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

Heather Little
Lois Miller
Jeffrey Smith

October 29, 2025

Background

- Education:
  - Washington (Heyne)
  - Chicago (Heckman, Hotz, LaLonde)
- Employment:
  - Western Ontario (1994-2001)
  - Maryland (2001-2005)
  - Michigan (2005-2017)
  - Wisconsin (2018-)
- Key influences on this paper: Heckman, Manski, McCloskey, Gelman
- Grumpy economists

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

"Empirical results hinge on analytical decisions that are defensible, arbitrary, and motivated."

- Simonsohn, Simmons, and Nelson (2020)

Epigrams - 3

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

- @crembrulemily

Overview

- Questions
- Policies
- Defining non-sampling variation
- Existing approaches for dealing with non-sampling variation
- Empirical examples from the literature
- Empirical application: effect of college quality
- Reflections on the path forward

Three related substantive questions

- How do current studies characterize the uncertainty due to non-sampling variation?
- How should researchers characterize the uncertainty arising from non-sampling variation?
- How empirically important is non-sampling variation relative to sampling variation overall?

Public policies ("evaluation policy")

- Data collection (e.g., surveys)
- Administrative data availability and linkage
- Proposal review at NSF / IES / NIH
- Evidence clearinghouse (e.g., WWC, CLEAR) grading of studies

Professional policies

- Pre-analysis plan requirements at journals
- Publication choices by editors and recommendations by reviewers
- Norms around the conduct and reporting of sensitivity analyses

Examples 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

Examples 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

Examples of non-sampling variation in macroeconomics

- Choice of values used for calibration
- Functional forms
  - Example: CES or Cobb-Douglas production function
  - Example: Utility additively separable in consumption and leisure or not
  - Example: Allow savings or not in the model
- Search model details
  - Example: Endogenous search effort or not
  - Example: Allow on-the-job search or not

Thinking about the examples

- Very different types of variation!
- The fuzzy boundary between substantive and design choices
- Which ones should vary between papers and which ones should vary within papers?

Baseline: 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?

Current practice: preferred single estimate

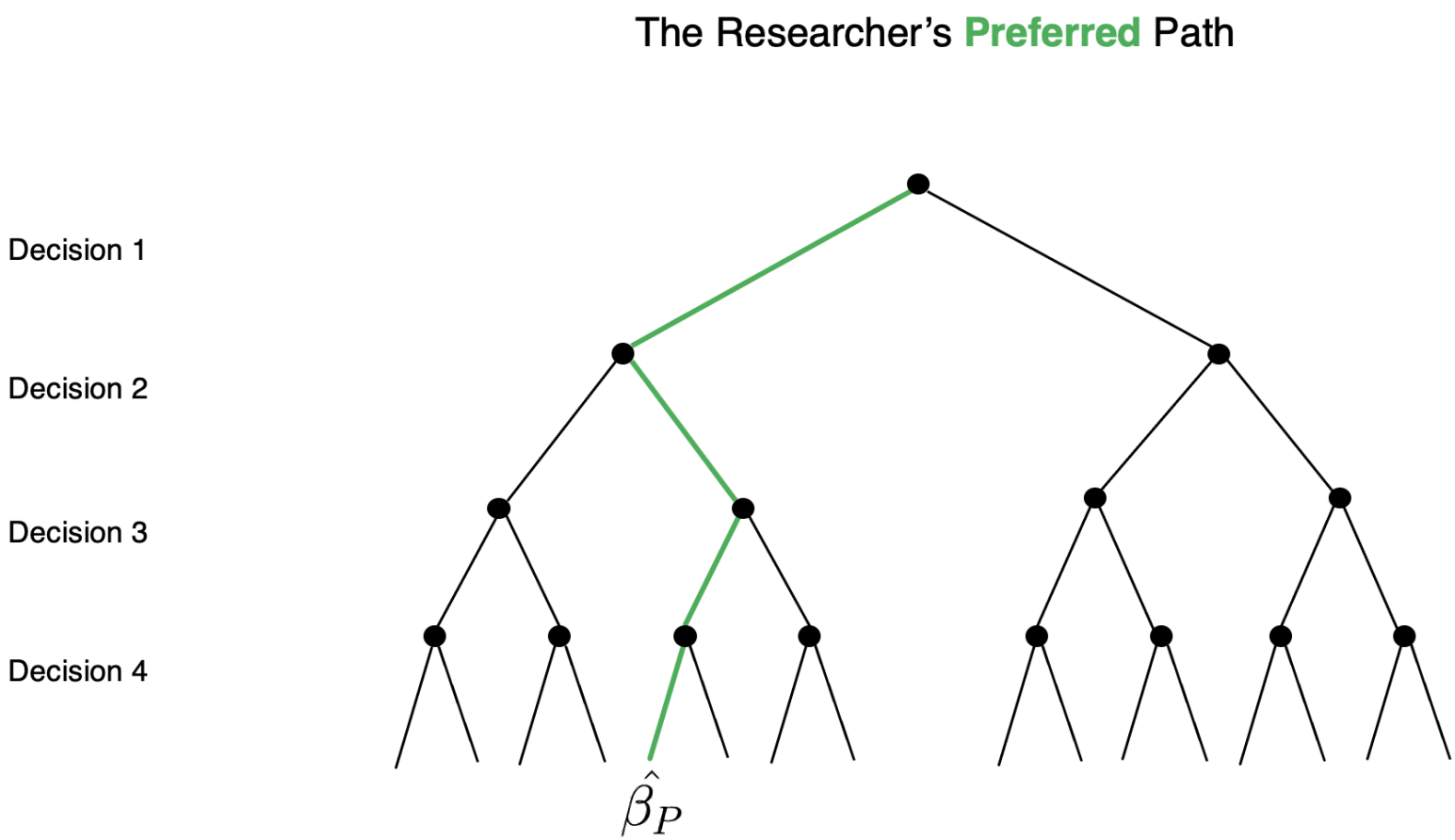
- Casual Bayesian based on authorial prior plus some reported sensitivity analyses and (one imagines) many unreported ones
  - Should the preferred estimate receive a weight of one and all others receive a weight of zero?
- Literature surveys of varying degrees of seriousness, depth, and formality
-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

Current practice: Sensitivity analyses

- "One step away" from preferred design choices
- Often little motivation for design choices considered (and not)
- Often only binary design choices considered
- Metric of uncertainty: Are the key conclusions "robust" to each choice?
- Robustness not well defined either qualitatively or quantitatively

Current practice: Sensitivity analyses (cont'd)

Sensitivity analysis typically changes only one or decision from the researcher's preferred specification, leaving much of the garden unexplored.



Current practice: Implicit variation across studies

- "Many steps away" all at once
- Differences in design choices across studies often poorly documented
- 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: Incorporate model selection in the standard errors

- Example: Guggenburger (2010) on Durbin-Wu-Hausman tests
- Example: Belloni, Chernozhukov, and Hansen (2014) in the machine learning literature
- Here the data drive the design choices, which get made differently in some samples than in others

Current practice: Some selected paths in the garden

- 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 research community weights for various paths through the garden

Current practice: All of the paths in the garden

- Some 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 paths, or a random subset of paths, while the studies on the previous slide sample the paths non-randomly based on researcher priors
- One part of the literature calls these "metaverse studies"
  - Is a garden or a metaverse a better metaphor?
- Another part of the literature plots "specification curves"

Empirical literature: 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 literature: 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 literature: 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

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

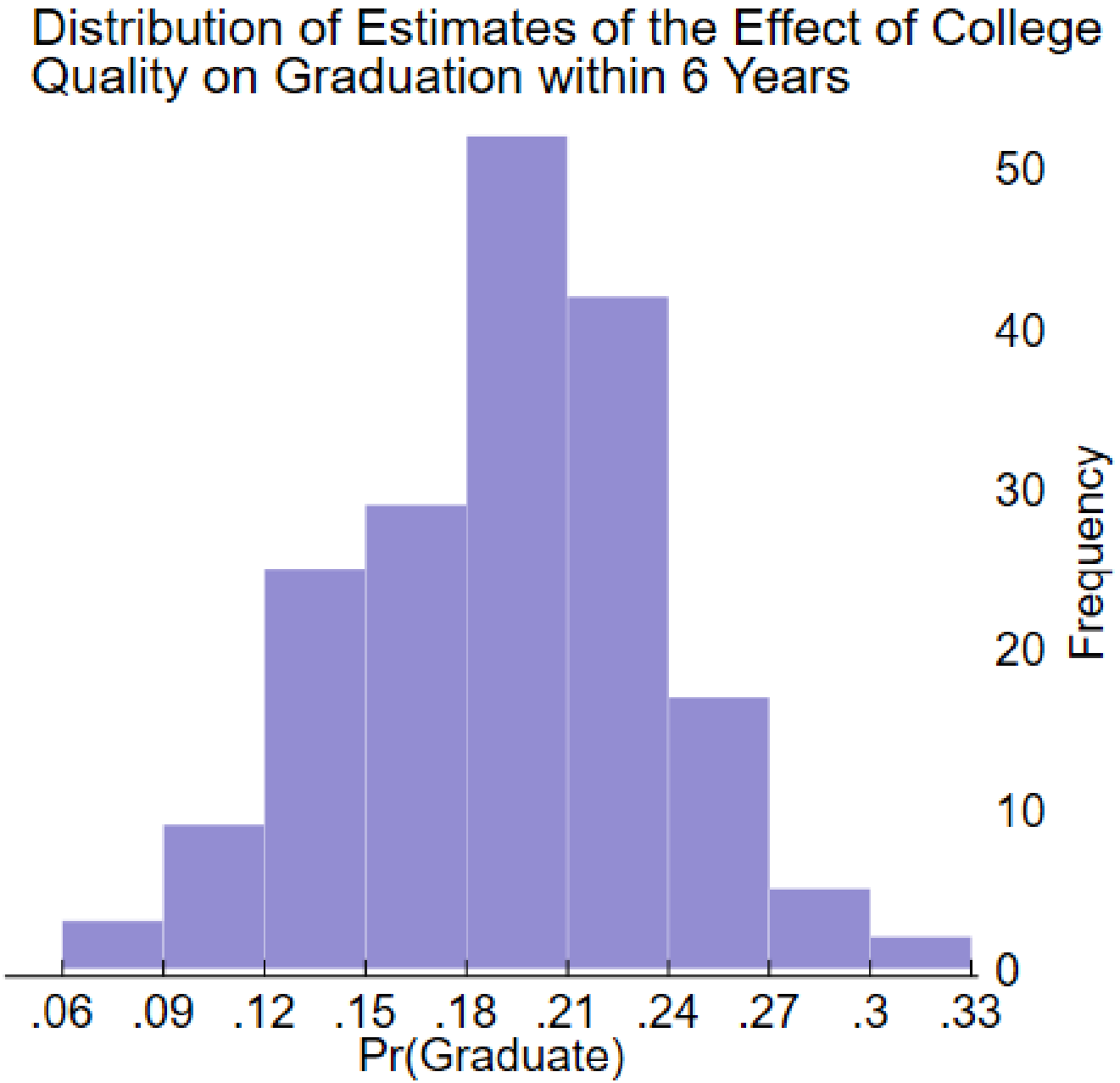
- Use NLSY-97, restrict sample as in Dillon and Smith (2020)
  - Graduated high school or received GED
  - Started at a 4-year college by age 21
  - Interviewed at least 5 years after starting
  - Has valid college quality index
  - Has valid ability measure (i.e. ASVAB)
- Last two restrictions are varied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

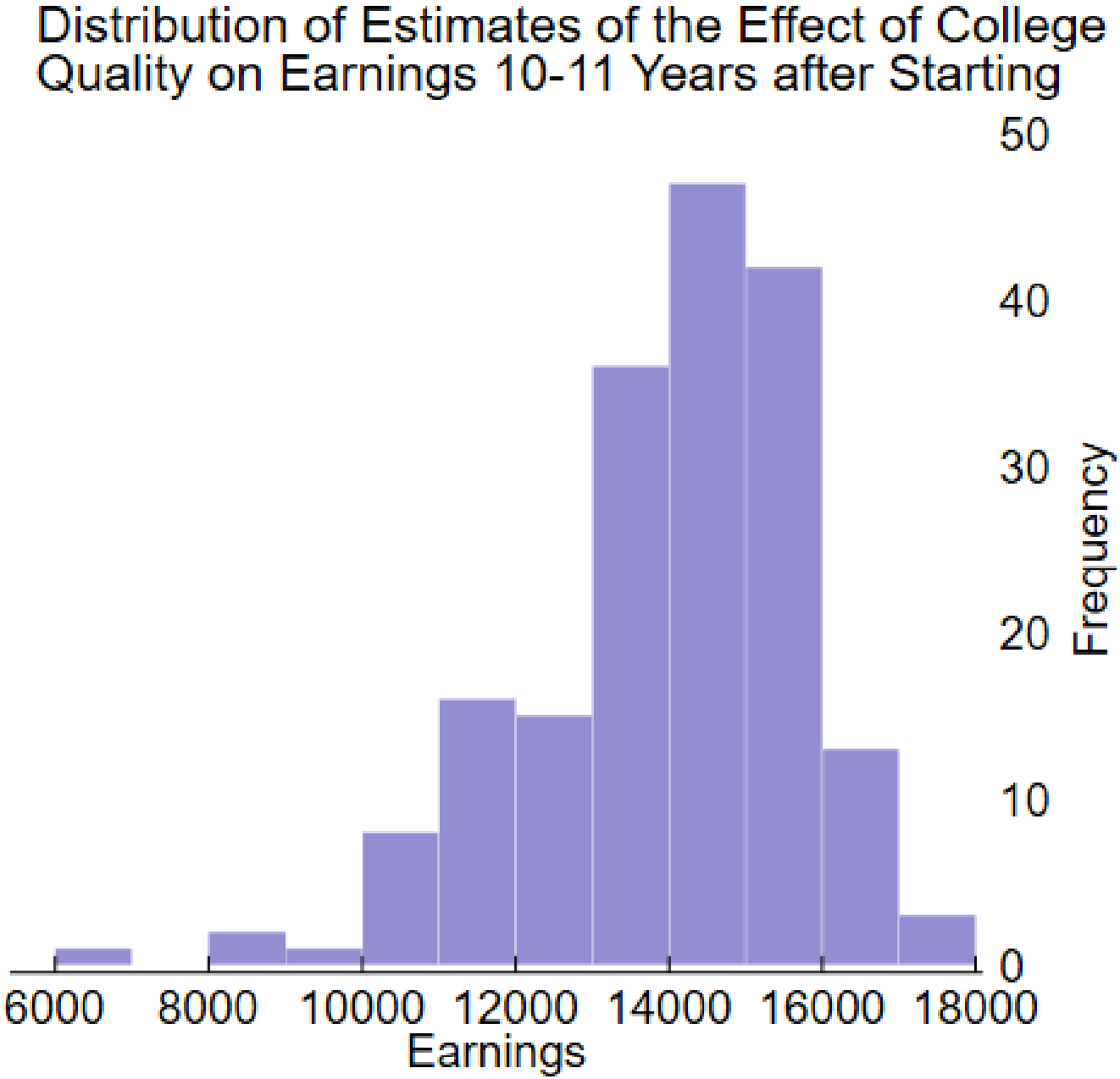
Empirical application: Relation to Dillon and Smith (2020)

- 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): 0.223 (0.048)
  - Earnings: 16,726 (3,376)
- College quality is measured from zero to one, so a 10 percentile increase in quality implies a 2.2pp increase in graduation probability and a \$1,673 increase in expected annual earnings

Empirical Application: College Quality Indices

- Follow Black and Smith (2006) approach and measurement model
- 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
50
40
30
20
10
0
Frequency
.06
.09
.12
.15
.18
.21
.24
.27
.3
.33
Pr(Graduate)


Distribution of Estimates of the Effect of College Quality on Earnings 10-11 Years after Starting
50
40
30
20
10
0
Frequency
6000
8000
10000
12000
14000
16000
18000
Earnings

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

Outcome	**N**	**Mean**	**SD**	**Min**	**Max**
Graduation (All)	184	0.192	0.047	0.069	0.308
Earnings (All)	184	14,022	1,839	6,195	17,570
Graduation (4+ Proxies)	126	0.197	0.042	0.105	0.308
Earnings (4+ Proxies)	126	14,479	1,520	8,644	17,570

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

	Graduation within 6 Years			Annual Earnings		
Method	**Effect**	**SE**	**N**	**Effect**	**SE**	**N**
Listwise deletion	0.144	0.068	774	18,963	5,384	681
Missing indicators (require ASVAB)	0.223	0.048	1,565	16,716	3,376	1,352
Missing indicators	0.247	0.041	1,964	16,026	3,038	1,672
Mean imputation (require ASVAB)	0.225	0.048	1,565	16,792	3,408	1,352
Mean imputation	0.253	0.041	1,964	16,535	3,042	1,672
Multiple imputation (require ASVAB)	0.217	0.048	1,565	16,393	3,402	1,352
Multiple imputation	0.231	0.042	1,964	16,042	3,068	1,672

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

Earnings Measure	Estimated Effect	SE	Sample Size
Levels, 10–11 years, include zeros	14,999	3,397	1,444
Levels, 10–11 years, drop zeros	16,716	3,376	1,352
Levels, 9–12 years, include zeros	14,488	3,103	1,498
Levels, 9–12 years, drop zeros	15,968	3,109	1,441
Winsorized 99th, 10–11 years, include zeros	14,506	3,325	1,444
Winsorized 99th, 10–11 years, drop zeros	16,189	3,291	1,352
Winsorized 95th, 10–11 years, include zeros	10,442	2,670	1,444
Winsorized 95th, 10–11 years, drop zeros	12,120	2,604	1,352
Winsorized 99th, 9–12 years, include zeros	13,875	2,988	1,498
Winsorized 99th, 9–12 years, drop zeros	15,350	2,988	1,441
Winsorized 95th, 9–12 years, include zeros	10,674	2,375	1,498
Winsorized 95th, 9–12 years, drop zeros	12,570	2,338	1,441
Logs, 10–11 years, drop zeros	16,517	3,653	1,352
Logs, 10–11 years, recode zeros to 1	9,843	15,266	1,444
Logs, 9–12 years, drop zeros	16,522	3,315	1,441
Logs, 9–12 years, recode zeros to 1	4,162	12,056	1,498
IHS transformation, 10–11 years	7,519	13,142	1,444
IHS transformation, 9–12 years	2,816	10,501	1,498

Estimated Effect of College Quality on Graduation within 6 Years and Annual Earnings, Using Various Sets of Control Variables

Functional Form	**Graduation within 6 Years**	**Annual Earnings**
Baseline from Dillon + Smith	0.223	16,716
All covariate main effects	0.222	16,602
All covariate main effects + lit-based interactions	0.223	16,073
LASSO linear terms only, CV	0.211	16,855
LASSO linear terms only, adaptive	0.211	16,855
LASSO linear terms only, plugin	0.277	23,612
LASSO linear terms only, BIC	0.271	21,709
LASSO include interactions, CV	0.195	17,931
LASSO include interactions, adaptive	0.195	18,450
LASSO include interactions, plugin	0.256	21,816
LASSO include interactions, BIC	0.256	18,572

Summary Statistics for Estimates of the Effect of College Quality on Graduation and Earnings, Varying All Applications

Outcome	N	Mean	SD	Min	Max
Graduation within 6 years	202	0.195	0.047	0.069	0.308
Annual earnings	220	14,218	2,450	2,816	23,612

Standard errors (i.e., just sampling variation) from the Dillon and Smith (2020) path through the garden: 0.048 and 3,376.

Empirical application: Lessons

- CQ index composition matters
  - One of us remembered this from looking at this back in the 1990s using the 1979 cohort.
- Listwise deletion is evil
- Tread lightly when transforming earnings as the dependent variable
- Estimates less affected by the specification of the conditioning variables than one might expect
- 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

Reflections

- How much do "one step away" sensitivity analyses miss?
- Would pre-analysis plans solve the problem (or at least add value)?
- Should papers report a "fabulous error" that incorporates non-sampling variation from a metaverse study?
  - How to avoid implicitly penalizing studies with richer data and therefore more quantifiable design decisions?
  - How to weight the various different paths in the garden?
  - What if the number of paths from a given fork is not well-defined?
- Taking account of non-sampling variation may change the relative weight placed on different pieces of evidence
- Step 1 of the 12

Going forward in 2026

- More literature
- More "one step away" sensitivity analyses
- More depth on the full metaverse analysis
- "fabulous errors"

Thanks!

Please reach out if you have other questions or comments:

econjeff@ssc.wisc.edu