The fundamental barrier to observing a program effect <u>distribution</u> for individuals

- One can only observe an outcome with a program <u>or</u> without the program for a given individual at a given time.
- Hence it is not possible to <u>*observe*</u> individual program effects
- Therefore one can only <u>*infer*</u> a distribution of individual program effects based on assumptions.

The fundamental barrier to estimating a <u>variance</u> of program effects for individuals

- The effect of a program on an outcome variance is not necessarily the same as the variance of the program effects.
- To see this, note that:

$$Y_i(1) = Y_i(0) + B_i$$

$$Var\big(Y(1)\big) = Var\big(Y(0)\big) + Var(B) + 2Cov(B, Y(0))$$

and

$$Var\big(Y(1)\big) - Var\big(Y(0)\big) = \boldsymbol{Var(b)} + 2COV(B, Y(0))$$

Some Implications of Individual Impact Variation For the National Head Start Impact Study

Estimated Parameter	Cognitive Outcome Measure: Receptive Vocabulary (PPVT)	Cognitive Outcome Measure: Early Reading (WJ/LW)
<u>Mean Effect Size</u>		
For full sample	0.15***	0.16***
For lowest pretest quartile	**0.16*** **	0.17***
For other sample members	**0.08***	0.13**
<u>Individual Residual outcome variance</u> (in original units)		
Treatment group	**545*** **	433***
Control group	**667*** **	440***

NOTES: The full sample size varies by outcome from about 3500 to 3700 children and includes both three and four year olds. The statistical significance of individual estimates is indicated as *≤ 10 percent, ** ≤ 5 percent and *** ≤ 1 percent. Estimates that differ statistically significantly across subgroups at the 0.10 level are indicated in bold.

Part II

Defining, Identifying, Estimating and Reporting Cross-site Variation in Program Effects

A Cross-Site Distribution of Mean Program Effects

Theoretical Model

Level One: Individuals

$$Y_{ij} = A_j + B_j T_{ij} + e_{ij}$$

Level Two: Sites

$$A_j = \alpha + a_j$$

$$B_j = \beta + b_j$$

where:

Y_{ij} = the outcome for individual i from site j,
T_{ij} = one if individual i from site j was assigned to the program and zero otherwise,
A_j = the site j population mean control group outcome,
B_j = the site j population mean program effect,
e_{ij} = a random error that varies across individuals with a zero mean and a variance that can differ between treatment and control group members
β = the cross-site grand mean program effect,
b_j = a random error that varies across sites with zero mean and variance $\tau_b^2 = \tau_B^2$
α and a_j = the cross-site grand mean control group outcome and a random error that varies across sites with zero mean and variance τ_a^2, respectively

Some Important Goals of a Cross-Site Analysis

Goal #1

Estimate the cross-site grand mean program effect

Goal #2

Estimate the cross-site standard deviation of program effects

Goal #3

Estimate the cross-site distribution of program effects

Goal #4

Estimate the difference in mean program effects between two categories of sites (the simplest possible moderator analysis).

Goal #5

Estimate the mean program effect for each site

Estimating Impact Variation across Randomized Blocks[1]

Identification strategy

- Randomizing individuals within a "block" to treatment or control status provides unbiased estimates of the mean program effect for each block.
- This makes it possible to estimate program effect variation across blocks.
- Blocks can be studies, sites, cohorts or portions of the preceding.

Important distinctions

- Effects of program <u>assignment</u> vs. effects of program <u>participation</u>
- Variation in <u>effects</u> vs. variation in <u>effect estimates</u>

[1] By definition, randomized blocks have subjects randomized <u>within</u> them. When entire blocks are randomized they typically are called *clusters*.

Cross-site Variation in <u>Impacts</u>
vs.
Cross-site Variation in <u>Impact Estimates</u>

For Impact Estimation

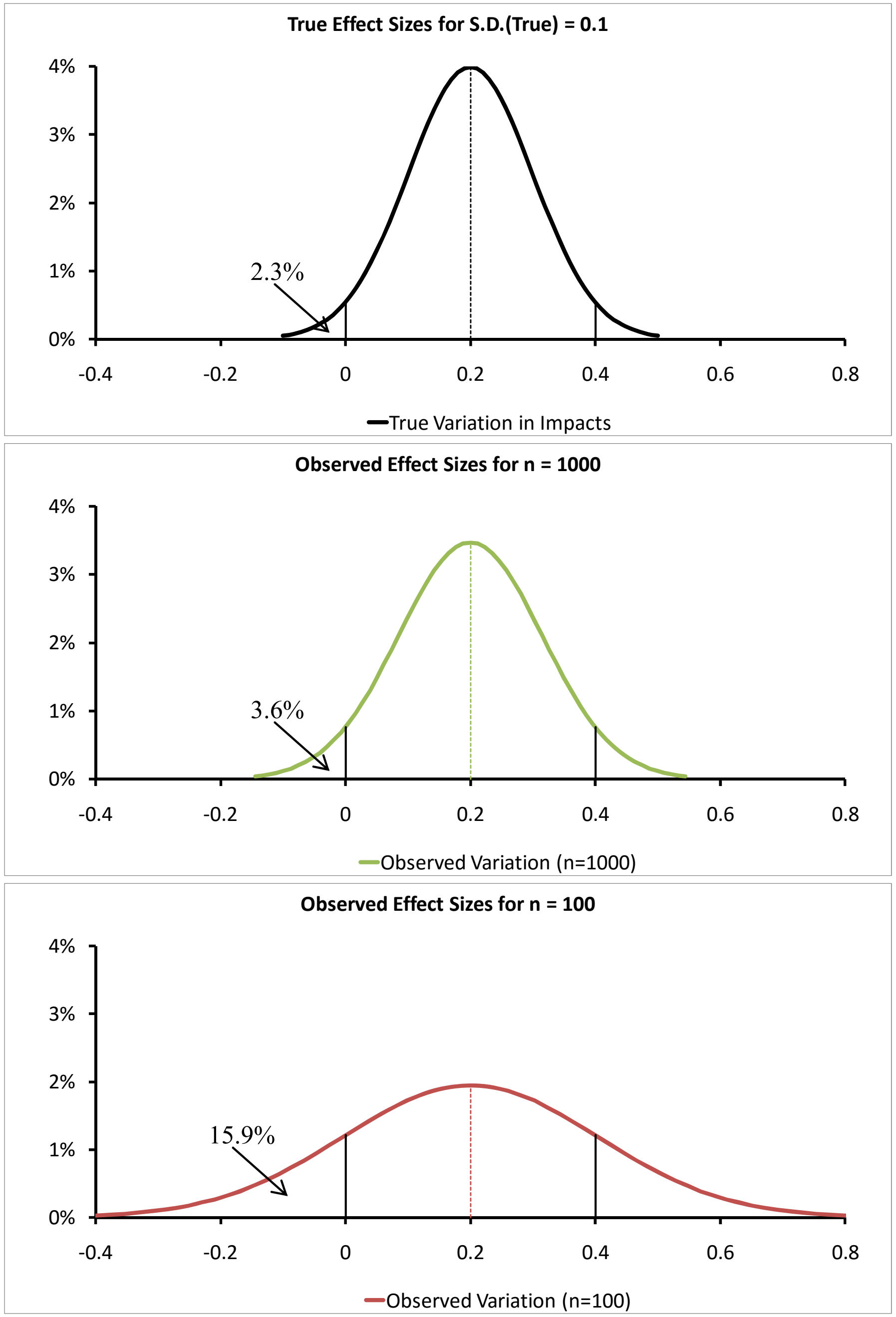
Var(impact estimates) = **Var**(impacts) + **Var**(impact estimation error)

$$= \tau_B^2 + V_j$$

Reliability(impact estimates) = **Var**(impacts)/**Var**(impact estimates)

$$= \frac{\tau_B^2}{\tau_B^2 + V_j}$$

Figure 1


True Effect Sizes for S.D.(True) = 0.1
4%
3%
2%
1%
0%
2.3%
-0.4
-0.2
0
0.2
0.4
0.6
0.8
True Variation in Impacts
Observed Effect Sizes for n = 1000
3.6%
Observed Variation (n=1000)
Observed Effect Sizes for n = 100
15.9%
Observed Variation (n=100)

Estimation Model:
FIRC

Fixed Site-Specific Intercepts, Random Site-Specific Program Effects and Separate Level-One Residual Variances for Ts and Cs (When necessary)

Level One: Individuals

$$Y_{ij} = \alpha_j + B_j \cdot T_{ij} + r_{ij}$$

Level Two: Sites

$$\alpha_j = \alpha_j$$
$$B_j = \beta + b_j$$

Why fixed site-specific intercepts?

- To account for cross-site variation in $\bar{T}_j$ and hence the potential for bias in estimates of $\beta \ and \ \tau_B^2$ due to a possible correlation between $\bar{T}_j$ and α_j

An Alternative Expression of the Impact Estimation Model

Site-Center All Variables

- This is equivalent to specifying fixed site-specific intercepts after one accounts for the degrees of freedom lost when site-centering the dependent variable

Level One: Individuals

$$Y_{ij} - \bar{Y}_{.j} = B_j\left(T_{ij} - \bar{T}_{.j}\right) + r_{ij} - \bar{r}_{.j}$$

Level Two: Sites

$$B_j = \beta + b_j$$

Specify a separate level-one residual variance for Ts and Cs

- Removes potential bias in cross-site variance estimates

How Many Level-One Residual Variances to Estimate?

A Cautionary Tale: Using Data from the *Head Start Impact Study*

- With a separate level-one residual variance for each site there appeared to be a huge amount of cross-site variation in program effects (which was highly statistically significant).
- With a single level-one residual variance for all sites and assignment groups there appeared to be much less cross-site variation in program effects (which was somewhat statistically significant).
- With a separate level-one residual variance for Ts and Cs the results were similar to those for a single variance.

Bottom Line

- Estimating too many variances reduces the sample size for each estimate and thereby increases the uncertainty about those estimates.
- This uncertainty (perhaps counter-intuitively) causes one to understate impact estimation error variance for each site (V_j) and thereby over-state true cross-site impact variation (τ_B^2).

Head Start Impact Study Example Of How Method Matters for Estimating Cross-Site Variation In Effects of Program Assignment

- **Sample size**: 119 centers, 1,056 children from the 3 year old cohort
- **Outcome**: Woodcock Johnson Letter Word Identification test score at the end of the first year after random assignment
- **Issue:** Massive difference in results from two different methods for estimating variation in effects of program assignment

  -**Method #1**: Site centering the treatment indicator for a random Head Start impact model with data pooled across blocks (<u>a single level-one residual variance</u>)

  -**Method #2:** A "split sample" model of Head Start impacts by site combined with a V-Known random-effects meta analysis (a separate level-one residual variance for each site)

Head Start Impact Study Results for Two Estimation Methods (Three-year-old Cohort)

Estimation Approach	Estimated Impact	True Impact Variation (τ)	Chi-sqr stat for τ	P-value
Single centering RE approach	6.071	35.737	125.705	0.296
Split sample + V-known approach	7.746	261.390	421.391	0.000

Key Results to Report From A Cross-Site Analysis Of Program Effects

Results to report

- Estimated grand mean program effect $(\hat{\beta})$
- Estimated cross-site standard deviation of program effects (τ_B)
- Estimated cross-site distribution of program effects (Adjusted Empirical Bayes Estimates)
- Estimated mean program effect for each site (Empirical Bayes Estimates)
- Estimated difference in mean program effects for two categories of sites $(\hat{\beta}_2 - \hat{\beta}_1)$

Empirical Example: MDRC's Welfare-to-Work Studies[1]

Research Design

- Secondary analysis of individual data from three MDRC multi-site randomized trials (GAIN, NEWWS and PI)

Study Sample

- 59 local welfare offices with an average of 1,176 randomized sample members per office (site)

Outcome Measure

- Total earnings (in dollars) during the first two years after random assignment

[1] Bloom, H S., C. J. Hill and J. A. Riccio (2003) "Linking Program Implementation and Effectiveness: Lessons from a Pooled Sample of Welfare-to-Work Experiments," *Journal of Policy Analysis and Management*, 22(4): 551 – 575.

Summary of Welfare-to-Work Parameter Estimates[1]

Estimated Cross-site Grand Mean Program Effect ($\hat{\beta}$)

- Point estimate = $875
- Estimated standard error = $137
- P-value $\leq$ 0.001
- 95 percent confidence interval = $606 to $1,144

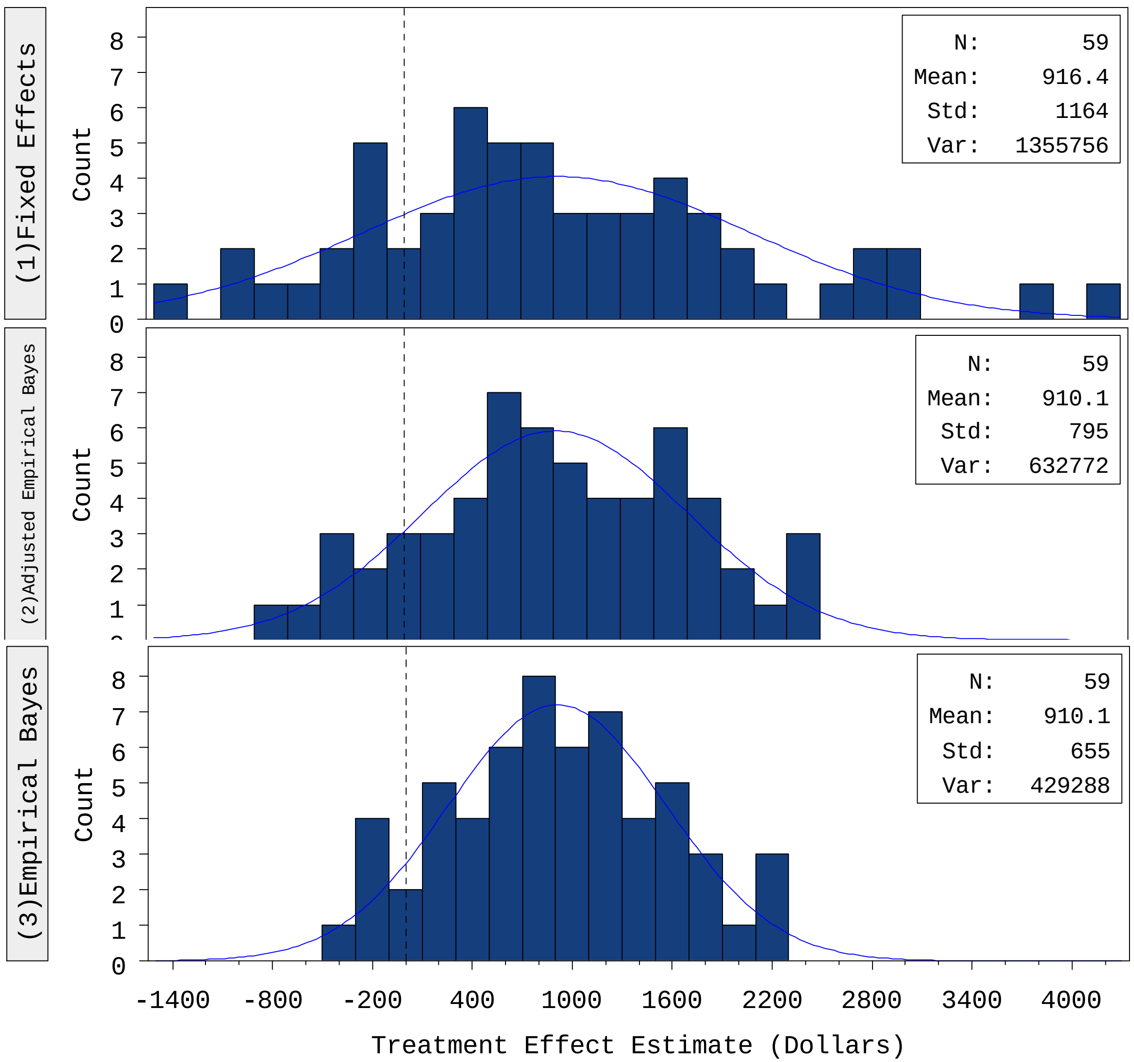
Estimated Cross-Site Standard Deviation of Program Effects ($\hat{\tau}_B$)

- Point estimate = $742
- P-value $\leq$ 0.001
- Asymmetric 95 percent confidence interval = $525 to $1,048

NOTE: Cross-site reliability = 0.497 and σ_T^2/σ_C^2 = 1.09

[1] From Bloom, Raudenbush, Weiss and Porter (under review).

Cross-Site Distribution of Welfare-to-Work Program Effects on Total Two-Year Earnings



Some Important Diagnostics

Assessing the Implications of Uncertainty

- It is important to assess the implications of uncertainty for interpreting one's findings about cross-site variation
- This uncertainty is a function of the study design that produced the findings

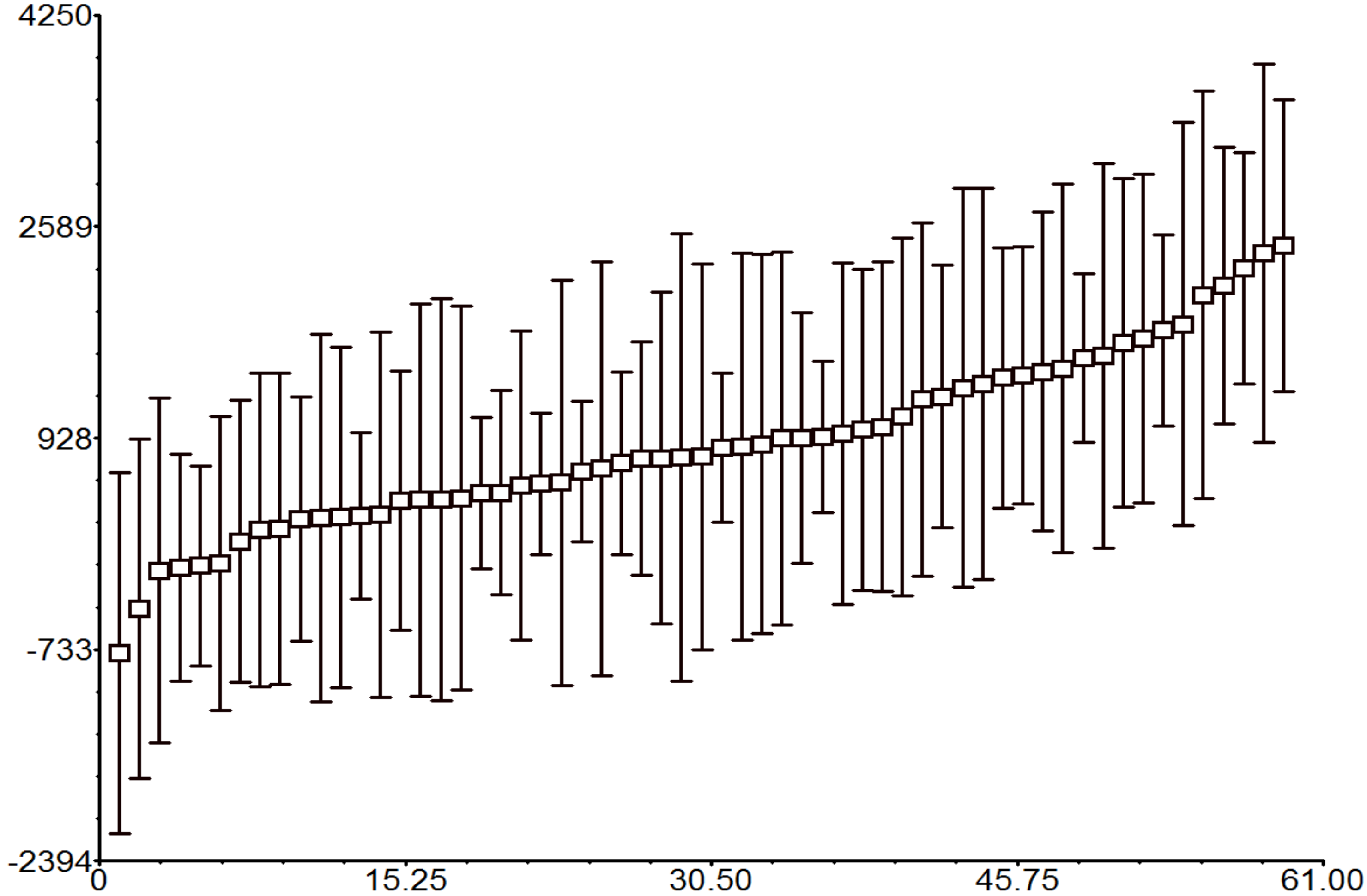
Caterpillar Plots

- graphically report confidence intervals of the OLS or Empirical Bayes estimates of the program effect for each site

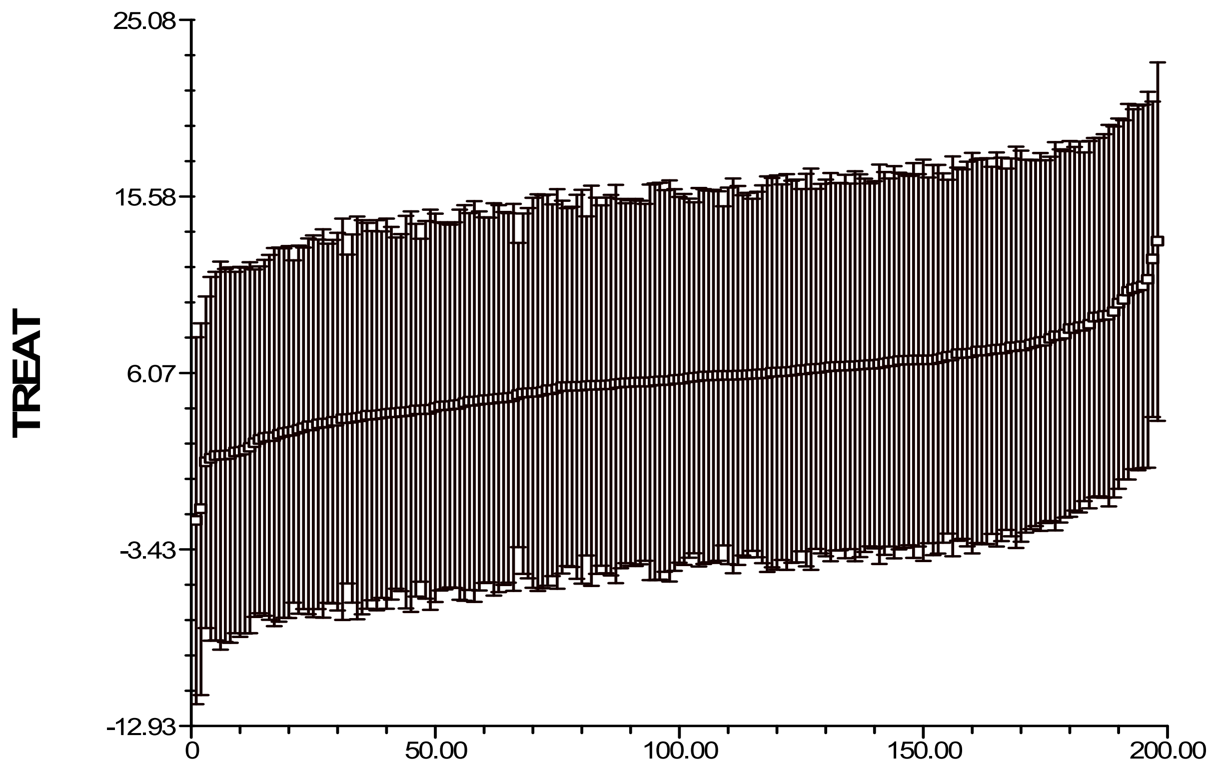
Likelihood Profile Graphs

- Superimpose a graph of the likelihood function for τ^2
- On a graph of the corresponding Empirical Bayes impact estimates for sites

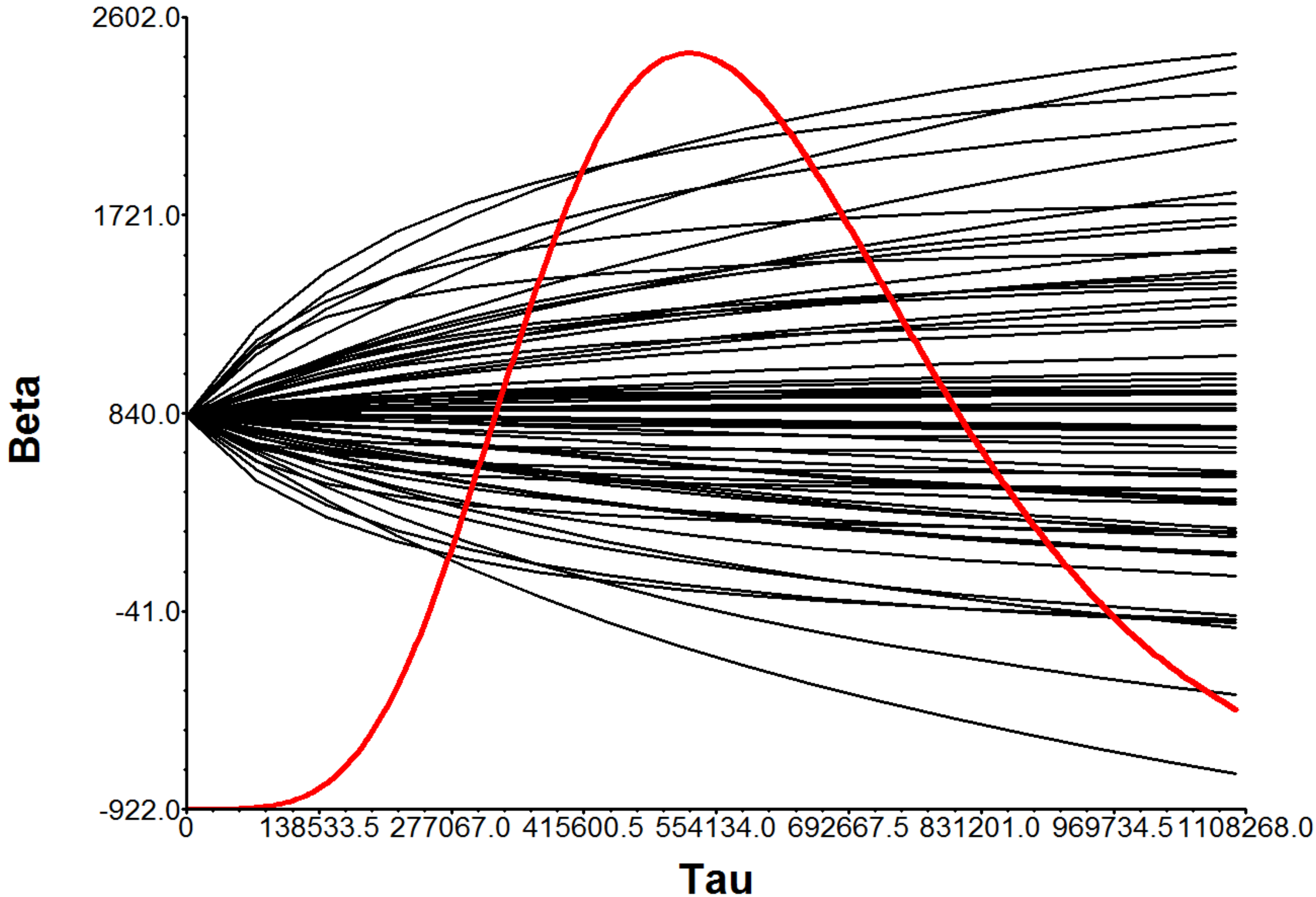
Caterpillar Plot of Empirical Bayes Estimates of Site-Specific Welfare-to-Work Program Effects



Caterpillar Plot For Empirical Bayes Estimates of Head Start Effects on Woodcock Johnson Letter Word Identification Scores



Likelihood Profile Graph for Empirical Bayes Estimates of Site-Specific Welfare-to-Work Program Effects



Profile Likelihood Graph For Empirical Bayes Estimates of Head Start Effects on Woodcock Johnson Letter Word Identification Scores