Quantifying Non-Sampling Variation: College Quality and the Garden of Forking Paths

Eleanor Dillon
Lois Miller
Jeffrey Smith

October 27, 2023

Epigrams - 1

"A game is a series of interesting decisions"
- Sid Meier, Designer of *Civilization!*

"So is an empirical economics paper"
- Us, the authors of this paper

Epigrams - 2

"they're not standard errors, they're fabulous errors, how dare you insult such an icon"

- @crembrulemily

Overview

- Motivation
- Categorizing non-sampling variation
- Existing approaches for dealing with non-sampling variation
- Our approach to non-sampling variation
- Empirical application: effect of college quality

Conceptual motivation: sampling variation

- Sampling variation results from the use of a single random sample from a population rather than the population
- Standard errors capture variability that would arise in estimates from repeated random samples
- Huge recent literature on heteroskedasticity off the diagonal (e.g. clustering)
- Huge recent literature on bootstrapping
- Should this be the only uncertainty we systematically worry about?

Conceptual motivation: population estimates

- Common to report standard errors even when using the population rather than a random sample
  - Many studies using jurisdiction-level data (but may still be sampling variation in jurisdiction-level aggregates if based on survey data)
  - Many studies using register data
- "This is common even in applications where it is difficult to articulate what that population of interest is, and how it differs from the sample." Abadie et al. (2020)
- If pressed, mumble something about super-populations
- What are these super-populations exactly?
- Why not take the sampling theory literally?

Conceptual motivation: design-based inference

- Abadie et al. (2020)
- Characterize the variation due to the "design" conditional on the sample or population under study
- Example: Who gets randomly assigned to treatment holding constant the set of units to be randomly assigned
- SATE versus PATE

Conceptual motivation: synthetic control

- Example: Which unit is treated among the units considered in a synthetic control exercise
- This is *not* sampling variation! Should it appear in the same parentheses below the estimate?
- Are the two examples really the same conceptual thing?

Empirical motivation: CETA evaluations

- CETA (= Comprehensive Employment and Training Act)
- MDTA begets CETA begets JTPA begets WIA begets WIOA
- Dept. of Labor commissions multiple studies by different evaluators using the same underlying CLMS data
- Wildly different impact estimates
  - Common data set $\rightarrow$ Can't be explained by sampling variation
- Barnow (1987) surveys the findings
- Dickinson, Johnson, and West (1987) shows how different study design choices led to different estimates
  - Example: Annual SSA earnings data but monthly enrollment. What is the before period?
  - Example: Definition of the comparison group

Empirical motivation: Heckman and Smith (2000)

- Observe that no one has ever done two experimental evaluations of the same program in the same place at the same time
  - In a sense this is not even possible
- Use data from the National JTPA Study to mimic the variation that could arise across experimental evaluations
  - Example: site selection
  - Example: method for dealing with earnings outliers
  - Example: survey versus administrative earnings measures
  - Example: weighting the 16 sites

Empirical motivation: Black, Daniel and Smith (2005) versus Dillon and Smith (2020)

- Same data set (NLSY-79) and one of the same researchers!
- A surprising number of different design choices
- Some choices that matter:
  - First versus last college attended
  - Trimming outlier values of earnings
  - Conditioning set, especially tract characteristics

Three distinct but related questions

- Produce a preferred single estimate of the parameter of interest
- Understand which design choices matter and which do not
- Characterize the uncertainty resulting from non-sampling variation
  - The second and third items are related but distinct

Current practice: preferred single estimate

- Casual Bayesian based on authorial prior plus some reported sensitivity analyses and (one imagines) many unreported ones
  - Should the preferred estimate receive a weight of one and all others receive a weight of zero?
- Literature surveys of varying degrees of seriousness, depth, and formality
- Formal Bayesian Model Averaging (BMA)
  - Economics example: Durlauf, Navarro, and Rivers (2016)
  - Key question: Whence the prior?
  - What counts as a model for BMA?

Current practice: which design choices matter

- Most common: Sensitivity analysis of a small number of choices
  - Whence the chosen design choices?
  - Typically just two options on each choice
- Meta-analysis (as economists do it)
  - Notable examples: Card, Kluve, and Weber (2010, 2018)

Current practice: characterizing uncertainty from non-sampling variation

- Most common: conduct a small number of sensitivity analyses
  - Whence the chosen design choices?
  - Metric of uncertainty: Are the key conclusions "robust" to each choice?
  - Robustness not well defined either qualitatively or quantitatively
  - Heckman and Smith (2020), doing the sensitivity analyses for the JTPA study
- Implicit variation in design choices across studies
  - End up with literatures with "mixed findings"
  - Few papers attempt to sort out where variation across previous studies comes from
  - Policy discussion overweights the study-of-the-week

Current practice: characterizing uncertainty from non-sampling variation (continued)

- Conventional standard errors that embody variation from model selection
  - Example: Guggenburger (2010) on Durbin-Wu-Hausman tests
  - Example: Belloni, Chernozhukov, and Hansen (2014) in the machine learning literature

Current practice: characterizing uncertainty from non-sampling variation (continued)

- Give the same data to different researchers and see what they do
  - Huntington-Kline et al. (2020)
  - Menkveld et al. (2021)
  - Schweinsberg et al. (2021)
- Could be framed as a way to generate a prior for BMA

Current practice: characterizing uncertainty from non-sampling variation (continued)

- Authors systematically characterize the non-sampling variation
  - Coker, Rudin, and King (2020)
  - Smith (2022) - not me but Gary Smith
- May attempt the full set of leaves, or a random sample of leaves, while the studies on the previous slide sample the leaves non-randomly based on researcher priors
- "Metaverse studies"
  - Is a garden or a metaverse a better metaphor?
- We will do more of this on our paper going forward.

Typology of non-sampling variation

- Measurement of variables
  - Example: Self-reported earnings or UI earnings
- Survey non-response
  - Example: Weighting versus ignorable non-response
- Item non-response
  - Example: Listwise deletion or imputation
- Functional form of estimating equation
  - Example: Logit or probit or LPM

Typology of non-sampling variation (continued)

- Variance estimator
  - Example: Conventional asymptotic approximation or bootstrap
- Population of interest
  - Example: Men or women
- Data set
  - Example: CPS or SIPP
- Identification strategy
  - Example: Conditional independence with different conditioning sets

Thinking about the typology

- Very different types of variation!
- Which ones should vary between papers and which ones should vary within papers?
- Should we describe the variation we consider more formally instead of calling everything a "sensitivity analysis"?

Empirical application: Effect of College Quality

- Estimate the effect of college quality on graduation and earnings outcomes, varying design choices
- Linear model with main effects

$$Y_i = \beta_0 + \beta_Q Q_i + \beta_X X_i + u_i$$

- Q_i = college quality index
- X_i = conditioning variables for CIA

Empirical application: data

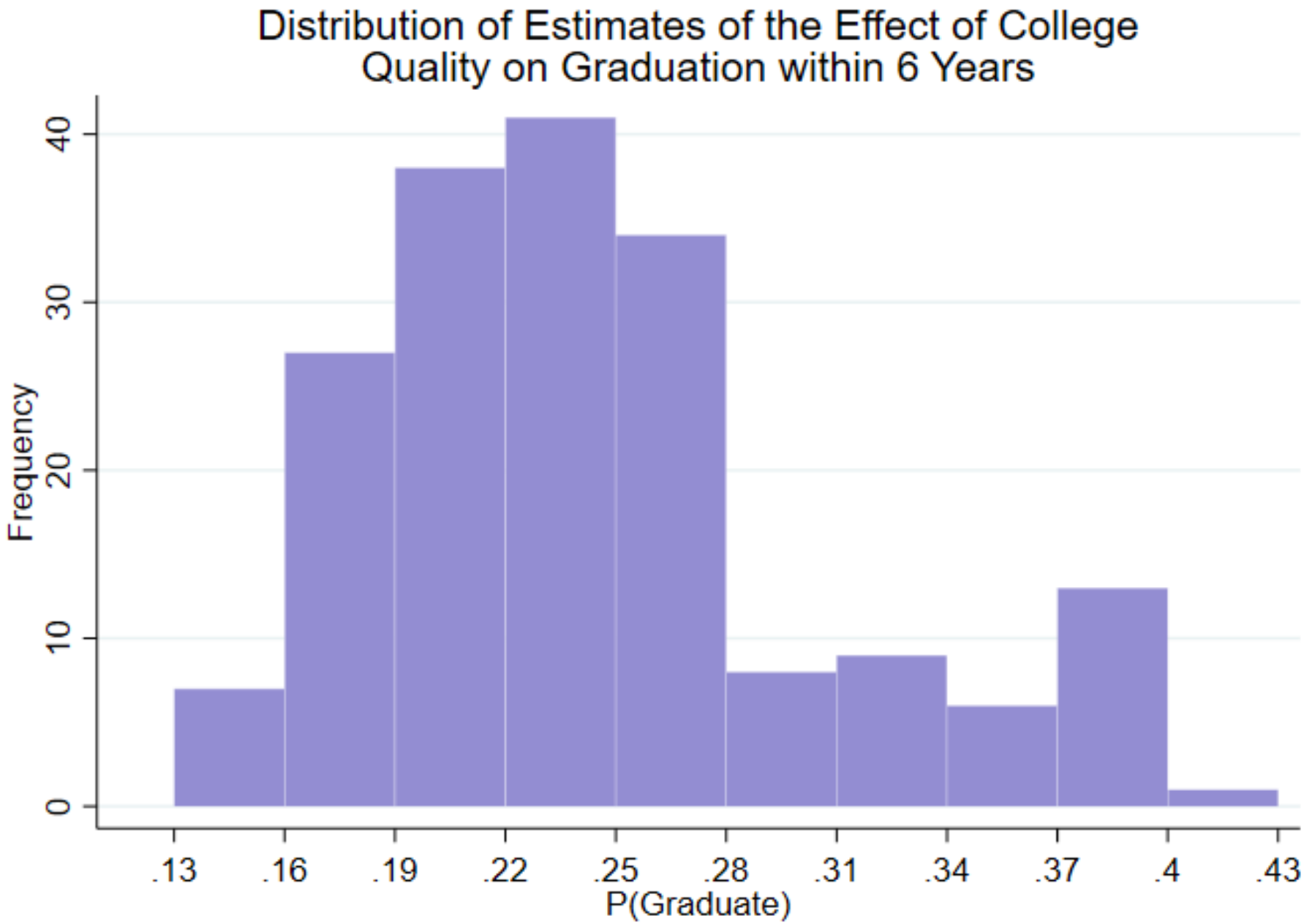
- Use NLSY-97, restrict sample as in Dillon and Smith (2020)
  - Graduated high school or received GED
  - Started at a 4-year college by age 21
  - Interviewed at least 5 years after starting
  - Has valid college quality index
  - Has valid ability measure (i.e. ASVAB)
- Last two restrictions are varied in some of our empirical exercises

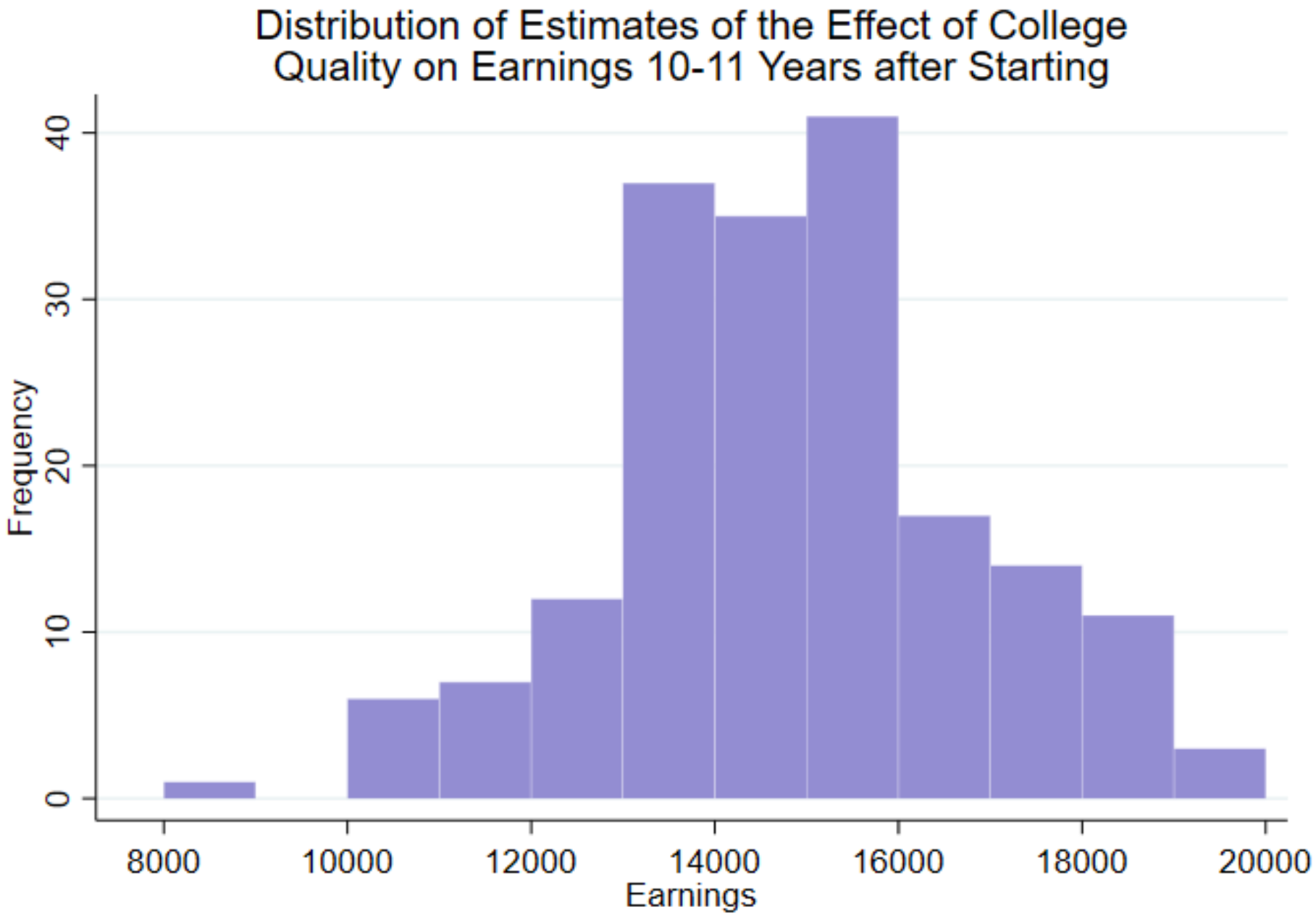
Comparison to Dillon and Smith (2020)

- We drop all college quality-student ability match terms so as to focus on one coefficient
- We focus on two outcomes: graduation within 6 years and earnings measured 10-11 years after staring college
- Estimates of single quality coefficient if using all other design choices from Dillon and Smith (2020)
  - Pr(Graduation): 0.275 (0.046)
  - Earnings: 15,539 (3,168)
- College quality is measured from zero to one, so a 10 percentile increase in quality implies a 2.7pp increase in graduation probability and a $1,553 increase in expected annual earnings

Empirical Application: College Quality Indices

- Follow Black and Smith (2006)
- We create all indices that combine 3, 4, or 5 of the proxies:
  - Pseudo-median SAT (mean of 25th and 75th percentiles)
  - Rejection rate
  - Average salary of faculty engaged in instruction
  - Faculty-student ratio
  - Share of faculty who are tenured or tenure-track
  - Tuition (posted price)
  - Total expenditures per student
  - Instructional expenditures per student
- We also include SAT alone and total expenditures alone
- This gives us 184 indices in total


Distribution of Estimates of the Effect of College Quality on Graduation within 6 Years
Frequency
0
10
20
30
40
.13
.16
.19
.22
.25
.28
.31
.34
.37
.4
.43
P(Graduate)


Distribution of Estimates of the Effect of College Quality on Earnings 10-11 Years after Starting
Frequency
0
10
20
30
40
8000
10000
12000
14000
16000
18000
20000
Earnings

Estimates of Effect of College Quality, Varying College Quality Indices: Summary Statistics

	Mean	SD	Min	Median	Max
Earnings	14,868	2,059	8,493	14,797	19,828
Pr(Graduate)	0.25	0.063	0.14	0.24	0.42

Empirical Application: Item Non-Response

- We use each of the following four ways of dealing with missing data from item non-response
  - Listwise deletion
  - Missing indicators
  - Mean imputation
  - Multiple imputation

- In Dillon and Smith (2020), students were dropped from the sample if they didn't have a valid ability measure (ASVAB score)
- In this paper, we include both ways

Empirical Application: Item Non-Response

- Listwise deletion: Drop any observation that is missing any conditioning variable
- Missing indicators: Change missing to zero and include additional indicator variable for missing each conditioning variable
- Mean imputation: Replace missing values with mean of other observations
- Multiple imputation:
  - Impute continuous varables with linear regression, dummy variables with logit, and categorical variables with multinomial logit
  - 114 replications for graduation, 55 for earnings, following von Hippel (2018)
- Covariates with missing values: HS GPA, SAT, indicators for bad behavior in high school, indicator for living in MSA, HH income quartile, parental education

Estimates of the Effect of College Quality on Graduation within 6 Years, Varying Handling Item Non-response

	Coefficient	Std. Error	Sample size
Do Not Require ASVAB			
List-wise deletion	0.162	0.067	880
Missing indicators	0.297	0.039	2,335
Mean imputation	0.303	0.039	2,335
Multiple imputation	0.272	0.040	2,335
Require ASVAB			
Missing indicators	0.275	0.046	1,857
Mean imputation	0.276	0.046	1,857
Multiple imputation	0.268	0.046	1,857

Thoughts so far

- CQ index composition matters
  - (one of us remembered from looking at this back in the 1990s using the 1979 cohort)
- Listwise deletion is evil
- Uncertainty due to non-sampling variation potentially of the same order of magnitude as sampling variation
  - In a larger dataset (e.g. LEHD) non-sampling variation could easily dominate overall uncertainty
  - Should papers report a "fabulous error" that incorporates non-sampling variation?
  - Taking account of non-sampling variation may change the relative weight placed on different pieces of evidence

Discussion questions

- Are there other related literatures that we are missing out on?
- Should we present descriptive evidence on what is done about non-sampling variation in some well-defined population of published papers?
- Should we recommend changes in common practice in empirical economics inspired by our findings? If so, what changes?
- Two papers?
  - "College Quality and the Garden of Forking Paths"
  - "The Number in Parentheses"

Estimated Effect of College Quality on Annual Earnings, using Various Measures of Earnings

Earnings meausure	Estimated Effect	SE	Sample size
Levels, 10-11 years, include zeros	13,700	3,143	1,713
Levels, 10-11 years, drop zeros	15,043	3,118	1,593
Levels, 9-12 years, include zeros	13,676	2,947	1,772
Levels, 9-12 years, drop zeros	14,759	2,940	1,702
Logs, 10-11 years, drop zeros	16,711	4,068	1,593
Logs, 10-11 years, recode zeros to 1	17,783	13,846	1,713
Logs, 9-12 years, drop zeros	16,561	2,872	1,702
Logs, 9-12 years, recode zeros to 1	13,265	11,155	1,772
IHS transformation, 10-11 years	14,146	12,307	1,713
IHS transformation, 9-12 years	10,939	9,914	1,772