CALIFORNIA STATE UNIVERSITY
FULLERTON
MIHAYLO COLLEGE
OF BUSINESS AND ECONOMICS

Department of Economics
Working Paper Series

2020/006

The Influence of Hidden Researcher Decisions in Applied Microeconomics

Nick Huntington-Klein, CSU Fullerton
Andreu Arenas, University of Barcelona & IEB
Emily Beam, University of Vermont
(continued next page)

CALIFORNIA STATE UNIVERSITY, FULLERTON
Mihaylo College of Business and Economics, *Department of Economics*
800 N. State College Blvd., Fullerton, CA 92834-6848 / T 657-278-2228 / F 657-278-3097

Marco Bertoni, Padova University
Jeffrey R. Bloem, University of Minnesota
Pralhad Burli, Idaho National Laboratory
Naibin Chen, Pennsylvania State University
Paul Greico, Pennsylvania State University
Godwin Ekpe, Northern Illinois University
Todd Pugatch, Oregon State University
Martin Saavedra, Oberlin College
Yaniv Stopnitzky, University of San Francisco

March, 2020

CALIFORNIA STATE UNIVERSITY, FULLERTON
Mihaylo College of Business and Economics, *Department of Economics*
800 N. State College Blvd., Fullerton, CA 92834-6848 / T 657-278-2228 / F 657-278-3097

The Influence of Hidden Researcher Decisions in Applied Microeconomics

Nick Huntington-Klein[a,*], Andreu Arenas[c], Emily Beam[c], Marco Bertoni[d], Jeffrey R. Bloem[e], Pralhad Burli[f], Naibin Chen[g], Paul Greico[h], Godwin Ekpe[i], Todd Pugatch[j], Martin Saavedra[k], Yaniv Stopnitzky[l]

[a]California State University, Fullerton
[b]University of Barcelona & IEB
[c]University of Vermont
[d]Padova University
[e]University of Minnesota
[f]Idaho National Laboratory
[g]Pennsylvania State University
[h]Pennsylvania State University
[i]Northern Illinois University
[j]Oregon State University
[k]Oberlin College
[l]University of San Francisco

Abstract

Researchers make hundreds of decisions about data collection, preparation, and analysis in their research. We use a many-analysts approach to measure the extent and impact of these decisions. Two published causal empirical results are replicated by seven replicators each. We find large differences in data preparation and analysis decisions, many of which would not likely be reported in a publication. No two replicators reported the same sample size. Statistical significance varied across replications, and for one of the studies the effect's sign varied as well. The standard deviation of estimates across replications was 3-4 times the typical reported standard error.

Keywords: Replication, Metascience, Research

JEL: C81, C10, B41

*Corresponding Author. Email: nhuntington-klein@fullerton.edu. Address: 800 N. State College Blvd., Fullerton, CA, 92831. Additional thanks to Andrew Gill, Seth Gitter, and Virginia Wilcox.

I. Introduction

A primary goal of empirical work in the social sciences is to generate results that are believable, generalizable, and replicable. However, estimated results for the same research question can vary considerably from study to study, and across different specifications within studies.

Results should naturally vary due to the use of different samples or empirical methods. However, variation may also come from the hundreds of decisions made in the process of analysis, from data cleaning to variable definition to model design. The wide range of decisions behind every analysis can be thought of as "researcher degrees of freedom."

Only some fraction of these choices can be described explicitly in the paper. Peer reviewers may not be capable of judging all of these choices, readers may not be aware how these choices are made and so be able to incorporate them into their understanding of the results, and revisiting these choices after publication occurs rarely, if it is even possible. And these choices are necessarily made at some point in the process, even if analysis does not change after viewing the results, so the is issue distinct from "p-hacking" or "The Garden of Forking Paths" (Gelman and Loken, 2014; Silberzahn et al., 2018), and would not be solved by the use of preregistered analysis.

In this paper we attempt to measure the extent of variation in these researcher degrees of freedom in the context of applied microeconomics studies that attempt to isolate a causal effect. We take a "many-analysts" approach where multiple researchers use the same data set to answer the same research question. This allows us to examine how different the choices made in data manipulation and analysis are between researchers, and to examine the impact of these choices on the eventual results.

This study is a part of a modern metascientific literature in the social sciences with a particular concern for the quality of results. Such metascientific studies in economics sometimes examine published work either through the lens of its unconsidered statistical properties (e.g. Ioannidis et al., 2017; Young, 2018). More often these checks operate through attempts to replicate published papers, either focused on individual papers, or on many at

once (Dewald et al., 1986; Camerer et al., 2016; Chang and Li, 2017).

Replication studies allow us to determine variation in results that arises from the use of different samples, when testing the same hypothesis in new data, or from major analysis choices, when testing the same hypothesis in the same data with a new method (Hamermesh, 2007; Clemens, 2017; Christensen and Miguel, 2018). These replication studies can also reveal variation in results that arise from errors in code, when attempting a "pure" replication (Hamermesh, 2007) to reproduce the original tables and figures. The most well-known example of the latter is likely the attempt by (Herndon et al., 2014) to replicate Reinhart and Rogoff (2010), which uncovered a coding error in the original study.

Replication studies in economics are relatively rare (Hamermesh, 2017; Berry et al., 2017), and the incentives for performing them are not well-aligned (Duvendack et al., 2017; Gertler et al., 2018; Mueller-Langer et al., 2019). These kinds of replication will also have difficulty in uncovering variation in results that is driven by researcher degrees of freedom. Pure replications will intentionally make the exact same choices as in the original study unless a clear error is spotted, and replications using new data or methods generally attempt to make the same choices except for the specific data set or method being changed, so as to isolate the source of any difference. Incentives to replicate, which favor results that overturn an original study (Dewald et al., 1986; Hamermesh, 2007; Gertler et al., 2018), do not favor looking into these choices. Even if results are found to be sensitive to researcher degrees of freedom, as long as the original choices are not obviously incorrect, it is difficult to make a convincing case to an editor that the results have been overturned.

These researcher degrees of freedom, however, may have significant impact on the results of a study even if the choices made are not wrong, but are simply one reasonable option of many. In psychology, Simmons et al. (2011) find that researcher flexibility in experimental design and data analysis allow nearly any hypothesis to be supported. Lenz and Sahn (2017) find that, in a major political science journal, 30-40% of significant results would have been insignificant if certain covariates were excluded from the model, and that the inclusion of

these covariates was generally not justified in the text.

One approach to probing researcher degrees of freedom is the "many-analysts" approach, in which many analysts use the same data set to answer the same research question, without knowledge of the methods used by other analysts. Looking at methodological variation between researchers for whom publication is not contingent on their results allows for an idea of how much variation in results can be driven by good-faith decision-making among informed researchers.

Many-analysts studies include Silberzahn et al. (2018), which recruited psychological researchers to examine whether a data set of referee calls in soccer showed evidence of discrimination against darker-skinned players. Analysts were very different in the way they constructed their models and also in their results. Botvinik-Nezer et al. (2019) recruited teams to examine a data set of fMRI imaging data on subjects playing a gambling task to evaluate nine ex-ante hypotheses. Again, results, and support for each of the hypotheses, differed widely across teams. Both of these are preceded in publication outside of academia by Cohn (2016), who had five pollsters evaluate the same poll for the 2016 US presidential election, again with widely varying results.

In this paper we apply the many-analysts approach to the kinds of causal inference estimates commonly generated in applied microeconomics. Unlike previous many-analysts studies, we allow researcher freedom in the actual construction of the data set. The processing and cleaning of administrative, governmental, or otherwise externally generated data is a common feature of applied microeconomic research, and is a likely source of researcher degrees of freedom.

We use two studies published in high-quality journals as a basis, and produce seven replications of each study. We find considerable variation both in results and in the construction of data. No two replicators ended up with the same sample size, and in several cases large differences in sample construction were driven by decisions that may have likely gone unmentioned in an eventual publication, or at least overlooked by readers. Analysis decisions

also showed large differences. Most analytic differences, like the use of linear probability models vs. logit, would have likely been mentioned in publication for reviewer scrutiny, but many of them, like the construction of bins when generating a control for education, would likely be overlooked by a reader.

The actual effect on results was mixed. In one of the studies, six of the seven replications had very similar point estimates and overlapping confidence intervals. In the other, results varied much more widely, with both significant positive and negative coefficients.

II. Methods

The methods for this study include (i) selecting papers, and analyses within those papers, to replicate, (ii) designing instructions and information to present to replicators, (iii) recruiting replicators, and (iv) evaluating replicator work.

II.i. Selecting Replication Tasks

Project organizers developed a list of desirable attributes for studies to replicate. These included:

- Studies should be published in well-regarded economics journals,[1] with a preference for publication in the last 20 years.
- Studies should produce a single causal estimate of interest.
- Studies should not be so well-known that replicators are likely to recognize them from the instructions.
- Studies must use publicly-available data, and ideally data that microeconomists would be used to working with. Because organizers anticipate that most replicators will be

[1] The IDEAS/RePEc Aggregate Rankings for Journals (https://ideas.repec.org/top/top.journals.all.html) list was used as a guide.

American, public American data sets are favored.

- If studies rely on highly specific domain knowledge or obscure methods that replicators would not know on their own, it should be simple enough to explain to replicators in instructions, or be secondary to analysis so it can be removed for a simplified replication.

- The two studies selected should be from different subfields of applied microeconomics.

Because the easiest of these criteria to use in a literature search is that the studies use publicly-available data, organizers used Google Scholar to search for the names of publicly available data sets commonly used in applied microeconomics research, including, generically, "Census", as well as the National Longitudinal Survey of Youth, the National Education Longitudinal Study, the Panel Study of Income Dynamics, the American Community Survey, and the Current Population Survey. We did not require studies to use one of these data sets.

In search results, studies that were published in top economics journals, and seemed likely to satisfy the other criteria based on the title and abstract, were examined more closely.

After examining 50 papers, five were selected for full examination as being strong candidates for satisfying all criteria: Black et al. (2008), Fairlie et al. (2011), Moretti (2000), Bernal and Keane (2011), and Lochner and Moretti (2004). From these five, Black et al. (2008) and Fairlie et al. (2011) best satisfied all criteria.[2]

Given these two studies, we isolate and simplify the analyses to be given to replicators. More detailed instructions are in the next section.

Black et al. (2008) is a study of the effect of compulsory schooling on teenage pregnancy. The authors use variation in compulsory schooling policy in two environments - the United

[2]Moretti (2000) relies on a non-public version of its data set and so was not selected. Analysis in Bernal and Keane (2011) relies on specific data-cleaning and analytic steps that would be difficult to relate to replicators without removing the possibility for researcher choice. Lochner and Moretti (2004) and Black et al. (2008) are very similar studies, and from these two Black et al. (2008) was selected out of concern that the Lochner and Moretti (2004) study, with more than 1,000 citations, was well-known enough to be recognized by some replicators.

States and Norway - and estimate the effect of those changes on teenage pregnancy rates. They then attempt to distinguish whether the effect operates by improving human capital or through the "incarceration effect."

Keeping in mind that the goal of this study is not to test the robustness of the original results, we base replication instructions on a simplified version of the analysis. We focus on their primary analysis of the United States, which uses US Census data from 1940 to 1980, calculates women's age at first birth, and excludes apparent births age 14 or below. State- and decade-level variation in compulsory schooling laws identifies the effect of compulsory schooling. Instructions are based on a replication of the top half of Black et al. (2008) Table 2, column 3, where "birth by age 18" is the outcome variable. We simplify their analysis by looking at only one compulsory schooling margin - whether the state has a compulsory schooling age of 16 or higher, as opposed to 15 or lower.

Fairlie et al. (2011) is a study of the effect of employer-based health insurance on entrepreneurship. The authors use variation in age to identify the effect. Men aged 65 or older qualify for Medicare, and those aged 64 and 11 months or younger generally do not. Medicare reduces the need for employer-based health insurance, and so authors look for a jump in entrepreneurship at age 65 exactly. Current Population Survey data is used to identify men who turn 65 within the four months they are included in the survey.

Instructions are based on one of their analyses, which is shown in Fairlie et al. (2011) Table 6. Men who can be observed having just turned 65 are compared to those observed just under 65 in terms of the rates of self-employment, conditional on being employed at all. We simplify the task for replicators by narrowing the sample window from 1996-2006, as in Fairlie et al. (2011), to May 2004-December 2006. This avoids combining data across samples where variable definitions have changed.[3]

[3]While the goal of this paper is not to judge the quality of the original findings being replicated, in preparing instructions for this study we happened to find that the Fairlie et al. (2011) analysis is sensitive to the decision of which years of data to include in the analysis. The effect using their original analysis can even reverse sign depending on the time period used. We did not check if this issue extends to the other analyses in that paper. However, this is an excellent example of results being affected by seemingly

II.ii. Replication Instructions

We construct sets of instructions for each replication. The goal of these instructions is to ensure that each replicator knows what the data set and research question of interest are, as well as some identifying assumptions, without restraining their choices too much. Replicators were encouraged to perform each analysis as if they were writing their own paper for publication. The full text of each set of instructions is in Appendix A.

For both sets of instructions, replicators are told to use any statistics package, and that they should use assistants if they would normally use assistants in their work. They are also told that their analysis should be independent, and should not attempt to identify the original study, or to match (or mismatch) with fellow replicators. The goal is to uncover "how you would estimate this effect, if you'd had this question, this idea for identification, and had chosen this particular sample."

They are also told to focus on a single "headline" result, of the kind that might be reported in an abstract. Replicators were not directed to perform robustness checks and alternate analyses.

Instructions for the Black et al. (2008) replication direct replicators to download Census 1% files from IPUMS for 1940-1970, and the 5% files from 1980. IPUMS provides data files in already mostly-usable format, and the different census years are already appended together. Data should then be limited to female subjects aged 20-30.

Replicators are given the background theory that compulsory schooling laws may reduce the incidence of teenage pregnancy for a number of reasons, and told to estimate the effect of compulsory education age in a state on the proportion of women in that state who have a teenage pregnancy, under the identifying assumption that trends in teen pregnancy are unrelated to the decision to change compulsory schooling policy.

Removing several potential sources of researcher degrees of freedom, replicators are given

innocuous researcher choices.

the definition of a teenage pregnancy as "having a child by age 18," and to determine the compulsory schooling law being applied as the law in place in the mother's birth state when they are 14 years of age. A table of compulsory schooling laws by state and decade, from Black et al. (2008), is given to replicators in Word format, and for women who turned 14 between policy years, they are told to use the most-recent policy. Replicators are also told to look specifically at the margin of compulsory schooling at age 16, comparing policies requiring students to stay until they are 16+ against policies requiring some age 15 or below.

Instructions for the Fairlie et al. (2011) direct replicators to download Current Population Survey (CPS) monthly files from the National Bureau of Economic Research (NBER) for the months of May 2004 through December 2006. The use of NBER individual monthly files, rather than pre-compiled and combined files from, for example, Integrated Public Use Microdata Series (IPUMS) means that replicators will have to import the raw files and combine them into a single data set, a data-cleaning task in which there may be different researcher decisions made.

Data should then be limited to male subjects who can be observed "in the exact month that they turn 65," meaning subjects observed both at the ages 64 and 65 in the four-month CPS rolling panel.

Replicators are given the background theory that employer-provided health insurance may be a barrier to entrepreneurship, and given the background information that Medicare eligibility occurs at exactly 65 years of age for most people. It asks for the effect of Medicare eligibility on the rate of self-employment, conditional on being employed at all. They are given the shared identifying assumption that nothing else of importance changes between the ages of 64-and-11-months and 65.

II.iii. Replicator Recruitment

Replicator recruitment began in May 2018. There were two main methods for requesting participation from replicators and directing them to the sign-up website.[4] In both cases, to improve recruitment success, the recruitment message stressed that the replication project was designed to reduce the time necessary to complete the task.

First, we used the U.S. News and World Report ranking of economics departments to develop a list of 138 top economics departments. We sent an email to the chair of each department, asking them to forward on a recruitment message to their faculty, or to only the applied microeconomists. We do not know how many department chairs complied with this request.

Second, we posted a message on Twitter asking for interested researchers to sign up. The link from the tweet to the recruitment website was clicked 638 times.

On the recruitment form, replicators were asked whether they had any published or forthcoming work in applied microeconomics, whether they had performed replication work before, whether they typically used student assistants, whether they would want to complete one replication or two, and what their typical fields of interest were. The pool of replicators was intended to represent people actually producing applied microeconomics research, and so recruitment was limited to those with published or active work in the field.

In total, 51 qualified researchers signed up to complete a replication, 37 of which came from Twitter and 14 from email.

Replication tasks were assigned to replicators first on the basis of field of interest. If replicators indicated that their primary fields of research were relevant to one of the replication tasks but not the other, they were assigned to that task. Replicators who listed topics relevant to neither or both tasks were randomly assigned. Replicators who agreed to do both replications were assigned their first task in the same way. Replicators were not given

[4] https://sites.google.com/view/replication2018

information from organizers about who else had been recruited, or results from any of the other replicators. The initial due date for replication was the end of January, 2019, or seven months after recruitment. This due date was eventually pushed back to March, 2019.

Of the original 51 qualified researchers, 12 finished a replication: 10 finished one replication each, and two finished two replications each, for a total of seven completed replications of each task. Four of the successful replicators had been recruited by email, and eight had been recruited from Twitter. Project organizers, who knew the content of the original studies, did not contribute any replications. In one case organizers provided assistance to a replicator who was having difficulty importing data files. Upon completion, replicators were asked to complete an exit survey. When those who had signed up to provide a replication but did not complete one gave reasons for not finishing, they reported almost uniformly that they did not have the available time they had expected to work on the project, and their decision was unrelated to the content of the task.

II.iv. Analysis

Replicators return to the organizers their raw data files, code for data processing and analysis, and a primary result of interest. Organizers then perform a descriptive analysis of the results and code.

Analysis proceeds first by taking the produced analyses and comparing them in absolute terms, analyzing the degree of overlap between analyses as well as in terms of features like included controls and sample size.

Organizers were able to successfully replicate the reported results of all replicators using the provided code, with one exception, where due to version control issues a line of code included in analysis was omitted from the submitted code. Code was later updated to the final version, after that replicator viewed the Results section and notified organizers.

Then, organizers analyzed the submitted code of each replicator line by line. This allowed organizers to code the decisions made by each replicator in the process of cleaning the data

and generating variables for inclusion in analysis.

Results consist of a description of the differing decisions made by replicators, and the implications of those decisions for sample construction and analysis.

III. Results

In total, there were fourteen completed replications: seven for the compulsory schooling and teenage pregnancy study based on Black et al. (2008), and seven for the health insurance and self-employment study based on Fairlie et al. (2011). We do not compare the replication results to the original study because matching the original study is not the goal, and the instructions were not designed to exactly match the original study.

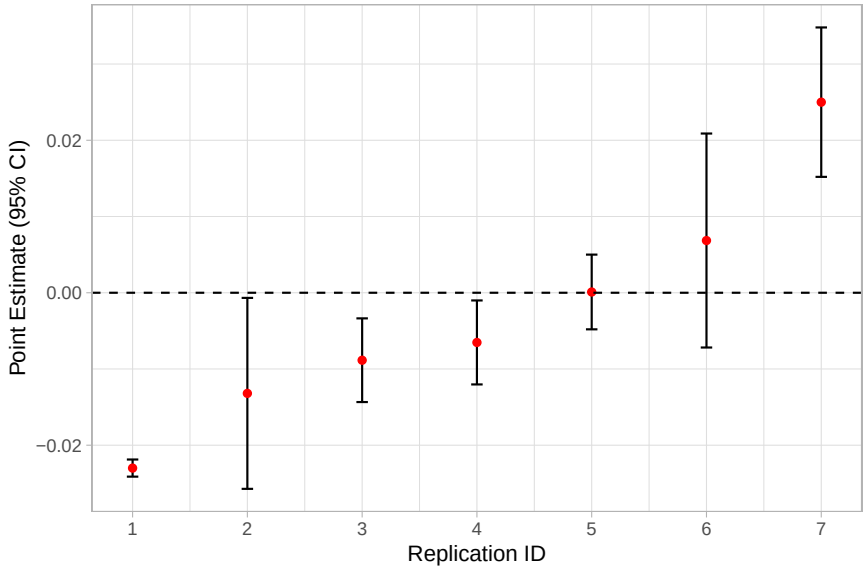
Figures 1 and 2 show the confidence intervals estimated in each study for the preferred estimates that replicators selected, based on reported point estimates and standard errors.[5] Many replicators performed additional analyses or robustness tests, but we will focus on the estimators they reported as preferred, which in the instructions were said to be the result that would be put in the abstract if these were individual studies being written up.

Results based on the reported point estimates are mixed. Estimates in the compulsory schooling study vary widely across replications. Four are statistically significant at the 95% level and negative, one is statistically significant and positive, and two are insignificant, one of which has a point estimate very near zero.

The results imply that different researchers answering the same question using the same data set may arrive at starkly different conclusions. Three (of seven) would likely conclude that compulsory schooling had a negative and statistically significant effect on teen pregnancy, two would find no significant effect, and one would find a positive and significant effect. This variation in results demonstrates the potentially crucial role played by researcher

[5] In some cases, coefficients from nonlinear models were reported in the original replications; in these cases we calculate marginal effects after the fact for comparability.

Figure 1: Results from Compulsory Education Study

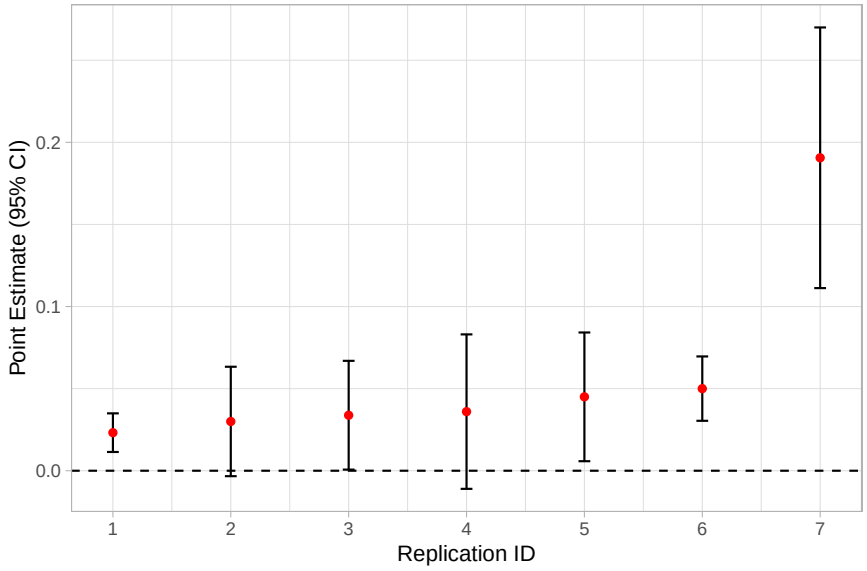


degrees of freedom in applied microeconomics research.

Estimates in the health insurance replication are more consistent. One replication has a markedly larger estimate than the others, and its confidence interval does not overlap with any others. Among the other six, while no two estimates are the same, point estimates are within a fairly narrow band, and all of them have overlapping confidence intervals. Statistical significance does vary across replications. Even in this case, with largely overlapping confidence intervals across replications, five researchers (of seven) would likely conclude a significant effect of employer-based health insurance on entrepreneurship, while the other two would find no evidence of this effect.

Differences between the point estimates are not the only matter of interest. We are also interested in the extent to which choices made by replicators differed in the ways they put together their data and designed their analyses, and how these choices affected the differences in results.

Figure 2: Results from Health Insurance Study



III.i. Compulsory Schooling Study

In this section we examine the general study procedures that each replication in the compulsory schooling study took, the ways in which they constructed their data, and the decisions made during analysis.

III.i.i. Compulsory Schooling Study Procedures

There were seven completed replications of the compulsory schooling study. All seven were completed by the primary replicators, and none reported using assistants. All seven reported that they were familiar with the methods used in the replication, and four of the seven reported that the topic was similar to the work they normally do. All seven replications were completed in Stata. The number of replications was not large enough to look for systematic differences between replicators based on their characteristics.

III.i.ii. Compulsory Schooling Study Data Construction

Table 1 shows the steps taken in data construction in the compulsory schooling study. In all cases, replicators gathered US Census Data Files from IPUMS and limited the data to adult

women subjects born in states with available policy information, as instructed.

Aside from these shared decisions, replicators made different decisions about cleaning the data. The instructions said that women with first births age 14 or below should be dropped. However, one study instead dropped those with first births age 13 or below. Another kept these women but coded them as not being teenage pregnancies. One study did not limit the data to women aged 20-30, as instructed. One study did not match policy dates to individuals in the exact way described in the instructions. Afterwards, one replicator reported having made these decisions because they misread the instructions. Other cases may be due to thinking the differing decisions were more appropriate.

Policy information was provided to replicators. Six replicators left a strange Ohio policy point in the data. In the original study, the table of compulsory schooling laws is 18 for Ohio in every year but 1944, when it is 8. This may be a typo rather than a real policy. We carried this 8 through in the version of the table given to replicators. One replicator (6) changed this to 18 in their main analysis, although several others did point out that it was odd, and said that they might have changed it if they were truly working on their own, but thought that might go against the instructions.

Replicators were instructed to perform the analysis as though they were designing it themselves, and so naturally some data construction decisions not in the instructions are made differently across replicators. One replicator each made the decision to drop subjects in group quarters, to drop the second 1970 census sample, to drop women who never had a child, and to keep only families where the woman was head or spouse to the head with an in-house child. There was also some minor variation in variable definitions, with one study counting women giving birth at age 18 exactly as not having a child by age 18.

These sample construction decisions led to different sample sizes from every replicator. No two replications had the same sample size, although most are similar. The smallest sample size is 831,139, driven by dropping women without children and one of the census samples. The largest is 4,271,245, driven by including women outside the age range of 20-30.

All other samples are fairly similar but not exactly the same, ranging from 1.64 million to 1.70 million.

The small differences in sample size may still be important. Even among the five studies with similar sample sizes, the point estimates vary widely, and even the sign is not consistent. Some of this may be due to differences in analysis rather than differences in sample construction. To account for this, we run the same basic two-way fixed effects model on all seven constructed data sets with no sample weights or other controls.[6] This reduces differences between estimates, suggesting that some of the differences are due to analysis rather than sample construction. But important differences remain, and the sign is still not consistent. Of particular interest are replications 1 and 7, which do not differ in sample construction in any obvious way, and which likely would have reported identical data construction procedures if these were real studies, but for which sample sizes differ by about 4,000. When using identical models they have similarly-sized estimates of opposite signs.

In the case of this compulsory schooling study, replicators did not perfectly agree on the proper approach to constructing the sample. Some of the differences between approaches, such as dropping women who never had children, would have been reported in a research paper and so could have been evaluated by a reader. Others may not have been. Studies here that are identical on all data-construction decisions we checked still result in different sample sizes and different point estimates of the treatment effect.

[6]We use the Stata command regdhfe with "child by age 18" as the dependent variable, allowing the definition to be different for replication 5, as in Table 1. Being treated is the only independent variable. Birth year and state are absorbed fixed effects.

Table 1: Sample Creation and Shared-Variable Definition Decisions in Compulsory Schooling Study

Decisions Affecting Sample Size: (* indicates that this is specified in instructions)							
Study	1	2	3	4	5	6	7
US Census Data Files from IPUMS*	X	X	X	X	X	X	X
Adult women subjects only*	X	X	X	X	X	X	X
Ages 20-30 only*	X	X	X	X	X		X
Drop women with first-birth age 14 or below*	X	X	X			X	X
... Drop women with first-birth age 13 or below				X			
... Women with first-birth age 14 or below coded as not being teenage births					X		
Excludes states without policy information (AK, HI)	X	X	X	X	X	X	X
Excludes anyone living in group quarters					X		
Drops all observations from second 1970 Census sample		X					
Drops women without children ever		X					
Keeps only household heads or spouse to household head, with in-house child related to household head						X	
Decisions Affecting Shared Variable Definitions:							
Changed strange-looking Ohio policy data point						X	
Matches policy years to individuals as in instructions*	X	X		X	X	X	X
Counts age-18 births as "child by age 18"	X	X	X	X		X	X
Sample size used in estimation	1,664,643	831,189	1,696,522	1,701,516	1,669,105	4,271,245	1,640,645
Point estimate	-0.023	-0.0132	-0.0089	-0.0065	0.0001	0.0068	0.025
Point estimate under simple shared model (fixed effects for state and birth year, nothing else)	-0.0164	-0.0229	-0.0089	-0.0113	-0.0086	0.0103	0.0177

III.i.iii. Compulsory Schooling Study Data Analysis

The compulsory schooling study, by design, is based on the concept of a policy that is administered at the state level, such that those in a given birth cohort are exposed, or not exposed, to a certain level of compulsory schooling based on the state they are born in. Except for the fact that treatment does not change monotonically over time within state, this is similar to a difference-in-difference setup. Accordingly, six of the seven replicators used a regression model with two-way fixed effects for state and birth cohort, the standard approach to estimating a difference-in-difference setup with multiple treated groups and variation in treatment timing. The seventh (replication 1) uses two-way fixed effects with state and year of observation.

While no two replicators performed the exact same analysis, all seven replicators made very similar choices in performing the analysis. In addition to all seven using a regression model with state fixed effects and a second set of time fixed effects, and all seven clustered standard errors at the state level (replication 5 additionally clustered at the birth year level). Despite a binary dependent variable, all seven used ordinary least squares rather than logit or some other nonlinear model. None of the seven used recent developments in difference-in-difference estimators or standard error adjustments. However, many of these developments (e.g. Goodman-Bacon, 2018) were very new at the time replications were performed, and the analysis in question is not exactly the same as difference-in-differences.

The main points of difference between the analyses were whether the second set of fixed effects should be for birth year or year of observation, the Stata commands used to estimate the model, the choice of additional control variables, and the use of sample weights. The two-way fixed effects model was most commonly estimated using the reghdfe command (replications 1, 3, 5) and regress (2, 4, 6), with 7 using areg.

Choice of control variables varied considerably. Table 2 shows the choice of control variables in each replication. As previously mentioned, all studies include state fixed effects and all but one include fixed effects for birth year. Four studies additionally control for the

Table 2: Control Variables Included in Compulsory Schooling Study

Study	1	2	3	4	5	6	7
State fixed effects	X	X	X	X	X	X	X
Birth year fixed effects		X	X	X	X		
Race		X		X	X		
Year fixed effects	X			X		X	X
Age fixed effects						X	X
Year-by-age fixed effects							X
State linear time trends					X		
Spouse is household head						X	
Person sample weights used		X				X	X
Point estimate	-0.023	-0.0132	-0.0089	-0.0065	0.0001	0.0068	0.025
Point estimate under prepared data from Replication 4	-0.0393	-0.0082	-0.0112	-0.0065	0.0028	-0.0280	-0.0069

time of observation in some way, either with age or year fixed effects as in 1, 4, 6, and 7. Three studies (2, 4, and 5) include dummies to control for race. One study (5) includes prior time trends by state.

Because all studies used the same design, differences in point estimates can only be driven by differences in data construction or the choice of controls or regression command. To see how much variation is left in point estimates after accounting for data construction differences, we fix the cleaned data to be that from replication 4, chosen arbitrarily from the seven.

After restricting data to be the same, differences between replications remain (see the final row of Table 2). This indicates that the choice of control variables, even when selecting across different sets that all may seem reasonable, still has a meaningful effect on the published coefficient.

III.ii. Health Insurance Study

In this section we examine the general study procedures that each replication in the health insurance study took, the ways in which they constructed their data, and the decisions made during analysis.

III.ii.i. Health Insurance Study Procedures

There were seven completed replications of the health insurance study. Four were completed by the primary replicators, two (2 and 3) were completed with graduate student assistance, and one (1) had most coding done by a graduate student assistant. Six of the replicators (other than replicator 1) reported that the statistical work was similar to the work they normally do, with two of those reporting that the topic was similar to the work they normally do. Six of the replications were completed in Stata, and one (6) was completed in R. The number of replications was not large enough to look for systematic differences between replicators based on their characteristics.

III.ii.ii. Health Insurance Study Data Construction

Table 3 shows the data construction decisions made by replicators working on the Health Insurance study. In all cases, replicators used monthly NBER CPS files from May 2004 to December 2006, limited to men only. Since CPS subjects are interviewed four months in a row, this will produce a small-T rolling panel data set with approximately four observations per individual.

After this point, data construction procedures diverge. The biggest point of divergence is in defining the age range. The instructions specify that subjects should be "observed in the exact month that they turn 65." However, replications 1-3 and 6 include a wider range of subjects in the data set. The widest range is in replication 1, which includes subjects aged 54-76. After the fact, replicators reported their reasoning for this decision. Two reported misreading the instructions, and the other two reported that they thought their age range

choice was more appropriate. Similarly, the instruction that subjects must be employed was implemented as instead being in the labor force in replication 4. This replicator reported that this was due to a misreading of the instructions, but that they would have likely made the same choice if writing their own paper.

Replicators were instructed to perform the analysis as though they were designing it themselves, and so naturally some data construction decisions not in the instructions are made differently across replicators. In particular, replicators implemented different kinds of checks on the plausibility of the data. Some dropped individuals with inconsistent data, or who did not appear all four times in the CPS sample, or who were missing income data. Replicators also differed on whether they defined self-employment status using the first worker-class variable in the data, or using both worker-class variables.

The sample sizes differ between the replications, and no two replications have the same sample size. The biggest reason for this is the choice of age ranges, which would have been reported if these replications were written in their own studies. However, even if all age ranges are narrowed to match the instructions, sample sizes are still, in order, 3,543; 5,604; 5,212; 2,493; 1,628; 4,322; and 2,016 (mean 3,545, standard deviation 1,567).

Despite the large differences in sample sizes, the differences in effect sizes are much smaller here than for the compulsory schooling study.[7] Replication 7 is the only outlier. However, the differences are large enough that some results are statistically significant at the 95% level (1, 3, 5, 6, 7), while others are not (2, 4).

Differences in effects may be due to differences in analysis in addition to differences in sample construction. To account for this, we perform the same basic analysis using the data sets from all seven replications, simply comparing the proportion of people who are self-employed above 65 vs. below 65 with no controls or sample weights. However, this consistent comparison may actually exaggerate differences, as the studies with large age ranges generally adjust for them with age controls. Accordingly, estimates still vary widely

[7] We calculated marginal effects ourselves in cases where authors reported logit or probit coefficients.

after making the model consistent. So we also perform the shared analysis while narrowing the age ranges to match the instructions. After doing so, while there is still some variation in the effect, results are very similar. This suggests that the differences in results for this replication study are largely due to differences in modeling, and the decision of how wide of an age range is included in the sample.

Table 3: Sample Creation and Shared-Variable Definition Decisions in Health Insurance Study

Decisions Affecting Sample Size: (* indicates that this is specified in instructions)							
Study	1	2	3	4	5	6	7
NBER CPS Monthly Files May 2004-December 2006*	X	X	X	X	X	X	X
Men only*	X	X	X	X	X	X	X
Observed both before and after turning 65*				X	X		
... Observed between aged 64 to 65						X	
... Observed between ages 63 to 66							X
... Observed between ages 60 to 70		X	X				
... Observed between ages 54 to 76	X						
Must be employed*	X	X	X[a]		X	X	X
Must be in labor force				X			
Drop observations with panel-inconsistent data					X		X
Drop those not observed four times					X		
Drop those for whom May 2004 is month-in-sample 4 or 8, or for whom Dec 2006 is MIS 1 or 5	X						
Drop missing income	X					X	
Decisions Affecting Shared Variable Definitions:							
Treatment compares age 65+ to 64-*		X	X	X	X	X	X
Treatment compares age 65 to other ages	X						
Self-employment given by first worker-class variable	X			X	X	X	X
Self-employment given by both worker-class variables		X	X				
Sample size used in estimation	156,533	90,035	85,400	2,493	1,628	12,288	2,016
Point estimate	0.0232	0.03	0.0338	0.0360	0.0450	0.0501	0.1906
Point estimate under simple shared model	0.1267	0.0751	0.0788	0.0089	0.0522	0.0474	0.0121
Point estimate under simple shared model and shared age range	0.0007	0.0063	0.0074	0.0089	0.0522	0.0091	0.0121

[a]The initial code turned in for this replication had dropped the line in which the sample was limited to the employed, and this X was added after the replicator read this results section. However, because the provided result already used data that included that line, this adjustment did not change anything else.

III.ii.iii. Health Insurance Study Analysis

The health insurance study analysis, by design, looks at people just above and just below an age cutoff, which lends itself to a regression discontinuity design (RDD), albeit one with very little variation in the running variable. Some replicators explicitly used regression discontinuity, while others compared the raw average above and below the cutoff, in effect a regression discontinuity with a zero-order polynomial.

Analysis decisions were more heterogeneous for the health insurance study than for compulsory schooling. Table 4 shows the decisions that replicators made.

Four replicators explicitly used regression discontinuity, but each was different, fitting a linear (2 and 4), quadratic (7), or cubic (3) RDD. The other three used a binary treatment indicator (zero-order polynomial RDD). The exception is replication 1, which as mentioned in the previous section compared 65 to others rather than above/below.

There was also variation in the use of nonlinear models. The dependent variable, self-employment, is binary. Most replicators used linear probability models, but 1 and 3 used probit, while 6 used logit. 2 and 4 used Stata's rdrobust function to run regression discontinuity, and the rest used Stata's regress function to either perform RDD or use a binary treatment indicator.

There are also many differences between analyses in the controls used, in a way that does not fit well into Table 4, because constructs like race and education are controlled for in different ways in different replications. Controls include:

- Replication 1: Family income (midpoint of bins, treated linearly), education (all included levels), race (white/black/other), marital status, citizenship status, presence of own-children under 18 in household, industry indicators for agriculture, financial services, real estate, health services
- Replication 2: Education (8th grade / college degree / postgraduate degree / less-than-8th-grade)

- Replication 3: Race (all included levels), education (all included levels), marital status, metropolitan area/non, state
- Replication 4: None
- Replication 5: Month of interview, year of interview, race (white/nonwhite), education (below/HS/above), marital status, metropolitan area/non
- Replication 6: Family income (16 bins), number of people in the household, marital status (all include levels), education (all included levels), citizenship status, census region
- Replication 7: Month in sample, race (all included levels), education (below/HS/College grad), date.

The list of differences between analyses is long, and are compounded by the fact that some analyses, like cubic RDD, rely on the wide age ranges discussed in the previous section. To evaluate the impact of analysis differences on point estimates, I estimate each model using the data from Replication 5. This process does require changing some of the analyses: specifically, dropping polynomial terms for age in RDD and other contexts. The two replications using rdrobust act strangely in this case, and either cannot run (replication 2), or produce a surprising result that may be due to the command being applied to this particular data (4). The non-rdrobust models, however, produce results with a similar spread to the original estimates, with the exception of Replication 1, which drops much closer to 0 (see the final row of Table 4). This is consistent with the previous section in suggesting that much of the variation is due to differences in age ranges combined with analytical choices.

Table 4: Analysis Decisions in Health Insurance Study

Study	6	7	4	3	2	5	1
Decisions Affecting Analysis:							
Explicit Regression Discontinuity...							
... Linear		X		X			
... Quadratic							X
... Cubic			X				
Above/Below Binary					X	X	
Linear Probability Model		X		X	X		X
Probit/Logit	X		X			X	
Heteroskedasicity-Robust Standard Errors							X
Clustered SEs (Individual)					X		
Clustered SEs (State)		X	X				
Stata regress					X		X
Stata probit	X		X				
Stata rdrobust		X		X			
R glm(link = 'logit')						X	
Person Sample Weights			X				X
Point estimate	0.0232	0.0300	0.0338	0.0360	0.0450	0.0501	0.1906
Point estimate under data from Rep. 5 (and RDD terms restricted to linear)	0.0488	NA	0.0353	-0.0203	0.0210	0.0307	0.0121

IV. Conclusion

Given the same data and research question of interest, we find considerable variation across researchers in the way that they clean and prepare data and design their analysis. In one of the two studies we examine, this led to considerable variation in results.

In both studies, the estimated sampling variation within studies was small relative to variation between studies. In the compulsory education study, the average reported standard error was 25.1% as large as the standard deviation of reported effects across studies. This figure was 32.5% in the health insurance study. In both cases standard errors omit a major source of variation in estimates.

It is not surprising that different researchers would carry out an analysis in different ways. Replicators were asked after completing their replication about their reasoning for the analytic and data cleaning choices that were not covered by the instructions and differed among replicators. The most common reasons included familiarity with a given model, differing intuitive or technical ideas about which control variables are appropriate or whether linear probability models are appropriate, and differing preferences for parsimony.

There is nothing inherently wrong about these choices or reasons, although the fact that researchers do not seem to agree on these issues implies additional sources of uncertainty in estimates. These differences only rise to a real cause for concern when they are about things that either would be unlikely to be reported in the resulting study, or would be reported but paid little attention by reviewers and readers. If invisible researcher choices are different and consequential, that means that empirical results in applied microeconomics reflect variation in sample and methods, as expected, but also reflect variation in researcher choice. And while this variation is not the same thing as "p-hacking," as these choices are not necessarily related to an attempt to find a particular result, this does makes it easier for an unscrupulous researcher to attempt many different analyses to get a desired result without detection.

Some of the differences between replicators would be very likely to be reported and given attention, such as the choice of age ranges in the health insurance study or the use of state

linear time trends in the compulsory schooling study. Others, like the large variation in selection of controls, would be likely to be reported, but as discussed by Lenz and Sahn (2017) in the context of the political science literature, these choices are often not justified in economics papers, and they might not receive strict attention from reviewers. The specific construction of those controls such as how coarsely to bin an education control, which also varied, is unlikely to receive notice from a reviewer.

The existence of these analytic differences suggests both that more space be given in economics papers to justifications of decisions like the use of control variables and variable definitions. Another approach is the use of model averaging, in which multiple possible ways of designing the model are averaged together to produce a result (Moral-Benito, 2015).

The biggest issue highlighted by these results is the considerable differences between researchers in the way the data was cleaned and prepared. No two researchers had the same sample size in their analysis. Nearly all of the decisions driving data construction would be likely to be omitted from a paper, or skimmed over by a reader.

It is also not clear that these differences are because replicators were making wrong decisions. So, even in the case where a reader or reviewer goes through the data preparation code for the paper, they might not see any problem. There is not a well-known set of "best practices" for data cleaning and preparation in economics, at least not to the extent that there is with analysis.

Differing approaches to data cleaning are particularly concerning given recent increases in the use of proprietary and otherwise nonstandardized data sources in economics (Currie et al., 2020). If only a handful of researchers have access to a given data set, data cleaning decisions are more difficult to observe and analyze.

While there are inherently always going to be researcher degrees of freedom affecting results, there are several possible approaches that could be taken to address differences at the data preparation stage. The first is simple transparency. American Economic Association (2020) has already taken steps to improve access to code, and it may also be advisable to

make the "Data Appendix" a more standard feature of economics papers, even for studies using standard data sources, and for that appendix to detail all the decisions made in the process of preparing data.

In the case of standard publicly-available data sources, the use of pre-prepared data would improve standardization across researchers. For example, the Current Population Survey Merged Outgoing Rotation Group Files provided by the National Bureau of Economic Research (2020) are popular, and remove any researcher degrees of freedom from the process of merging together CPS files, although of course not from any point later in the data preparation process. Standard pre-prepared data for other common data sources may help make results more consistent.

Variation in results driven by researcher degrees of freedom in either analysis or data preparation naturally comes from hundreds of decisions that are seemingly innocuous, and would be difficult to evaluate or test. So any fixes can only be partial. It is relieving that, in this paper, one of the two studies had little variation in results despite major differences in researcher choices. But the other study did have major differences. Strengthening the validity of results in applied microeconomics could come through additional transparency in all stages of analysis, a stronger sense of "best practices" in all stages of analysis, standardized data sources, and the use of tools like model averaging. These approaches may allow us to account for researcher degrees of freedom in our understanding of results.

V. References

References

American Economic Association. 2020. Data and Code Availability Policy.

Bernal, Raquel, and Michael P. Keane. 2011. Child Care Choices and Children's Cognitive Achievement: The Case of Single Mothers. Journal of Labor Economics 29(3):459–512.

Berry, James, Lucas C. Coffman, Douglas Hanley, Rania Gihleb, and Alistair J. Wilson. 2017. Assessing the Rate of Replication in Economics. American Economic Review 107(5):27–31.

Black, Sandra E., Paul J. Devereux, and Kjell G. Salvanes. 2008. Staying in the Classroom and Out of the Maternity Ward? The Effect of Compulsory Schooling Laws on Teenage Births. The Economic Journal 118(530):1025–1054.

Botvinik-Nezer, Rotem, Felix Holzmeister, Colin F. Camerer, Anna Dreber, Juergen Huber, Magnus Johannesson, Michael Kirchler, and 190 more. 2019. Variability in the Analysis of a Single Neuroimaging Dataset by Many Teams. Tech. Rep. 8641719, biorXiv.

Camerer, Colin F., Anna Dreber, Eskil Forsell, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, Johan Almenberg, Adam Altmejd, Taizan Chan, Emma Heikensten, Felix Holzmeister, Taisuke Imai, Siri Isaksson, Gideon Nave, Thomas Pfeiffer, Michael Razen, and Hang Wu. 2016. Evaluating Replicability of Laboratory Experiments in Economics. Science 351(6280):1433–1436.

Chang, Andrew C., and Phillip Li. 2017. A Preanalysis Plan to Replicate Sixty Economics Research Papers That Worked Half of the Time. American Economic Review 107(5): 60–64.

Christensen, Garret, and Edward Miguel. 2018. Transparency, Reproducibility, and the Credibility of Economics Research. Journal of Economic Literature 56(3):920–980.

Clemens, Michael A. 2017. The Meaning of Failed Replications: A Review and Proposal. Journal of Economic Surveys 31(1):326–342.

Cohn, Nate. 2016. We Gave Four Good Pollsters the Same Raw Data. They Had Four Different Results. The New York Times.

Currie, Janet, Henrik Kleven, and Esmée Zweirs. 2020. Technology and Big Data Are

Changing Economics: Mining Text to Track Methods. Working Paper 26715, National Bureau of Economic Research.

Dewald, William G., Jerry G. Thursby, and Richard G. Anderson. 1986. Replication in Empirical Economics: The Journal of Money, Credit and Banking Project. The American Economic Review 76(4):587–603.

Duvendack, Maren, Richard Palmer-Jones, and W. Robert Reed. 2017. What Is Meant by "Replication" and Why Does It Encounter Resistance in Economics? American Economic Review 107(5):46–51.

Fairlie, Robert W., Kanika Kapur, and Susan Gates. 2011. Is Employer-based Health Insurance a Barrier to Entrepreneurship? Journal of Health Economics 30(1):146–162.

Gelman, Andrew, and Eric Loken. 2014. The Statistical Crisis in Science: Data-dependent Analysis– "A Garden of Forking Paths"–Explains Why Many Statistically Significant Comparisons Don't Hold Up. American Scientist 102(6):460–466.

Gertler, Paul, Sebastian Galiani, and Mauricio Romero. 2018. How to Make Replication the Norm. Nature 554(7693):417–419.

Goodman-Bacon, Andrew. 2018. Difference-in-Differences with Variation in Treatment Timing. Working Paper 25018, National Bureau of Economic Research.

Hamermesh, Daniel S. 2007. Viewpoint: Replication in Economics. Canadian Journal of Economics/Revue canadienne d'économique 40(3):715–733.

———. 2017. Replication in Labor Economics: Evidence from Data, and What It Suggests. American Economic Review 107(5):37–40.

Herndon, Thomas, Michael Ash, and Robert Pollin. 2014. Does High Public Debt Consistently Stifle Economic Growth? A Critique of Reinhart and Rogoff. Cambridge Journal of Economics 38(2):257–279.

Ioannidis, John P. A., T. D. Stanley, and Hristos Doucouliagos. 2017. The Power of Bias in Economics Research. The Economic Journal 127(605):F236–F265.

Lenz, Gabriel S., and Alexander Sahn. 2017. Achieving Statistical Significance with Covariates and without Transparency. Tech. Rep., University of California Berkeley.

Lochner, Lance, and Enrico Moretti. 2004. The Effect of Education on Crime: Evidence from Prison Inmates, Arrests, and Self-Reports. American Economic Review 94(1):155–189.

Moral-Benito, Enrique. 2015. Model Averaging in Economics: An Overview. Journal of Economic Surveys 29(1):46–75.

Moretti, Enrico. 2000. Do Wages Compensate for Risk of Unemployment? Parametric and Semiparametric Evidence from Seasonal Jobs. Journal of Risk and Uncertainty 20(1): 45–66.

Mueller-Langer, Frank, Benedikt Fecher, Dietmar Harhoff, and Gert G. Wagner. 2019. Replication studies in economics—How many and which papers are chosen for replication, and why? Research Policy 48(1):62–83.

National Bureau of Economic Research. 2020. CPS Merged Outgoing Rotation Groups.

Reinhart, Carmen M., and Kenneth S. Rogoff. 2010. Growth in a Time of Debt. American Economic Review 100(2):573–578.

Silberzahn, R., E. L. Uhlmann, D. P. Martin, P. Anselmi, F. Aust, E. Awtrey, Š. Bahník, and 59 more. 2018. Many Analysts, One Data Set: Making Transparent How Variations in Analytic Choices Affect Results. Advances in Methods and Practices in Psychological Science 1(3):337–356.

Simmons, Joseph P., Leif D. Nelson, and Uri Simonsohn. 2011. False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. Psychological Science 22(11):1359–1366.

Young, Alwyn. 2018. Consistency without Inference: Instrumental Variables in Practical Application. Unpublished.

Appendix A. Full Replicator Instructions

Appendix A.i. Compulsory Schooling Study Instructions

Data:

U.S. Census public sample files for 1940-1980, limited to female subjects age 20-30. Please download 1940 1% sample, 1950 1% sample, 1960 1% sample, both 1970 1% state samples, and 1980 5% sample from https://usa.ipums.org/usa/index.shtml rather than from any other source. IPUMS extracts come in .gz compressed format. See this guide if you are unable to decompress the extract for use: https://fileinfo.com/extension/gz. Additionally, use the table provided on page 3-4 of compulsory education policy by state for every decade from 1924 to 1974. This table lists the age at which students are allowed to drop out of school; a higher number is a more restrictive policy and requires that students stay in school longer. Limit the sample to those born in one of the states present in the provided table.

General Question:

There are several reasons to believe that education may reduce the incidence of teenage pregnancy. Education raises the level of human capital, which increases the opportunity cost of young pregnancy. Additionally, education takes a lot of time, which may reduce the time available for risky behavior. What effect does education have on teenage pregnancy?

Specifically:

Compulsory schooling laws are likely to increase the amount of education attained, especially for marginal students who are more likely to become pregnant. Does the compulsory education age in a state affect the proportion of women in that state who have a child by age 18?

Estimate and Identification:

Calculate the age at first birth as the difference between current age and the age of the eldest own-child in household. Drop anyone for whom age-at-first-birth is 14 or below (including the negative values). Assume that women are exposed to the most recently listed

policy in their state of birth as of the year they turn 14. For example, a woman who was born in Arkansas and turned 14 in 1938 is subject to the compulsory schooling law listed for Arkansas in 1934 in the policy table.

Estimate the effect of compulsory schooling policies with a required age of 16 or higher, as opposed to those with an age of 15 or lower, on the probability that a woman will have a child by age 18. Make the identifying assumption that trends in teen pregnancy are unrelated to the decision to change compulsory schooling policy.

Reminders:

- You may use any statistics package you like.
- If you would typically use graduate students for a given task (data cleaning, coding, etc.) we encourage you to use them for that task in this project as well.
- Unless necessary, we ask that you try not to ask us for clarification on how we would like the analysis done, as we want the analysis to be independent. Similarly, do not try to guess how other researchers will approach this task in order to match (or avoid matching) their approach, or try to seek out the original study to see how it was done. The idea is that we want to see how you would estimate this effect, if you'd had this question, this idea for identification, and had chosen this particular sample.

Turn in when done:

- All code used, including (1) any IPUMS-provided code that imports the data into your statistics package of choice, (2) any code that cleans or merges the data, and (3) code that runs the estimation. Please include instructions as to the order code should be run if there are multiple code files. It should be possible to recreate your results from the raw IPUMS files by running the provided code in the determined order. Please comment your code if possible. At minimum, add comments reading "ESTIMATE BEGINS HERE" and "ESTIMATE ENDS HERE" so that the exact causal analysis run can be easily located.

- A table of summary statistics of relevant variables.
- One estimate of the causal effect of interest, as you would present it if you were submitting this paper for publication. Please select a single "preferred estimate" rather than several estimates produced under differing assumptions (robustness tests). What's the estimate you'd mention in the abstract or intro of this paper if you were publishing it? Use that one. Your preferred estimate may be a point estimate with standard errors, a confidence interval, an estimated model, or a mix of these, as appropriate.
- The IPUMS extract used. Please hold on to the .gz file provided by IPUMS to reduce file size.
- Please email completed project materials to nhuntington-klein@fullerton.edu. We recommend hosting data extracts on a cloud service like Dropbox and including a link to the file in your email, rather than trying to include the data extract as an attachment. Let us know if you have difficulty with this.
- We are aiming for a completion date of January 2019. Please let us know if this is an issue.

State	Compulsory schooling age in...					
(US Postal Code)	1924	1934	1944	1954	1964	1974
AL	16	16	16	16	16	16
AZ	16	16	16	16	16	16
AR	15	16	16	16	16	16
CA	16	16	16	16	16	16
CO	16	16	16	16	16	16
CT	16	16	16	16	16	16
DE	16	16	16	16	16	16
DC	14	16	16	16	16	16

FL	16	16	16	16	16	16
GA	14	14	14	16	16	16
ID	18	18	16	16	16	16
IL	16	16	16	16	16	16
IN	16	16	16	16	16	16
IA	16	16	16	16	16	16
KS	16	16	16	16	16	16
KY	16	16	16	16	16	16
LA	14	14	14	16	16	16
ME	17	17	14	16	16	16
MD	16	16	16	16	16	16
MA	16	16	16	16	16	16
MI	16	16	16	16	16	16
MN	16	16	16	16	16	16
MS	14	17	16	16	0	0
MO	16	16	14	16	16	16
MT	16	16	16	16	16	16
NE	16	16	16	16	16	16
NV	18	18	18	18	17	17
NH	16	16	16	16	16	16
NJ	16	16	16	16	16	16
NM	16	16	16	17	17	17
NY	16	16	16	16	16	16
NC	14	14	14	16	16	16
ND	17	17	17	17	16	16
OH	18	18	8	18	18	18
OK	18	18	18	18	18	18

OR	16	18	16	18	18	18
PA	16	16	18	17	17	17
RI	16	16	16	16	16	16
SC	14	14	16	16	0	16
SD	17	17	17	17	16	16
TN	16	16	16	16	16	17
TX	14	14	16	16	16	17
UT	18	18	18	18	18	18
VT	16	16	16	16	16	16
VA	14	15	15	16	16	17
WA	16	16	16	16	16	16
WV	16	16	16	16	16	16
WI	16	16	16	16	16	18
WY	16	17	17	16	17	17

Appendix A.ii. Employer-Based Health Insurance Study Instructions

Data:

Current Population Survey monthly files for May 2004-December 2006, limited to male subjects who can be observed in the exact month that they turn age 65. Please download data from http://www.nber.org/data/cps_basic.html rather than from any other source.

General question:

Many people in the United States rely on their employers for health insurance. This may act as a deterrent to starting a business, since entrepreneurs must find their own health insurance. How does the availability of non-employer based health insurance affect business creation?

Specifically:

Americans who are age 65 (or above) are eligible for health insurance through Medicare. Americans who are age 64 and 11 months (or below) are generally ineligible for Medicare. Does Medicare increase the incidence of business ownership among men in the United States?

Estimate and identification:

Estimate the effect of eligibility for Medicare on the probability of being self-employed, conditional on being employed at all. Use the fact that people become eligible for Medicaid in the month that they turn 65. Make the identifying assumption that nothing else of importance changes between age 64 and 11 months and age 65.

Reminders:

- You may use any statistics package you like.
- If you would typically use assistants for a given task (data cleaning, coding, etc.) we encourage you to use them for that task in this project as well.
- Unless necessary, we ask that you try not to ask us for clarification on how we would like the analysis done, as we want the analysis to be independent. Similarly, do not try to guess how other researchers will approach this task in order to match (or avoid matching) their approach, or try to seek out the original study to see how it was done. The idea is that we want to see how you would estimate this effect, if you'd had this question, this idea for identification, and had chosen this particular sample.

Turn in when done:

- All code used, including (1) any NBER-provided code that imports the data into your statistics package of choice, (2) any code that cleans or merges the data, and (3) code that runs the estimation. Please include instructions as to the order code should be run if there are multiple code files. It should be possible to recreate your results from the raw CPS .DAT files by running the provided code in the determined order. Please

comment your code if possible. At minimum, add comments reading "ESTIMATE BEGINS HERE" and "ESTIMATE ENDS HERE" so that the exact causal analysis run can be easily located.

- A table of summary statistics of relevant variables.
- One estimate of the causal effect of interest, as you would present it if you were submitting this paper for publication. Please select a single "preferred estimate" rather than several estimates produced under differing assumptions (robustness tests). What's the estimate you'd mention in the abstract or intro of this paper if you were publishing it? Use that one. Your preferred estimate may be a point estimate with standard errors, a confidence interval, an estimated model, or a mix of these, as appropriate.
- Please email completed project materials to nhuntington-klein@fullerton.edu. Let us know if you have difficulty with this.
- We are aiming for a completion date of January 2019. Please let us know if this is an issue.