I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 17744

The Sources of Researcher Variation in Economics

Nick Huntington-Klein
Claus C. Portner
et al.

FEBRUARY 2025

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 17744

The Sources of Researcher Variation in Economics

Nick Huntington-Klein
Seattle University

Claus C. Portner
Seattle University

et al.

For a complete list of authors please check pages 1 to 3.

FEBRUARY 2025

ISSN: 2365-9793

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

The Sources of Researcher Variation in Economics

We use a rigorous three-stage many-analysts design to assess how different researcher decisions—specifically data cleaning, research design, and the interpretation of a policy question—affect the variation in estimated treatment effects. A total of 146 research teams each completed the same causal inference task three times each: first with few constraints, then using a shared research design, and finally with pre-cleaned data in addition to a specified design. We find that even when analyzing the same data, teams reach different conclusions. In the first stage, the interquartile range (IQR) of the reported policy effect was 3.1 percentage points, with substantial outliers. Surprisingly, the second stage, which restricted research design choices, exhibited slightly higher IQR (4.0 percentage points), largely attributable to imperfect adherence to the prescribed protocol. By contrast, the final stage, featuring standardized data cleaning, narrowed variation in estimated effects, achieving an IQR of 2.4 percentage points. Reported sample sizes also displayed significant convergence under more restrictive conditions, with the IQR dropping from 295,187 in the first stage to 29,144 in the second, and effectively zero by the third. Our findings underscore the critical importance of data cleaning in shaping applied microeconomic results and highlight avenues for future replication efforts.

Keywords: metascience, applied econometrics, causal inference, research methods

Corresponding author:
Nick Huntington-Klein
Department of Economics
Seattle University
901 12th Ave.
Seattle, WA, 98122
USA
E-mail: nhuntington-klein@seattleu.edu

The Sources of Researcher Variation in Economics*

Nick Huntington-Klein Claus C. Pörtner Yubraj Acharya
Matus Adamkovic Joop Adema Lameck Ondieki Agasa
Imtiaz Ahmad Mevlude Akbulut-Yuksel
Martin Eckhoff Andresen David Angenendt
José-Ignacio Antón Andreu Arenas Erkmen Giray Aslim
Stanislav Avdeev Andrew Bacher-Hicks Bradley J. Baker
Imesh Nuwan Bandara Avijit Bansal David Bartram
Katarzyna Bech-Wysocka Christopher Troy Bennett
Andu N. Berha Inés Berniell Moiz Bhai
Shreya Bhattacharya Markus Bjoerkheim Jeffrey R. Bloem
Margaret E Brehm Martín Brun Florent Buisson
Pralhad Burli Andrew M. Camp Nicola Cerutti
Weiwei Chen Jeffrey Clement Matthew Collins
Lee Crawfurd John Cullinan Lachlan Deer
Reid Dorsey-Palmateer Nicolas J. Duquette

Diego Marino Fages Grace Falken Christine Farquharson

Jan Feld Yevgeniy Feyman Nathan Fiala Anne Fitzpatrick

Andrey Fradkin Evaewero French Wei Fu Luca Fumarco

Sebastian Gallegos Julio Galárraga Aaron M. Gamino

Romain Gauriot Victor Gay Savas Gayaker Jules Gazeaud

Alexandra de Gendre Gregory Gilpin Daniele Girardi

Dan Goldhaber Mark N. Harris Blake H. Heller

Daniel J. Henderson Arne Henningsen Junita Henry

Clément Herman Øystein Hernæs Andrew J. Hill

Felix Holzmeister Martijn Huysmans M. Saad Imtiaz

Anil K. Jain Niklas Jakobsson José Kaire

Kalyan Kumar Kameshwara Daniel H Karney Sie Won Kim

Valentin Klotzbücher Christoph Kronenberg Daniel LaFave

David Lang Ryan Lee Maxime Liégey Dede Long

Jan Marcus Gabriele Mari Ian McCarthy

Laura Meinzen-Dick Erik Merkus Klaus M. Miller

Lukas Mogge S. M. Woahid Murad Rafiuddin Najam

Elias Naumann Job Nda Nmadu Gorkem Turgut Ozer

Jayash Paudel Filippos Petroulakis Christian Peukert

Visa Pitkänen Simon Porcher Manab Prakash

Andrew Adrian Yu Pua Todd Pugatch Daniel S. Putman
Veeshan Rayamajhee Obeid Ur Rehman Maira Emy Reimão
Anna Reuter Michael David Ricks Fernando Rios-Avila
Abel Rodriguez Julian Roeckert Ivan Ropovik Jayjit Roy
Nicolas Salamanca Margaret Samahita Aparna Samudra
Vassiki Sanogo Orkhan Sariyev Henning Schaak
Joel E. Segel Hans Henrik Sievertsen Mike Smet
Brock Smith Lucy C. Sorensen Lisa Spantig
Krzysztof Szczygielski Anirudh Tagat Huseyin Tastan
Martin Trombetta Madhavi Venkatesan Antoine Vernet
Eden Volkov Gary A. Wagner Yue Wang Zachary Ward
Tom Waters Ellerie Weber Stephen E Weinberg
Kristina S. Weißmüller Christian Westheide
Kevin M. Williams Xiaoyang Ye Jisang Yu
Muhammad Umer Zahid Raffaele Zanoli

*Corresponding author Nick Huntington-Klein, nhuntington-klein@seattleu.edu, +1 (206) 296-5815. Department of Economics, Seattle University, 901 12th Ave., Seattle, WA, 98122. Huntington-Klein and Pörtner are the project organizers. This project was supported by the Alfred P. Sloan foundation grant G-2022-19377. The Seattle University IRB determined this study to be exempt from IRB review in accordance with federal regulation criteria. We would like to thank Kian Farzaneh, Amrapali Samanta, and Erica Long for research assistance and seminar participants at the Center for Education data and Research (CEDR)/Center for Analysis of Longitudinal Data in Education Research (CALDER) at the University of Washington, Institut national de recherche en sciences et technologies du numérique (INRIA), La Universidad de las Americas,

Abstract: We use a rigorous three-stage many-analysts design to assess how different researcher decisions—specifically data cleaning, research design, and the interpretation of a policy question—affect the variation in estimated treatment effects. A total of 146 research teams each completed the same causal inference task three times each: first with few constraints, then using a shared research design, and finally with pre-cleaned data in addition to a specified design. We find that even when analyzing the same data, teams reach different conclusions. In the first stage, the interquartile range (IQR) of the reported policy effect was 3.1 percentage points, with substantial outliers. Surprisingly, the second stage, which restricted research design choices, exhibited slightly higher IQR (4.0 percentage points), largely attributable to imperfect adherence to the prescribed protocol. By contrast, the final stage, featuring standardized data cleaning, narrowed variation in estimated effects, achieving an IQR of 2.4 percentage points. Reported sample sizes also displayed significant convergence under more restrictive conditions, with the IQR dropping from 295,187 in the first stage to 29,144 in the second, and effectively zero by the third. Our findings underscore the critical importance of data cleaning in shaping applied microeconomic results and highlight avenues for future replication efforts.

Ludwig-Maximilians-Universität München (LMU Munich), Western Washington University, and the 2024 Annual Meeting of WEAI for helpful comments and suggestions. We would also like to thank the researchers Mira Chaskes, Jennifer A. Heissel, Elaine L. Hill, Rajius Idzalika, Joshua D. Merfeld, and Ethan Sawyer, who contributed but did not want an authorship slot, the researchers who wished to remain anonymous, and the researchers who enlisted in the study but were ineligible or unable to complete all three rounds of the project. Data and code for this project are available at https://github.com/many-economists/analysis. Preregistration for the project is available at OSF: https://doi.org/10.17605/OSF.IO/CJ9YX. Suggested citation: Huntington-Klein, Pörtner, et al. (2025), "The Sources of Researcher Variation in Economics."

1 Introduction

Skepticism about empirical results in economics is not new, but has received increasing attention over the last decade with concerns about replicability, publication bias, power, and p-hacking (Leamer 1983; Brodeur, Cook, and Heyes 2020; Lang Forthcoming). Even in journals with data availability policies, code and data are, more often than not, either not available or do not reproduce the published results, and "policing replications" that test sensitivity of published results are rare (Herbert et al. 2021; Ankel-Peters, Fiala, and Neubauer 2023). Even experimental economics results are not immune; a high percentage of studies cannot be replicated when tested using new data (Camerer et al. 2016).[1]

A broader issue is that researchers face myriad choices regarding data collection, data cleaning, variable selection, and estimation methods, each of which can substantially affect published results. For example, researchers are more likely to present marginally significant results over marginally insignificant ones (Brodeur, Lé, et al. 2016; Brodeur, Cook, and Heyes 2020). Even without conscious manipulation, these numerous "researcher degrees of freedom" can lead equally competent researchers to substantially different conclusions (Simmons, Nelson, and Simonsohn 2011). In psychology, the variation introduced by researcher choices might outweigh the population variation typically considered when estimating standard errors (Holzmeister et al. 2023). Similarly, in finance, these methodological choices—many of which remain unreported—explain substantial variation in estimated effects across 80 different studies of the same policy change (Black et al. 2024).

Our goal is to understand the relative importance of the various researcher degrees of freedom in explaining estimate variation.[2] We use a "many-analysts" design, where researchers inde-

[1] Experiments in social psychology, and psychology more broadly, perform even worse, leading to discussions about an ongoing "replication crisis" (Open Science Collaboration 2015).

[2] Traditional replication work asks whether a study's results are robust to re-evaluation, whereas researcher degrees of freedom focuses on whether different researchers would perform the same study differently. These fields intersect when replication failures arise because both the original and replication analyses made rea-

pendently perform the same research task. We additionally have those researchers perform the task multiple times, under progressively stricter restrictions on their choices. Our chosen task, common in applied econometrics, is to estimate the causal effect of a policy implemented at a specific time and affecting only some individuals. We isolate researcher degrees of freedom at each stage of the research process to examine where researcher choices vary most and where they most strongly impact results. We also examine whether differences in researchers' characteristics and their analytic and data cleaning choices can explain the variation in results.

The three main contributions of this paper are: first, we introduce multiple iterations of the research task, second, the initial stage provides researchers with more freedom in relation to data processing than is common in many-analyst studies, and, finally, we have a substantially larger number of researchers who complete the project than most prior many-analysts efforts. By introducing multiple iterations of the task, each time restricting the amount of choice that researchers can make and so reducing researcher degrees of freedom, we can both observe the overall amount of variation in estimates between researchers, as is common in many-analysts designs, and separately evaluate the influence of choice in research design and in data cleaning.

In a "many-analysts" design, organizers provide multiple teams of researchers with the same data and have them independently try to answer the same research question (Silberzahn et al. 2018). Many-analysts studies have been conducted in microeconomics (Huntington-Klein et al. 2021; Borjas and Breznau 2024), finance (Menkveld et al. 2024), religion (Hoogeveen et al. 2023), neuroimaging (Botvinik-Nezer et al. 2020), political science (Breznau et al. 2021), machine learning (W. Chen and Cummings 2024), ecology and evolutionary biology (Gould et al. 2023), psychology (Boehm et al. 2018; Bastiaansen et al. 2020; Schweinsberg et al. 2021), and medical informatics (Ostropolets et al. 2023), among others.[3]

sonable but divergent choices (Bryan, Yeager, and O'Brien 2019).

[3]See also Magnus and Morgan (1997) for an early example in the same vein in applied econometrics.

Most many-analysts studies find meaningful variation in both methods and conclusions across researchers. However, participating in such studies requires considerable time and effort, which has limited the size of most prior studies and prevented them from moving beyond demonstrating the existence of variation to exploring its causes and potential remedies.[4] To achieve sufficient statistical power to both establish the presence of variation *and* examine its sources, our goal was for at least 90 researchers to complete all steps of the project. In total, 146 research teams successfully completed all tasks, exceeding this requirement

Three common explanations for researcher variation are task difficulty, researcher experience or characteristics, and peer review. The more complex or difficult-to-analyze scenarios are, the less researcher agreement (Menkveld et al. 2024; Ortloff et al. 2023). Higher-quality or more experienced teams tend to agree more and draw more abstract codebooks and conclusions, and replicators with more coding skill find more errors (Menkveld et al. 2024; Ortloff et al. 2023; Broderick, Giordano, and Meager 2020). Researcher political orientation and personality also affect findings, both in many-analysts work and outside (Borjas and Breznau 2024; Jelveh, Kogut, and Naidu 2024; Sulik et al. 2023). However, some many-analysts studies show that researcher characteristics explain only a small share of the variation (Breznau et al. 2021). Finally, peer review may increase agreement if there is an option to revise, although if instead outside evaluation is used as a measure of researcher quality, peer review scores do not necessarily predict outlier results (Menkveld et al. 2024; Gould et al. 2023).

Other work uses simulation to explore numerous analytical or data-cleaning combinations and measure resulting estimate variation. Like many-analysts studies, the aim of these simulations is to identify how different choices influence results, but they are necessarily limited to decisions identified by the organizers, and treat all combinations equally. One particularly relevant example examines the sensitivity of results in an observational psychological data set to various

[4] A notable exception is Menkveld et al. (2024), which had 164 teams test the same hypotheses on the same data.

preprocessing and modeling choices and finds significant variation (Klau et al. 2023). A similar attempt to separate researcher variation into modeling and preprocessing components is also done in a many-analysts design in Huntington-Klein et al. (2021), although in a limited way.

We employ a three-staged design to evaluate multiple sources of variation: data preparation, research design, and the interpretation of the research question. Each stage allows a narrowing degree of researcher choice, with randomized peer reviews in between stages. The first stage allows researchers substantial freedom in answering the research question, while the second stage specified the research design more precisely, and the third round in addition provided a pre-cleaned data set. The goal is to incorporate the mechanisms proposed by the literature and to respond to the critique of prior studies (Auspurg and Brüderl 2021). We also collect researcher characteristics to explore their role in estimate variation, although not in a controlled way. We do not address the difficulty of the research task as a potential source of researcher variation.

Our results show that while researchers varied considerably in their data preparation and modeling choices, the reported policy effects were relatively similar to each other, at least in the center of the distribution. The IQR of policy impacts in the first stage, where researchers had full freedom, was only 3.1 percentage points, although there were substantial outlier estimates. The second stage showed *less* agreement than the first, with an IQR of 4.0 percentage points, with the reduction in agreement driven by some researchers not fully adhering to the specified research design. In the final stage, where data was pre-cleaned to eliminate errors in data preparation, the IQR fell to its lowest level at 2.4 percentage points. We considered this a meaningful improvement in agreement, although the reduction in the variance of estimated effects was not statistically significant. Specifying a research design considerably improved agreement in reported sample sizes, with the IQR of sample sizes falling from 295,187 originally to 29,144 in the second stage to effectively 0 for the final stage. In contrast to these changes, we

found no impact of peer review or researcher background or experience on reported effects.

Some of the observed differences stem from decisions that are typically scrutinized, such as research design and control variables. Other arise in less examined areas, like functional form, data cleaning, and sample limitation decisions. When researchers are required to use the same design, their results became more similar, especially when that shared design is adhered to. Agreement rose sharply when data was pre-cleaned, suggesting that data cleaning decisions are a major source of variation. More standardized data-cleaning procedure and greater transparency in cleaning code could substantially improve consistency and credibility in applied microeconomics.

The rest of the paper is organized as follows. We first present the research design in Section 2. This is followed by a description of the collected data and characteristics of the participating research teams. Section 4 presents the results. Finally, we discuss the implications of our results and suggest areas of future research.

2 Design

We have the same set of researchers complete the same research task at least three times to isolate the influence of different sources of researcher variation. The research task is to estimate the effect of the Deferred Action for Childhood Arrivals (DACA) program on the probability that those affected by the program work full-time. The details and restrictions differ between the three main rounds, which we will refer to as Task 1, Task 2, and Task 3. The intuition behind this design is that if the removal of a specific kind of researcher freedom meaningfully reduces the variation in results between researchers, then that degree of freedom is a meaningful contributor to researcher variation. Following each task, a subset of researchers are randomized into peer review pairs, and given the opportunity to revise their work.

The following goals and instructions are shared across all tasks:

- Estimate the causal effect of the DACA policy on the probability of working full-time, among the group affected by that policy (see Appendix Section A below for more details).
- Use American Community Survey (ACS) data to estimate the effect, using data no older than 2006 and no newer than 2016.
- Procure ACS data from IPUMS (Ruggles et al. 2024), selecting only one-year files and using harmonized variables.
- Optionally, combine the ACS data with a data set on the presence or absence of other relevant policies by state and year, provided by the organizers.
- Use a statistics package or language that allows results to be immediately replicated.

Researchers were also given background information on DACA and its eligibility criteria, guidance on how to use the IPUMS website, instructed to use assistants for any work they would normally use assistants for, and to complete their analysis as though it had been their own idea, rather than attempting to match or not-match other researchers, or asking the project organizers how they would like the analysis to be performed.

Task 1 gives researchers a large amount of freedom in how they complete the research task, with the instructions above comprising the entirety of the limitations on researchers in Task 1. Each successive task removes a degree of freedom from the researcher and further specifies how the analysis is to be performed.

Task 2 specified the research design more precisely, with the goal of examining whether researcher variation arises from an imprecise statement of the research question, as in Auspurg and Brüderl (2021), or is due to differences in research design choices. Instead of allowing

any research design to identify the causal effect of interest, Task 2 gave specific definitions for which individuals comprised a "treated" group and which comprised an "untreated" comparison group.[5] Researchers are then instructed to estimate the effect by comparing how outcomes for the "treated" group changed from before DACA was implemented to afterwards against how outcome for the "untreated" group changed. This can be thought of as a difference-in-differences style design, although the phrase "difference-in-differences" was not used in the instructions.

To show the researcher variation introduced by decisions made in the data cleaning and variable definition process, Task 3 provides a pre-cleaned data set, prepared by the organizers, while maintaining the same research design limitations as in Task 2. In principle, a researcher following the Task 2 instructions should arrive at the same sample size, number of treated individuals, and number of untreated individuals as in Task 3, as well as the same definition for the outcome variable.[6] Hence, differences in the data set and in the results between Task 2 and Task 3 should be a result of differences in the data cleaning and preparation process. The data set offered a pre-prepared treated/untreated-group indicator as specified in Task 2, limited the data set only to the treated and untreated group, prepared and cleaned all variables in the data set that did not already come pre-cleaned, handled missing-data flags, merged in state policy data, and offered standardized simplified recodings of demographic variables. Researchers were instructed to not further clean the data or limit the sample.

Following each of the research tasks, 2/3 of the researchers are randomly assigned to peer review and 1/3 not assigned to peer review. Those in peer review are randomly assigned in

[5]Although eligibility criteria for DACA were explicitly given in Task 1, Task 2 further limits the treated group by narrowing the acceptable age range. The limitation was more impactful for defining the untreated comparison group, though. Many researchers did use a treated/untreated group approach in Task 1 before it was specified in Task 2, but researchers defined the untreated group in highly diverse ways, as will be shown in the Results section.

[6]The Task 2 instructions do leave some leeway for definition of some variables, in particular control variables like education or race, which have a specific recoded version available in Task 3 that is not specified in the Task 2 instructions. However, the definitions of the treated and untreated comparison groups should be the same between Task 2 and Task 3.

pairs. Those pairs were given work performed by the other member of their pair; the other person's response to the research survey (see below) as well as a brief write-up representing their work, usually including a regression table. Each member performed a blind review of the provided work, and provided a written assessment of that work, which was shared with the original researcher. Reviewers were instructed to produce a review "as though (they) were the reviewer of a journal article," and to judge the work as though they were reviewing for a journal where a study of this kind "could be published if the work was of high quality." Following peer review, researchers have an opportunity to revise their work in light of the peer review (or for any other reason). Importantly, revision is not mandatory, nor is satisfying one's peer reviewer, and the majority of researchers choose not to submit revisions.

This form of peer review does not match what is typically done in peer review work for journal publications. In particular, revision is not mandatory, all reviewers have themselves completed a study with the same goal and data and so have extensive background information, and all reviewers are themselves also reviewed by the same person. These features will all affect interpretation of the peer review results. The non-mandatory nature of the revision means that the between-round revision work is only visible for a small subset of the researchers, and the paired nature of the reviews means we cannot separate the effect of being reviewed from the effect of reviewing someone else.

Following each research task and revision, researchers filled out a survey about their work.[7] This survey asked them to report their findings, additional information like sample size and standard errors, and choices made in the process of doing the analysis like sample restrictions, treated-group definitions, estimator, and standard error adjustments. Researchers were also asked to justify why they had made these choices.

[7]Note that the design of this study, and this survey, predates Sarafoglou et al. (2024). We, therefore, do not include questions related to researchers' subjective assessments of topics such as methodology choices and consistency of results.

There are several papers that use the same ACS data set to identify the effect of DACA on various outcomes, although, to the best of our knowledge, no prior work has been done on the effect of DACA on the probability of working full-time. The design used in Tasks 2 and 3 was most directly inspired by Amuedo-Dorantes and Antman (2016), although the designs do not match exactly, and the outcomes of interest are not the same. There is also prior research on topics such as educational and economic attainment, health care use and outcomes, and marriage-partner decisions (Jones 2020; Giuntella and Lonsky 2020; Amuedo-Dorantes and Wang 2024). Researchers are informed that such previous studies exist and that they can optionally look into previous studies for background as they would normally do when performing research, although no specific previous study is listed. The instructions emphasize that any previous study does not constitute a "right answer" that researchers should be trying to match.

More detailed instructions for the research task and description of the limitation between task rounds are in Appendix Section A, and full instructions for each task, as well as post-task survey text and the peer-reviewing instructions, are available online at https://osf.io/9p7j6/, which also offers sufficient information for interested researchers to attempt the tasks themselves. This research design and analysis plan has been preregistered (Pörtner and Huntington-Klein 2022). Analyses that were not preregistered will be noted in the results section as they are performed. Data processing and analysis as well as table and figure creation for this paper were performed using R.[8]

[8]We used the following R packages: **data.table**, **tidyverse**, **rio**, **fixest**, **car**, **modelsummary**, and **vtable** (Barrett et al. 2024; Wickham et al. 2019; Becker et al. 2023; Bergé 2018; Fox and Weisberg 2019; Arel-Bundock 2022; Huntington-Klein 2021).

3 Recruitment and Descriptive Statistics

Researcher recruitment criteria focused on identifying people who have produced applied microeconomic research, including non-academic applied microeconomics research. Researchers qualified for the project if they satisfied any of the following criteria: they are academic faculty working in applied microeconomics; they are a graduate student *and* have a published or forthcoming paper in applied microeconomics; or they hold a PhD *and* work in a job where they write non-academic reports using tools from applied microeconomics to estimate causal effects.[9] Participation was not limited on the basis of country, career stage, or demographics such as sex, race, or sexual or gender identity.

For our simulation-based power analysis, we assumed that each research task would have 5% smaller between-researcher variation in effects than the previous round and determined the statistical power needed to detect a linear relationship between task number and the squared deviation of effects (variance of estimated effects across researchers). With 90 researchers finishing all tasks, we would have 90% power to detect this effect. For comparisons of only two different research tasks, 90 researchers would give 85% power to detect a decline in variance from one stage to the next of 15% or more, a reasonable effect size given previous many-analyst studies. We further assumed that attrition rates would be roughly 50%, which would suggest recruiting 180 eligible researchers to achieve adequate power. We revised that goal to 200 to account for our assumptions potentially being optimistic and obtained funding to support payments to 200 researchers.

The project was advertised to potential researchers through three avenues: (1) social media posts on Twitter and LinkedIn, (2) emails to professional organizations, and (3) emails to United States economics department chairs. For emails to department chairs, we gathered

[9]This qualification allows those employed in, for example, central banks, the World Bank, and private sector research to participate.

the list of all 286 economics departments listed in the U.S. News and World Report. We emailed the 264 departments for which we could locate email addresses for a front desk or (preferably) department chair, asking for the message to be passed on to all relevant faculty. The recruitment message described the project and its goals, and provided a link to a website (https://nickch-k.github.io/ManyEconomists/) with further detail on project expectations and incentives for participation, and a link to a survey to determine eligibility for the project. As incentives for participation we offered, upon completion of all three tasks, a $2,000 payment for up to 200 of the participants and authorship on the eventual paper.

3.1 Participation and Attrition

A total of 362 people submitted applications for the project (Table 1 shows signups and attrition). Of those, 18.51% were ineligible for the project. Most ineligible people were graduate students without a forthcoming paper. This left 295 eligible participants, which was in excess of the 200 the available budget allowed for. We, therefore, randomly ordered the 282 participants who had signed up by the original due date, and added the 13 late signups at the end of this list. The first 200 on the list were told that they would be paid if they completed all stages of the project, but were not told their order. Everyone with a number above 200 were given their place in the list, and informed that they would be paid if they completed all stages of the project and sufficient numbers of those below them did not complete all steps. For example, if someone was number 206, payment was conditional on at least six participants with a lower number completing all steps.

Our initial assumption that attrition rates would be near 50% was almost exactly correct, with 49.49% of these initial 295 eligible researchers completing all three stages. Nearly all of the attrition occurred by the completion of Task 1. After 141 eligible researchers failed to complete Task 1, only a further 8 failed to complete Task 3. This means we have 146 researchers who

Table 1: Participation and Attrition

Round	Participants	Attrition
Original Signup	362	18.51%
Assigned Task 1	295	47.80%
The first replication task	154	2.60%
The second replication task	150	2.67%
The third replication task	146	

completed all three research tasks, well above the goal of 90, and that we were able to pay all participants who completed all steps.

The high recruitment numbers and the fact that nearly all attrition occurs before Task 1 is complete allow us to evaluate the impact of the payment incentive.[10] One potential concern with our incentive design is that payment and authorship are offered to anyone who completes all tasks, regardless of the quality of their work. We evaluate whether being guaranteed payment affects the probability of completing Task 1 using a regression discontinuity-style design. Researchers below the cutoff were not told their order, so we use a zero-order polynomial (average only) below the cutoff. The effect of the cutoff on completion rates is insignificant and positive using a linear slope above the cutoff and insignificant and negative using a quadratic specification above the cutoff. We also find no effect if we drop the late sign-ups from the regression discontinuity analysis.[11] This is not strong evidence that participants were simply signing up in an attempt to get a $2,000 payment for little effort.

[10] Seven researchers indicated that they did not want payment on their consent form. Of those, three finished the project, while the other four did not.

[11] Furthermore, Appendix Figure C.1 and Table C.1 show that immediately above the cutoff, completion rates are no different. Furthermore, local-polynomial regression discontinuity estimates of the effect of the cutoff are insignificant.

3.2 Researcher Characteristics

Table 2 shows the characteristics of the recruited sample, and how those characteristics changed with eligibility and attrition. Task 2 is omitted as an attrition stage since so few people dropped out between Task 1 and Task 2. One researcher completed all three research tasks, and appears in the above tables, but their work has been removed from the results in the rest of the paper, because a misunderstanding of the instructions meant that their work did not attempt to estimate the effect of DACA on the probability of employment.

The majority of researchers were recruited via social media. Upon signup, researchers were about 90% confident of their ability to finish all three tasks. Those recruited from social media reported a higher expectation of finishing all three tasks. More-confident researchers were slightly more likely to actually finish with the average confidence rates of those who did finish about 92%.

The majority of eligible researchers (83%) had PhDs. PhD holders were also more likely than other eligible researchers to complete all three tasks. These PhDs are split across faculty (62%) and other non-faculty researchers (22%), both of which were more likely than graduate students to finish all three rounds. Most of the researchers had at least one published paper.[12] About a third of initial researchers, and 40% of the final set of researchers, had done work in either immigration or labor economics, the fields closest to the research task at hand, with 5% having done work in both, although all researchers had done work in applied microeconomics generally.

[12] Researchers in the "faculty" or "non-faculty researchers" categories who do not hold PhDs were either people who had been hired to faculty roles without holding PhDs (such as ABDs, or people in a faculty position requiring only a Master's degree), or people with Master's degrees in non-faculty research positions who had published academic papers (some of whom were still graduate students). Researchers with "No Academic Papers" are non-academic researchers who produce work not intended for academic journal publication. Those with "No Published Academic Papers" have papers that are forthcoming, or are faculty who only have working papers and no publications.

Table 2: Researcher Recruitment Source, Professional Experience, and Demographics

	Round											
	Original Signup			Assigned Task 1			Finished Task 1			Finished Task 3		
Variable	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD
Recruitment												
Recruitment Source	347			285			150			142		
... Social media	270	78%		224	79%		124	83%		116	82%	
... Department email	31	9%		28	10%		13	9%		13	9%	
... Professional organization email	15	4%		10	4%		4	3%		4	3%	
... Other	31	9%		23	8%		9	6%		9	6%	
Certainty to Finish Task 1	355	90	11	292	90	10	153	92	8.4	145	92	8.3
Certainty to Finish Task 3	355	89	12	292	89	12	153	91	9.9	145	91	9.6
Professional Experience												
Degree	360			295			154			146		
... No graduate school	3	1%		0	0%		0	0%		0	0%	
... Some Grad School	14	4%		5	2%		3	2%		2	1%	
... Master's degree	78	22%		44	15%		17	11%		17	12%	
... Prof. Degree	3	1%		1	0%		0	0%		0	0%	
... PhD	262	73%		245	83%		134	87%		127	87%	
Occupation	361			295			154			146		
... Faculty	191	53%		182	62%		99	64%		98	67%	
... Grad. Student	69	19%		36	12%		13	8%		12	8%	
... Other	14	4%		11	4%		5	3%		3	2%	
... Other Researcher	87	24%		66	22%		37	24%		33	23%	
Research Experience	361			295			154			146		
... 1-5 Papers in Applied Micro	162	45%		152	52%		74	48%		70	48%	
... 6+ Papers	104	29%		102	35%		58	38%		57	39%	
... No Academic Papers	17	5%		4	1%		3	2%		3	2%	
... No Published Academic Papers	78	22%		37	13%		19	12%		16	11%	
Field	360			294			154			146		
... Immigration & Labor	27	8%		24	8%		9	6%		8	5%	
... Immigration	8	2%		6	2%		4	3%		4	3%	
... Labor	102	28%		85	29%		49	32%		47	32%	
... Neither	223	62%		179	61%		92	60%		87	60%	
Demographics												
Gender	359			294			154			146		
... Female	81	23%		64	22%		28	18%		26	18%	
... Male	274	76%		230	78%		126	82%		120	82%	
... Non-binary / third gender	1	0%		0	0%		0	0%		0	0%	
... Prefer not to say	3	1%		0	0%		0	0%		0	0%	
Race	360			294			154			146		
... White	188	52%		164	56%		100	65%		97	66%	
... Asian	79	22%		60	20%		25	16%		25	17%	
... Black or African American	27	8%		21	7%		4	3%		4	3%	
... Hispanic	25	7%		19	6%		10	6%		9	6%	
... Other or Multiracial	41	11%		30	10%		15	10%		11	8%	
LGBTQ+	360			294			154			146		
... Yes	18	5%		14	5%		7	5%		7	5%	
... No	323	90%		268	91%		137	89%		129	88%	
... Prefer not to say	19	5%		12	4%		10	6%		10	7%	

Note: Results for Tasks 2 are omitted because only four researchers dropped out between Task 1 and Task 2. Full results are available upon request.

The original enrollment was just under 80% male and more than 50% white, with the white share growing to 66% by the end of Task 3. The 80% male figure is similar to the share male found for faculty at a selected set of top economics departments in 2017 by Lundberg and Stearns (2019), and among all actively publishing economists in 2019 by Card et al. (2022). About half of the sample was situated in the United States, and about half was from another country. The representativeness of the racial mixture is difficult to assess for this reason; 66% white would be low if the entire sample were from the United States (Stansbury and Schultz 2023), but it is unclear what the population rate is in a 50% US/50% other location sample. Aside from being skewed towards the United States, the sample largely reflects the group of people who publish work in applied microeconomics. The US overrepresentation is likely driven by the emails sent to US economics departments, that the project was advertised and carried out in English, and that the project organizers are in the United States and advertised the project using their own social media.

4 Results

This section examines variation in effects, samples, and methods across researchers and conditions. We first establish that such variation exists and then evaluate potential explanations for it. Specifically, we describe the distribution of estimated effects and researcher choices and test our preregistered hypotheses.

Our preregistered hypotheses include the following: (1) the standard deviation of estimated effects, sample sizes, and treated and untreated group sample sizes will decrease from task to task; (2) peer review will lead subsequent estimates and sample sizes to become more similar to both the group as a whole and the reviewer's estimates; and (3) the standard deviation of

reported effects will exceed the mean reported standard error. A detailed description of the preregistered hypotheses and analyses is provided in Appendix Section B.

Importantly, the results are based on survey responses from researchers regarding their findings and methodological decisions. The project organizers did not cross-check these responses against researchers' actual coding, meaning there is no guarantee of consistency between the two.[13] Consequently, the variation presented here reflects what readers might encounter in published study descriptions. Any discrepancies arising from coding errors or misrepresentations in research reports are beyond the scope of this analysis but could be explored in future research.

The distribution of estimated effects, reported standard errors, and the size of the sample used, both overall and for the treated group are in Table 3. The effect distributions are shown in two ways: unweighted and using inverse-standard-error weights.[14] Several data points are dropped from the weighted analysis for researchers who did not report standard errors or reported zero. Other missing values are researchers who did not repond to a given question. The lower number of responses for the treated-group sample size question in Task 3 is due to researchers who skipped it because they assumed the answer was obvious.

4.1 Variation Across Researchers

The average effects are relatively similar across the three Tasks, but there is an increase in researcher agreement measured by the inter-quartile range (IQR) when we provided pre-cleaned data in Tasks 3. In Task 1, the mean unweighted estimated effect of DACA eligibility on the

[13]The exception is a small number of cases where the survey response could not be interpreted.

[14]The use of inverse-standard-error weights is not preregistered but follows meta-analytic standards, reducing the influence of estimates that may be outliers due to being estimated with a highly-noisy method, under the suggestion of Auspurg and Brüderl (2023). Weights are truncated at the 95th percentile (200, or a standard error of .005) so as to avoid any single researcher having too much influence on results. Not using the truncation leads to more agreement because a few researchers with very small standard errors make up a significant share of the weighted sample.

Table 3: Distribution of Reported Effects and Sample Sizes

Variable	N	Mean	SD	Min	Pctl. 25	Median	Pctl. 75	Max
Round: Task 1								
Effect Size (Unweighted)	145	0.053	0.095	-0.049	0.014	0.030	0.051	0.660
Effect Size (Weighted)	138	0.044	0.092	-0.049	0.012	0.026	0.043	0.660
Standard Error	139	0.019	0.055	0.000	0.005	0.007	0.013	0.460
Sample Size	145	828,318	3,056,037	681	61,600	179,960	356,787	29,536,580
Treated-Group Size	141	96,395	648,493	270	17,950	34,435	52,581	7,727,201
Round: Task 2								
Effect Size (Unweighted)	145	0.044	0.100	-0.390	0.015	0.032	0.058	0.850
Effect Size (Weighted)	141	0.046	0.069	-0.090	0.018	0.034	0.058	0.850
Standard Error	141	0.031	0.078	0.001	0.010	0.014	0.020	0.744
Sample Size	144	157,006	1,065,593	6,196	18,981	25,414	48,125	12,609,847
Treated-Group Size	140	31,948	221,175	3,519	5,953	11,157	15,832	2,627,183
Round: Task 3								
Effect Size (Unweighted)	145	0.045	0.101	-0.810	0.031	0.050	0.058	0.650
Effect Size (Weighted)	142	0.062	0.103	-0.810	0.036	0.051	0.060	0.650
Standard Error	144	0.059	0.268	0.000	0.015	0.018	0.026	2.747
Sample Size	145	16,904	1,756	7,833	17,379	17,382	17,382	17,832
Treated-Group Size	129	9,433	3,008	11	5,149	11,382	11,382	17,383

probability of working full-time was .053, which was above the 75th percentile because of high top-end estimates, with the unweighted median estimate .030. However, even with substantial researcher freedom, there was a reasonable amount of agreement outside the tails. The 25th to 75th percentile range of the unweighted effect was .014 to .051, an IQR of .037, or 3.7 percentage points in the effect.[15] Task 2 shows less agreement than Task 1, despite giving researchers less freedom, with the unweighted IQRs increasing to .043 and the coefficient of variation (CV) increasing from 1.7 to 2.3. For Task 3, agreement increases between researchers, with the 25th and 75th percentile unweighted effects .031 and .058 (IQR .027 with a median of 0.05), although the CV only declines from 2.3 to 2.2. Hence, from Round 1 to Round 3 we see considerable increases in agreement between researchers, although there are still substantial outliers.

The changes in sample and treated-group sizes across tasks reflect, to a large extent, the same

[15]The use of weights narrows the distribution of effects across all tasks: researchers reporting smaller standard errors also reported estimates that were more similar to each other.

pattern as for effect size. For tasks 1, the 25th and 75th sample size percentiles ranged from 61,600 to 356,787, with some researchers using millions of observations. In the absence of a specified control group, some researchers used nearly the entire ACS sample, including people very unlike the DACA-eligible group. Opposite the effect size, the IQR did reduce considerably in Task 2, where the instructions specified a treated and comparison group, although the 75th percentile (48,125) is still double the 25th (18,981) and some researchers still used millions of observations, resulting in an increase in the CV from 3.7 in Tasks 1 to 6.8 in Task 2.

The imposition of a shared definition for the treated group reduced the treated-group IQR from 34,631 in Taks 1 to 9,879 in Task 2. Theoretically, since there was a shared definition of the treated group in Task 2, the treated-group sample size should be similar in Tasks 2 and 3.[16] That they are not indicates that not all instructions were implemented in the same way across researchers, which will be explored further in Section 4.2. Despite a shared understanding of who was eligible for DACA and who should be in the treated group, only a shared data preparation that correctly implemented these rules for people led to sharp agreement in the size of the treated-groups sample.

With the conflicting changes between the first two tasks and the substantial number of outliers in both effect and sample sizes, Figure 1 and Figure 2 show the distributions of effects and sample sizes, respectively, across the three tasks.[17] For Task 2, both the the effect and sample size distributions are somewhat bimodal; for the effect size especially when weighted. One of these modes appears to be researchers reporting effect estimates and sample size of a similar level to those in Task 1, and others reporting effect estimates similar to what would later be found in Task 3. The bimodality is still present in Tasks 3, but with much more agreement

[16]Variation in the treated-group size in Task 3 is affected by researcher confusion in responding to the survey question. The survey question instructed researchers to not count individuals eligible for DACA as treated for the purposes of this question if they were in a pre-DACA year. However, many researchers counted these individuals as treated anyway, leading to variation in the Task 3 distribution, even though every researcher is at this point working with the same eligibility indicator.

[17]For sample size, the x-axes are on a log scale and that Task 3 is not shown in the graph because the sample is pre-specified.

and density at the higher mode. The decline in agreement and the bimodal result for Task 2 will be investigated further in Section 4.4.

Reported standard errors increase substantially from round to round despite relatively minor changes in average effects, which was driven primarily by the research design specification narrowing the samples used.[18] Figure 3 shows the reported effect sizes ranked from smallest to largest for each round, together with the calculated confidence intervals for each effect size based on the reported standard error. The distribution of effects narrows across rounds, as shown by the flatter specification curve, but confidence intervals increase across rounds and there are fewer statistically significant effects. Throughout, while there is general agreement on effect size in the middle of the distribution, researchers vary in whether the reported effect is statistically significant, with 78%, 60%, and 64% reporting results that were statistically significantly different from 0 in Tasks 1, 2, and 3, respectively. Those effects that are away from the modal effects are substantially more likely to have very large confidence intervals in Tasks 2 and 3, while there are relatively few studies with very large confidence intervals in Task 1.

The increasing agreement in effect size is necessarily driven by individual researchers changing their reported effects in subsequent rounds, but researchers were effectively unbound by their previous estimates as Figure 4 shows. There is little visible or linear statistical relationship between a researcher's reported effects in one task and the next. Only between Tasks 1 and 2 is there a statistically significant correlation, and that correlation is less than 0.2.

In addition to the preregistered descriptive analysis above, we preregistered a set of tests on

[18] Comparing our average standard errors to the variation in reported effects, as in Huntington-Klein et al. (2021) and Menkveld et al. (2024), suggests that reported standard errors alone substantially understate total uncertainty. Calculating this ratio across all three tasks shows a decline in the ratio of variance to mean standard error. However, this decline occurs partly because smaller shared samples in later rounds inflate the standard errors (increasing the denominator). Hence, more generally, because researcher variation need not scale in parallel with standard errors, using these ratios to capture researcher-induced uncertainty can be misleading.

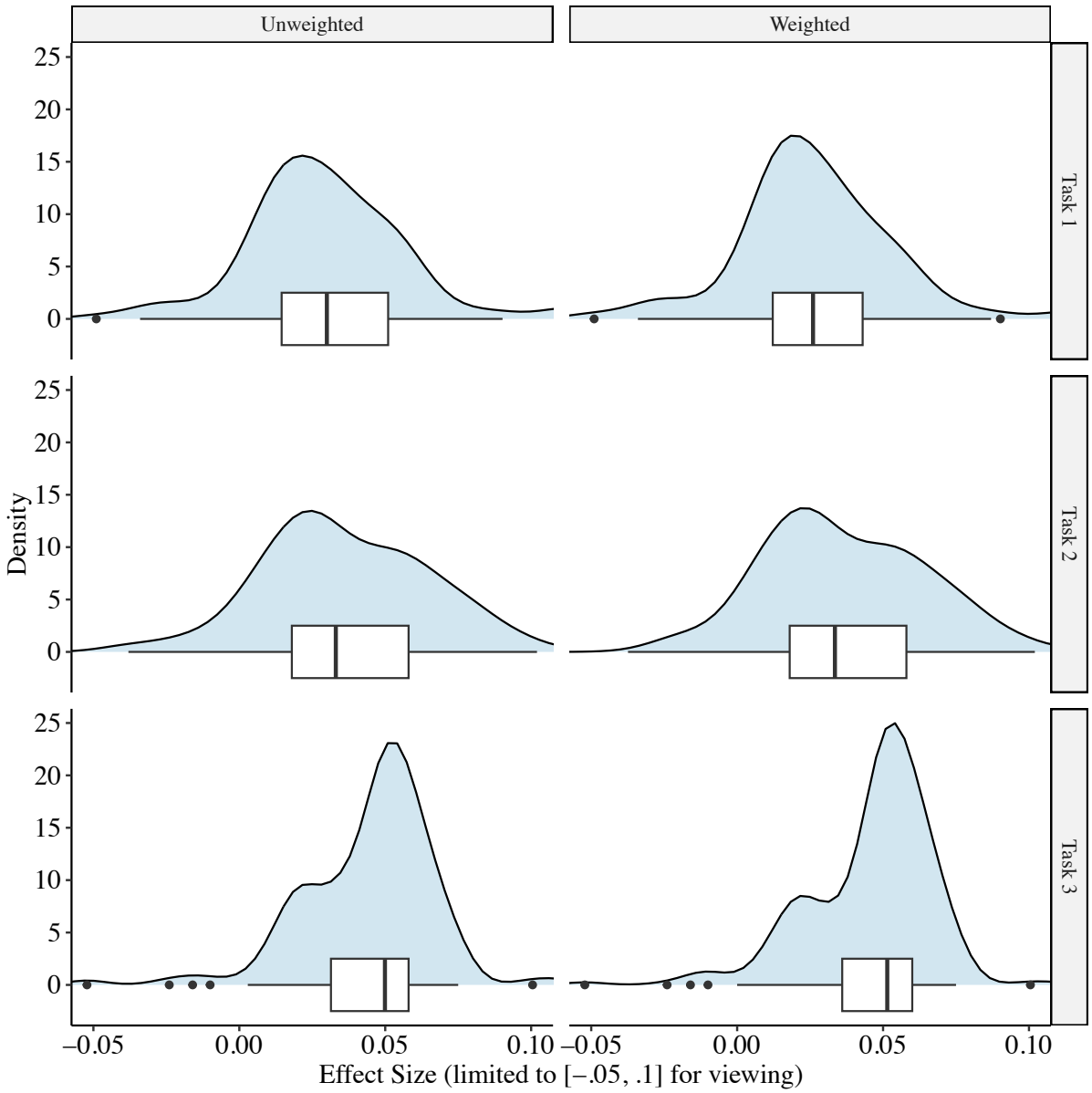


Figure 1: Distributions of Reported Effect Sizes by Task With the Weighted Distributions Using Inverse-Standard-Error Weights

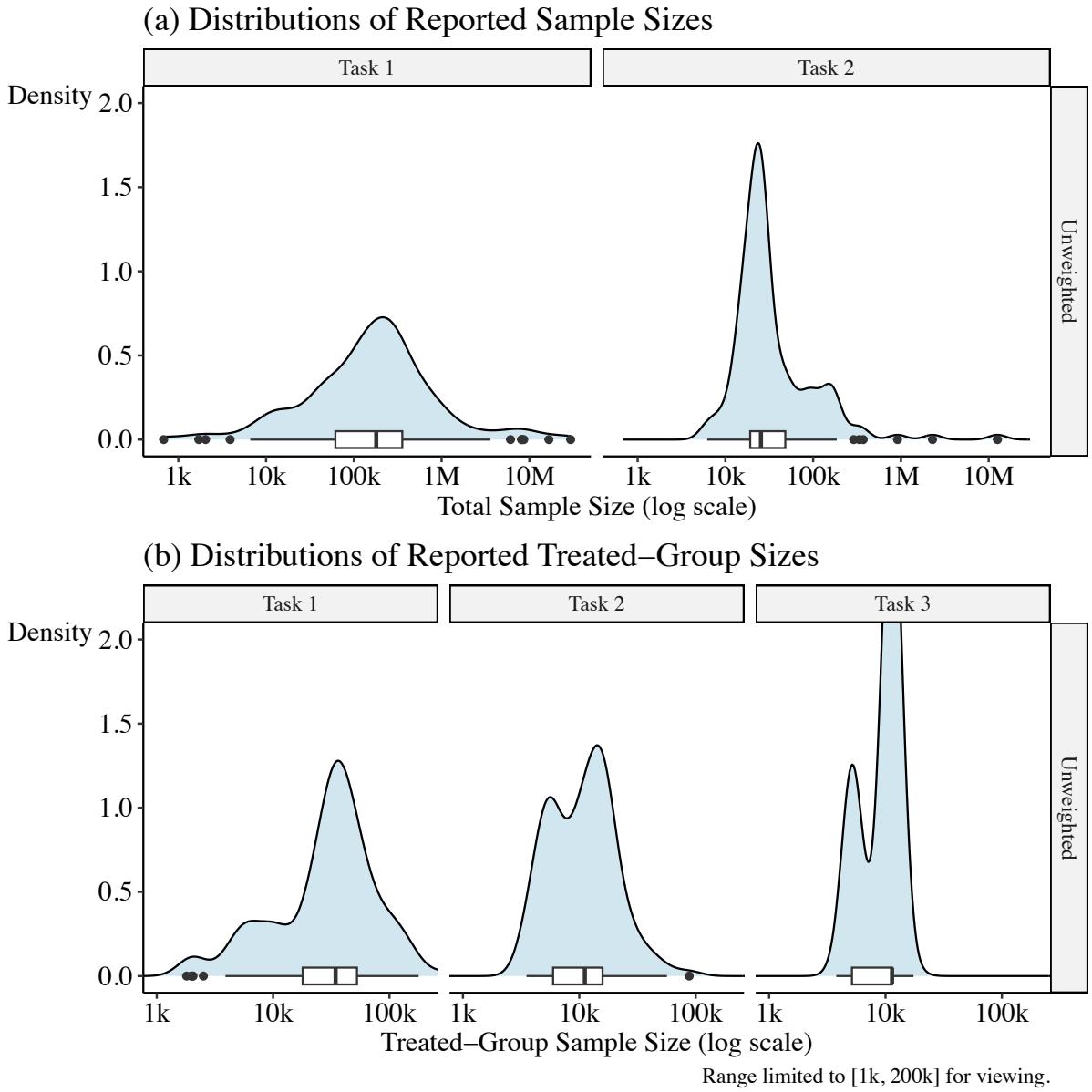


Figure 2: Distributions of Reported Sample Sizes and Treated-Group Sizes

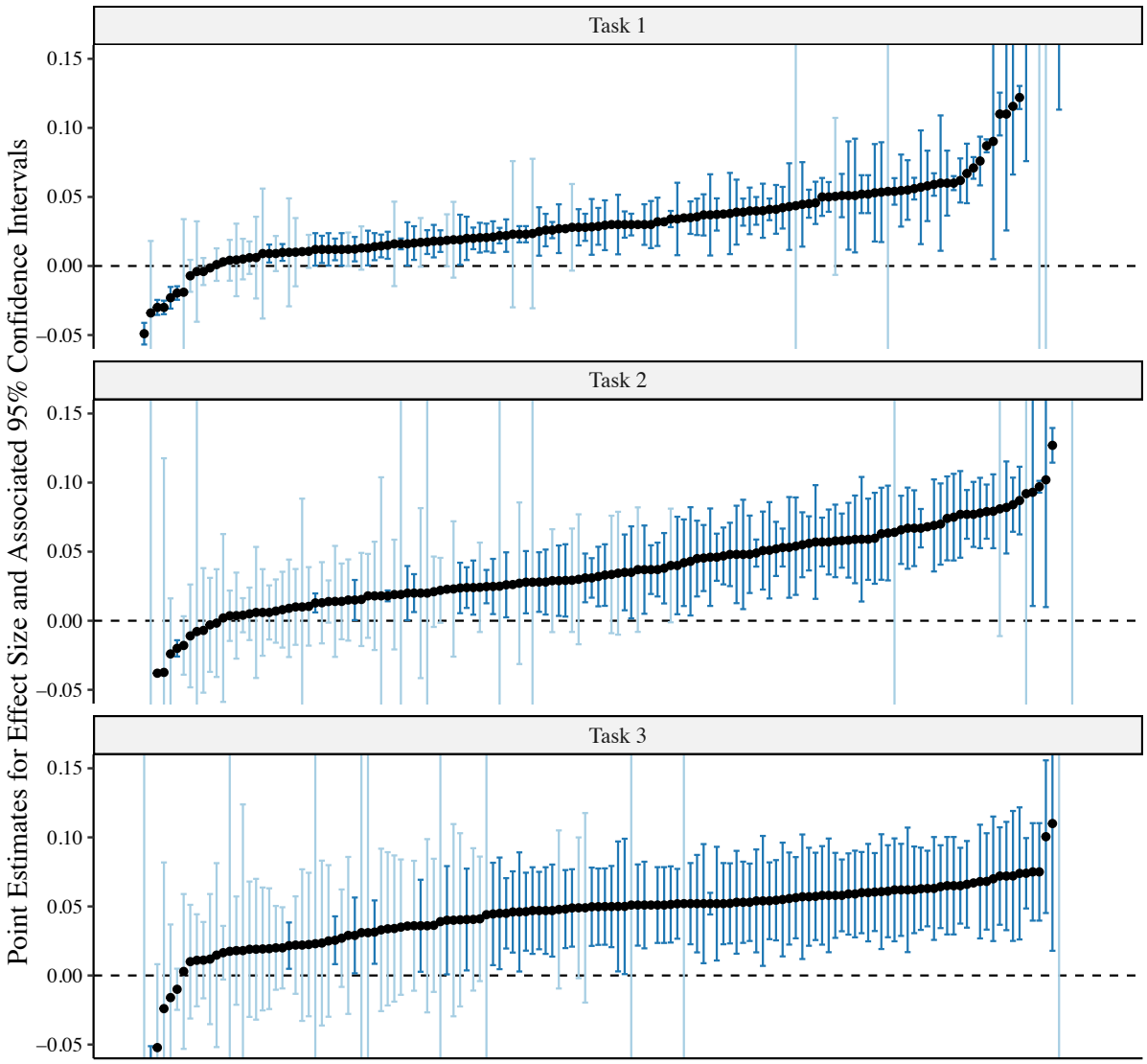


Figure 3: Specification Curve for All Reported Estimates by Task with Estimates Ordered From Smallest to Largest

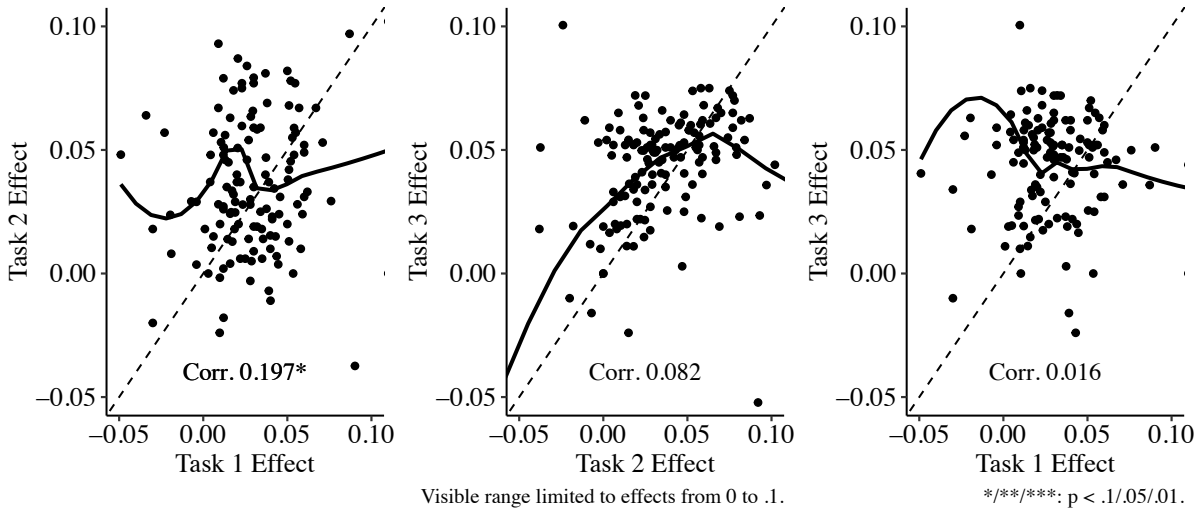


Figure 4: Same-Researcher Effect Sizes Across Tasks

effect and sample sizes. For effect size, we do not reject any of the preregistered Levene tests for the variation at the 95% level, where the null hypothesis is no change in variance from any stage to any later stage (including a comparison of each task to its revision stage, and comparing each main task to later main tasks). The lowest p-value of 0.197 comes from the comparison of Task 1 to Task 1 Revision. We do, however, reject the null hypothesis of equal variance across Task 1 and Task 2 at the 0.05 level for the sample size variance, consistent with the expected reduction when we specify the treated and comparison groups.

Furthermore, we preregistered regressing the squared difference to the round means against the round number. The squared differences to the mean in each round which provide us with a measure of the variance in effect and samples sizes across researchers. None of the coefficients on round are statistically significant, although the coefficients for sample sizes are negative as expected (See Appendix Table C.2). Finally, as shown in Appendix Section Section D, there is no statistically significant difference in variance between the reviewed and non-reviewed groups, nor does peer review demonstrate a consistent effect on results. Similarly, there are no statistically significant differences in the variance of sample sizes between the peer-reviewed

and non-peer-reviewed groups in the follow-up tasks.

4.2 Researchers' Analytic and Sample Choices

The standard approach when evaluating research, for example when peer reviewing a paper, is to focus on the researchers' analytical choices, such as estimation method and associated decisions. Given the differences in samples above, this section examines to what extent the different choices that researchers made in both analytical and sample choices can explain the variation in effects and sample sizes.

For most researchers, the choices of estimator, use of ACS sampling weights, or standard error adjustment did not change across tasks, and Table 4 therefore shows the combined choices, which generally are in line with common practices in applied econometric work. The dependent variable is binary, but linear regression was the most common estimator with 82% of entries, and 13% using logit or probit. In many cases, these linear regressions used a fully saturated (or nearly fully saturated) difference-in-differences design, which mutes the downsides of linear probability models. Other researchers used a matching estimator (sometimes combined with linear regression) or one of several newly-introduced estimators for difference-in-differences designs, like Callaway and Sant'Anna (2021). Despite IPUMS recommendation, only 25% used weights. The standard error adjustment varied considerably, with a slim majority clustering standard errors in some way, although at a range of different levels, and 17% used heteroskedasticity-robust but not cluster-robust standard errors.

The choices shown in Table 4 modestly explain estimated effects. Regressing effect sizes on the full set of indicators, which is not preregistered, produces an R^2 value of .162 for Task 1, .073 in Task 2, and .023 in Task 3. In Tasks 2 and 3, these choices more heavily influence whether the result is statistically significant at the 95% level, with R^2 values of .026 in Task 1, .227

Table 4: Estimation Methods

Variable	N	Percent	Variable	N	Percent
Method	437		S.E. Adjustment	438	
... Linear Regression	358	82%	... Cluster (State)	118	27%
... Logit/Probit	57	13%	... Cluster (State & Year)	58	13%
... Matching	11	3%	... Cluster (ID/Strata/Other)	65	15%
... New DID Estimator	7	2%	... Het-Robust	76	17%
... Other	4	1%	... Other/Bootstrap	23	5%
Weights	438		... None	98	22%
... No Sample Weights	329	75%			
... Sample Weights	109	25%			

Notes: This table shows details on estimation, not research design. "Difference-in-differences" implemented with linear regression, for example, counts here as linear regression.

in Task 2, and .338 in Task 3. The choice of estimation method drove most of the variation in significance (and was a significant predictor in Tasks 2 and 3), followed by standard error adjustment (which was significant only in Task 3).

The researchers disagreed substantially on the appropriate controls.[19] Across the 435 submissions there were 333 different unique sets of included covariates, with 64% choosing a set of covariates that no other researcher chose. Only those with *no* controls shared a covariate set with more than three other people, while 12% shared with two or three other people, and 17% shared with one other person. The most common included controls were for state, year, age, and sex, which more than 50% included in all three tasks. However, there was a large amount of variation in the sets of included covariates. In Task 1, for example, there are ten covariates with inclusion rates between .2 and .8, meaning that at least 20% of the researchers made a different decision on inclusion of the covariate than the majority.

This lack of agreement across researchers did, however, not substantially impact the effect estimates. The mean reported effects differ by only .023 percentage points when we compare

[19]Appendix Table C.3 shows the average rate of inclusion of covariates, as well as the estimated effects among analyses including those controls, in order of average effect size. Variables are included regardless of the functional form used to include them.

Table 5: Estimated Effects by Functional Form of Control Variable

			Effect		SE
Category	Control	N	Mean	SD	Mean
AGE	Linear Age	164	0.058	0.107	0.024
AGE	Age FE	36	0.024	0.022	0.040
AGE	Age Quadratic	33	0.035	0.089	0.015
EDUC	Linear Education	122	0.040	0.066	0.016
EDUC	Education FE	32	0.047	0.033	0.021
EDUC	Education Transform	61	0.045	0.064	0.017
STATE/YEAR	Linear Year	79	0.044	0.140	0.037
STATE/YEAR	Year FE	103	0.047	0.062	0.026
STATE/YEAR	State FE	155	0.046	0.102	0.031
STATE/YEAR	State FE x Year FE	56	0.037	0.027	0.018
STATE/YEAR	State FE x Linear Year	23	0.061	0.133	0.017

Note: SD is the standard deviation of reported effect estimates among all estimates including the listed functional form. SE is the mean of all standard errors reported for those estimates.

the covariate with the highest average effect estimates (Continuous Years in the USA) against the lowest (Race). This comparison likely overstates the impact of covariate selection since selecting the highest versus the lowest after estimates are known will bias towards a larger difference from noise alone. There do not appear to be major differences in the average reported standard errors either, or in the standard deviation of the effect distribution among researchers.

The chosen functional form explained more variation in average effects than the covariate choices, as shown in Table 5. For both age and the State/Year controls, the difference between the highest average-effect functional form variants and the lowest was greater than the difference between highest and lowest for covariates.

As Table 6 shows, there is large variation in which variables researchers based their sample selection on and how these variables were implemented, including in Task 2, where there is a

correct answer according to the instructions.[20] For each variable, the most-common option, listed at the top, is the "correct" answer for defining the treated group, with two exceptions. First, for Citizenship, there is a second justifiable answer because those who are "Non-Citizen or Naturalized After 2012" would have been eligible for DACA in 2012, but not when they were surveyed and therefore would have received a partial "dose" of DACA. Second, for "Years Continuous in USA," DACA requires that the immigrant have lived *continuously* in the United States for five years as of 2012, but most researchers used only year of immigration being before 2007 to satisfy this criterion, while others used the YRSUSA set of variables which specifically track living continuously in the country. For all other variables besides "Years Continuous in USA," the option matching the instructions was the most common, but we see plenty of variation. We also see considerable variation when there is not a clear "correct" option, like the analytic sample definition, where no specific usage of any one variable was used by more than 84% of the sample.

Showing the impact of these choices on estimated effects is difficult for any one variable because any specific alternative to the most common option has too few people using it to make a reasonable comparison. The two comparisons for which an alternative was common enough to compare are for the YRSUSA inclusion and the use of "< 2007" vs. "<= 2007" for year of migration, shown in 7. These choices are not associated with large differences in estimated effects. Effect differences are larger in Task 2. However, in Task 1, even though estimated effects are similar, sample sizes are considerably larger for the less-restrictive option, and so reported standard errors would be lower, and statistical significance more likely. For other

[20]This is the only part of the paper that does not rely on researcher responses to the survey. For each researcher's Task 1 and Task 2 code, organizers read the code directly and recorded some aspects of the sample definitions used for the overall analytic sample and for the definition of the treated group, including definitions that appeared to be the result of coding errors. The table allows for coding errors. For example the individuals reporting that they used only high school graduates or *non-veterans,* instead of veterans as per the instructions, likely did not intentionally choose to use non-veterans but rather coded "VETSTAT == 1," which indicates "non-veteran", perhaps based on a misunderstanding of the IPUMS documentation (veterans are VETSTAT == 2). However, an earlier version of this paper relied on researcher self-reports of sample limitations in the survey, and found similar rates at which Task 2 choices did not match the "correct answer", so coding errors alone do not account for these results.

Table 6: Sample Restriction Methods

	Task 1				Task 2			
	All		Treated		All		Treated	
Variable	N	Percent	N	Percent	N	Percent	N	Percent
Hispanic	144		144		144		144	
... Hispanic-Mexican	105	73%	109	76%	112	78%	113	78%
... Hispanic-Any	17	12%	17	12%	13	9%	13	9%
... Hispanic-Mex or Mex-Born	1	1%	2	1%	1	1%	1	1%
... None	21	15%	16	11%	18	12%	17	12%
Birthplace	145		145		145		145	
... Mexican-Born	103	71%	112	77%	114	79%	116	80%
... Hispanic-Mex or Mex-Born	2	1%	2	1%	1	1%	2	1%
... Non-US Born	4	3%	4	3%	3	2%	3	2%
... Central America-Born	1	1%	1	1%	1	1%	1	1%
... None	35	24%	26	18%	26	18%	23	16%
Citizenship	145		145		145		145	
... Non-Citizen	83	57%	117	81%	104	72%	118	81%
... Foreign-Born	2	1%	2	1%	2	1%	2	1%
... Non-Cit or Natlzd post-2012	4	3%	7	5%	7	5%	8	6%
... Other	11	8%	11	8%	6	4%	8	6%
... None	45	31%	8	6%	26	18%	9	6%
Age at Migration	145		145		145		145	
... < 16	21	14%	105	72%	77	53%	111	77%
... <= 16	10	7%	25	17%	18	12%	21	14%
... Other	24	17%	11	8%	8	6%	7	5%
... None	90	62%	4	3%	42	29%	6	4%
Age in June 2012	145		145		145		145	
... Year-Quarter Age	40	28%	117	81%	92	63%	118	81%
... Year-Only Age	18	12%	21	14%	22	15%	24	17%
... Other	2	1%	0	0%	0	0%	0	0%
... None	85	59%	7	5%	31	21%	3	2%
Year of Immigration	145		145		145		145	
... < 2007	15	10%	43	30%	34	23%	44	30%
... <= 2007	13	9%	52	36%	44	30%	58	40%
... < 2012	3	2%	1	1%	2	1%	1	1%
... <= 2012	2	1%	4	3%	2	1%	3	2%
... Any Year	7	5%	4	3%	3	2%	2	1%
... Other	5	3%	3	2%	0	0%	1	1%
... None	100	69%	38	26%	60	41%	36	25%
Education/Veteran	145		145		145		145	
... HS Grad or Veteran	0	0%	3	2%	85	59%	108	74%
... 12th Grade or Veteran	0	0%	0	0%	3	2%	3	2%
... HS Grad	13	9%	21	14%	6	4%	8	6%
... HS Grad or Non-Veteran	0	0%	0	0%	3	2%	4	3%
... Other	3	2%	6	4%	9	6%	11	8%
... None	129	89%	115	79%	39	27%	11	8%
Years Continuous in USA	145		145		145		145	
... Used YRSUSA	23	16%	55	38%	39	27%	55	38%
... No YRSUSA	122	84%	90	62%	106	73%	90	62%

Note: The table does not cover the full set of possible variables used to define samples. Some common limitations used by some researchers and not others include filtering out people living in group quarters or those out of the labor force, or dropping anyone with a recorded year of immigration before their recorded year of birth. Many researchers also chose to limit the sample based on current age as of the year of their inclusion in the ACS, as opposed to their age in 2012, which is shown, choosing many different acceptable age ranges.

Table 7: Effect and Samples by Sample Definitions

	Treated-Group Restrictions			All-Sample Restrictions					
	Effect Percentile			Effect Percentile			Sample Size Percentile		
Variable	25	50	75	25	50	75	25	50	75
Task 1									
Year of Immigration									
... < 2007	0.016	0.030	0.052	0.013	0.028	0.042	13,222	31,878	57,192
... <= 2007	0.013	0.029	0.052	0.019	0.037	0.057	44,073	96,406	209,528
Years Continuous in USA									
... Used YRSUSA	0.017	0.030	0.053	0.017	0.026	0.045	41,450	141,847	367,300
... No YRSUSA	0.012	0.030	0.046	0.014	0.030	0.053	67,068	190,052	352,245
Task 2									
Year of Immigration									
... < 2007	0.017	0.028	0.053	0.018	0.030	0.058	21,988	24,263	28,345
... <= 2007	0.018	0.034	0.056	0.022	0.038	0.060	22,398	25,588	32,630
Years Continuous in USA									
... Used YRSUSA	0.018	0.037	0.059	0.016	0.034	0.058	19,562	25,134	42,951
... No YRSUSA	0.015	0.029	0.057	0.016	0.031	0.057	18,750	25,639	49,356

comparisons, there are large differences in estimated effects and sample sizes across many of the different sample restriction choices, although in most cases these comparisons are based on very small samples (See Appendix Tables C.4 for Task 1 and Table C.5 for Task 2).

That the inclusion of different covariates did not have a major impact on estimated effects or that the choice of functional form had a greater impact than the selection of covariates likely do not generalize, but is specific to this research task. However, there is clearly substantial variation across researchers in what they believe the appropriate set of covariates should be and, for a given covariate, what the appropriate functional form is. We also see that, in the case of this particular study, these decisions did not fully explain the variation in effects between researchers.

4.3 Researcher Characteristics and Effects

As listed in our preregistration, the two project organizers individually evaluated the relationship between researcher characteristics and the effects they reported using a multiple-analysts approach, with the two project organizers taking the same data and research question and performing independent analyses. The preregistration called for independent analysis of the relationship between (a) researcher characteristics and reported research results in earlier stages, and (b) attrition from the study and reported research results in later stages. Because there was so little attrition from the study after Task 1, part b was dropped from the analysis. Full results from each project organizer can be found in Appendix B.

The two project organizers took very different approaches to the question of how researcher characteristics affected results, selecting different dependent variables and methods of analysis, and different sets of researcher characteristics. Despite this, both organizers found that researcher characteristics were not strong predictors of estimated effects. Across researcher demographics, occupation, and professional experience, there was no strong relationship between researcher background and either the level of the effect estimate they reported, the deviation of their estimate from the mean, or changes in their estimate from task to task. The only relevant difference we found is that the minority of researchers who used the R programming language were more likely to report outlier estimates than researchers who used Stata.

4.4 Bimodality in the Task 2 Effect Estimates

When designing the study, we expected that each task would show a narrower distribution of effects than the previous task, but while we see this pattern for sample sizes and some researcher choices, the distribution of effects became wider between Tasks 1 and 2. There is also an emerging bimodality in both effects and samples sizes, where most of the researchers

reported estimates that reflected the distribution of effects seen in Task 1, while a smaller group of researchers reported larger effects more like those in Task 3. This un-preregistered analysis examines potential explanations for these unexpected findings.

A major contributing factor to Task 2 bimodality is the ability to precisely implement the treated-group definition given in the instructions. Task 2 gave a very precise definition of who should be included as a part of the treated group, and we examine whether a given researcher followed the full set of treated-group definition instructions exactly or not. Any mismatch could be small, such as using "<= 16" instead of "< 16" for age at migration, or large, such as omitting that eligible people must be non-citizens. Figure 5 shows the distribution of effects for researchers who follow the definition precisely against researchers who had a mismatch in their criteria in any way. The graph indicates that the bimodality is driven by the group that precisely matched the treated-group definition. This implies that the bimodality in Task 2 may be explained in large part by a split between researchers who exactly followed the instructions, and so were more likely to match what a typical researcher found in Task 3, and those who did not.

The treated-group implementation does not fully explain researcher behavior. There are many other decisions made, and the treated-group implementation captures only one angle. Furthermore, the share of researchers matching exactly across all fields is fairly low at 20-25% as shown by Table 8, although recall that even very minor mismatches are counted as mismatches. Perfect-match rates were slightly higher among researchers whose work was closest to the field that the research task was in, immigration and labor, although this difference was not statistically significant at the 95% level.

Several other anticipated correlates did not explain the bimodal outcomes of Task 2. The Task 2 reported sample sizes and standard errors do not strongly explain the effects reported (See Appendix Figure C.2 and Appendix Figure C.3). The bimodality is also not a feature of some

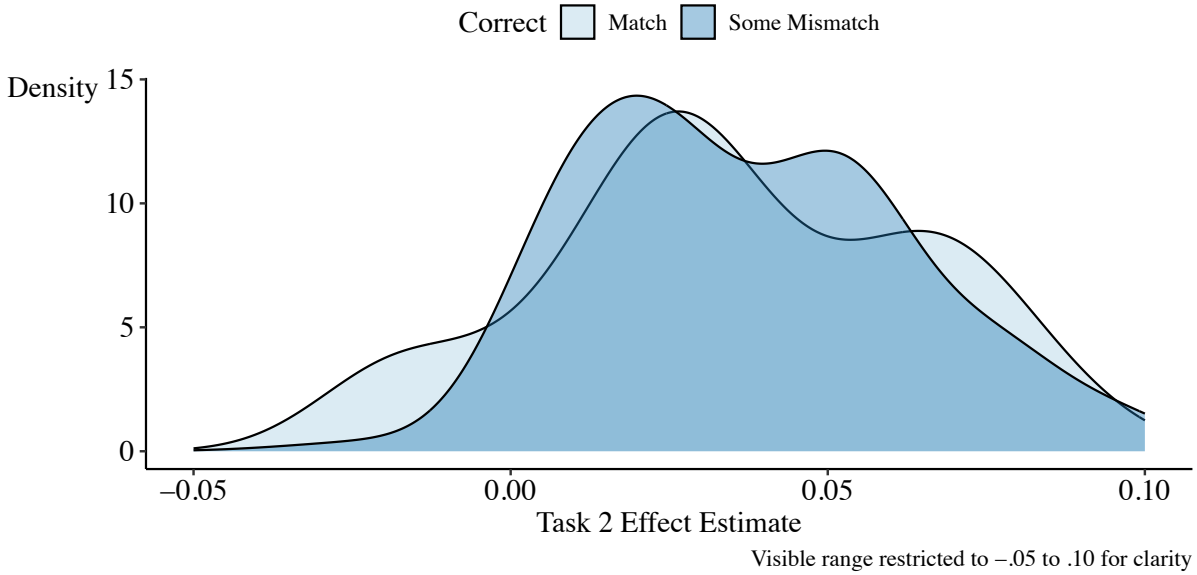


Figure 5: Task 2 Effect Distributions Among Those with Exact Treated-Group Definition Matchs vs. Those with Some Mismatch

Table 8: Share of Researchers Matching Treated-Group Definition Exactly by Field

Field	Share Match	Num. Match	Share Some Mismatch	Num Some Mismatch
Immigration & Labor	0.0%	0	100.0%	8
Immigration	100.0%	4	0.0%	0
Labor	31.9%	15	68.1%	32
Neither/Other	32.6%	28	67.4%	58

researchers trying to make their Task 2 results consistent with their Task 1, as Figure 4 in Section 4.1 shows.

5 Conclusion

We had 146 research teams perform the same causal inference task three times: first with few constraints, then using a shared research design, and finally with pre-cleaned data in addition to a specified design. We found substantial variation in researchers' choices, particularly in data cleaning and processing, research design, the definition of treated and comparison groups, and the selection and functional form of controls. Some of this variation appears to stem from data cleaning processes that do not align with the instructions for constructing the treated group. Variation was not strongly constrained by peer review or a shared research design, though providing pre-cleaned data did reduce the variation. However, there were also aspects where researchers behaved similarly. Linear regression modeling was widely used, and very few researchers employed unadjusted standard errors, although the specific adjustment or clustering level varied substantially.

Despite the variation in data preparation and modeling, researchers reported policy effects that were relatively similar, at least in the center of the distribution. In Task 1, where researchers had full freedom, the interquartile range (IQR) of policy impacts was only 3.1 percentage points, though substantial outlier estimates lay outside this range. Task 2 showed less agreement than Task 1, with an IQR of 4.0 percentage points, driven by some researchers not fully adhering to the specified research design. In Task 3, where data was pre-cleaned and errors in data preparation were eliminated, the IQR fell to its lowest level at 2.4 percentage points—an improvement in agreement, though not statistically significant. Specifying a research design considerably improved agreement in reported sample sizes, with the IQR decreasing from

295,187 in Task 1 to 29,144 in Task 2 and effectively 0 in Task 3. In contrast to these changes, we found no effect of peer review or researcher background or experience on reported policy effects.

The fact that different researchers approach the same research question differently is not inherently problematic, as long as disagreements are visible to readers, open to scrutiny, and understood as part of a broader discourse. However, problems arise when researcher variation reflects either (a) errors or (b) unexamined or invisible choices. In our study, when "standard" approaches existed—such as using linear modeling in a difference-in-differences setting with a binary outcome or adjusting standard errors—researchers tended to follow them. In the absence of well-established standards—such as in the choice of clustering level, covariate selection in this particular setting, or data cleaning—researchers diverged, sometimes with consequential effects and sometimes without significant impact.

A key result of this study is that the absence of standards in aspects of the research process, such as data cleaning, can lead to arbitrary variation. By "standards," we refer both to the process by which choices are made *and* to how these choices are reported. For data cleaning, for example, neither standardized practices nor transparent reporting of those practices are common. In principle, the use of replication packages and well-documented code can address the transparency issue, but in many cases, this information is either unavailable or not sufficiently examined in research discussions. Even when provided, there is often little emphasis on understanding *how* these choices influence research outcomes.

The optimal level of researcher variation is not zero, as individual researchers often have valid reasons for deviating from the methods and practices of others. However, such deviations should occur because there *is* a good reason to depart from an established template—not because no template exists in the first place. Without formal training in PhD programs or a

culture of reviewing and critiquing data cleaning and preprocessing in research papers, substantial unexplained inter-researcher variation will persist. Among broader systemic changes, incorporating data cleaning and preprocessing courses into the standard PhD applied economics curriculum could improve research quality and reduce researcher variability.[21]

This discussion presumes the existence of best practices, as codifying standards without empirical validation could reduce variability around wrong answers. However, the crucial point is that once an effort is made to formalize and disseminate best practices—much like how applied economists routinely learn about modeling—we establish a basis for evaluating and refining those practices.[22] Other disciplines have already made progress in this direction (e.g., Osborne 2012; Jafari 2022). Thus, economics would not need to start from scratch but could refine existing recommendations to suit applied microeconomics. The inclusion of discussions on data-cleaning best practices in at least one recent textbook is a step in this direction (Békés and Kézdi 2021).

Researchers are accustomed to critiquing research design and modeling choices, and we expect these decisions to be clearly documented in research writeups. What is less common, however, is testing whether different analytical choices yield different results. In this study, for instance, seemingly minor decisions—such as the functional form of covariates—proved more consequential than the choice of which covariates to include. This suggests an important future role for multiverse analysis, where researchers systematically assess the impact of alternative modeling decisions, or for many-analyst approaches, as in Section 4.3, when conducting original research (Steegen et al. 2016). Journals could also consider publishing studies that explore variations in analytical approaches to existing research, even if these are not framed as replications or direct

[21] While we have highlighted data-cleaning practices as a particularly fruitful area for standardization, similar efforts are needed in other areas, such as clustering levels. An example of progress in this direction is found in Abadie et al. (2023).

[22] Similar to how researchers learn about modeling through econometrics textbooks or applied literature (Abadie et al. 2023).

challenges to prior findings—a current barrier to publishing replications (Galiani, Gertler, and Romero 2017).

At a minimum, researchers should document their data-cleaning processes as thoroughly as they describe their modeling choices. Ideally, arbitrary data-cleaning decisions should also be subjected to multiverse analysis (as suggested by Steegen et al. 2016). Moreover, while an increasing number of journals require replication packages (for example, American Economic Association 2024), these packages often start from a pre-processed dataset and only include code for running statistical models. Requiring the inclusion of data preprocessing code in replication packages—and making this code accessible to peer reviewers and readers—would enhance transparency and accountability. In short, economics should treat data cleaning and preprocessing as just as critical to the research process as model selection.

References

Abadie, Alberto et al. (2023). "When Should You Adjust Standard Errors for Clustering?" In: *The Quarterly Journal of Economics* 138.1, pp. 1–35.

American Economic Association (2024). *Data and Code Availability Policy*. Accessed on July 10, 2024. URL: https://www.aeaweb.org/journals/data/data-code-policy.

Amuedo-Dorantes, Catalina and Francisca Antman (2016). "Can Authorization Reduce Poverty among Undocumented Immigrants? Evidence from the Deferred Action for Childhood Arrivals Program". In: *Economics Letters* 147, pp. 1–4.

Amuedo-Dorantes, Catalina and Chunbei Wang (2024). "Intermarriage amid immigration status uncertainty: Evidence from DACA". In: *Journal of Policy Analysis and Management*. DOI: https://doi.org/10.1002/pam.22640.

Ankel-Peters, Jörg, Nathan Fiala, and Florian Neubauer (2023). "Do Economists Replicate?" In: *Journal of Economic Behavior & Organization* 212, pp. 219–232.

Arel-Bundock, Vincent (2022). "modelsummary: Data and Model Summaries in R". In: *Journal of Statistical Software* 103.1, pp. 1–23. DOI: 10.18637/jss.v103.i01.

Auspurg, Katrin and Josef Brüderl (2021). "Has the Credibility of the Social Sciences Been Credibly Destroyed? Reanalyzing the "Many Analysts, One Data Set" Project". In: *Socius* 7.

– (2023). "Is Social Research Really Not Better than Alchemy? How Many-analysts Studies Produce "A Hidden Universe of Uncertainty" by Not Following Meta-analytical Standards". In.

Barrett, Tyson et al. (2024). *data.table: Extension of data.frame*. R package version 1.15.4. URL: https://r-datatable.com.

Bastiaansen, Jojanneke A et al. (2020). "Time to get Personal? The Impact of Researchers Choices on the Selection of Treatment Targets Using the Experience Sampling Msethodology". In: *Journal of Psychosomatic Research* 137, p. 110211.

Becker, Jason et al. (2023). *rio: A Swiss-Army Knife for Data I/O*. R package version 1.0.1. URL: https://github.com/gesistsa/rio.

Békés, Gábor and Gábor Kézdi (2021). *Data Analysis for Business, Economics, and Policy*. Cambridge, UK: Cambridge University Press.

Bergé, Laurent (2018). "Efficient estimation of maximum likelihood models with multiple fixed-effects: the R package FENmlm". In: *CREA Discussion Papers* 13.

Black, Bernard S et al. (2024). "The Road to False Positives: Sample Selection and Specification Choice in Randomized and Natural Experiments". In: *Available at SSRN 5024145*.

Boehm, Udo et al. (2018). "Estimating Across-trial Variability Parameters of the Diffusion Decision Model: Expert Advice and Recommendations". In: *Journal of Mathematical Psychology* 87, pp. 46–75.

Borjas, George J and Nate Breznau (Dec. 2024). *Ideological Bias in Estimates of the Impact of Immigration*. Working Paper 33274. National Bureau of Economic Research. DOI: 10.3386/w33274. URL: http://www.nber.org/papers/w33274.

Botvinik-Nezer, Rotem et al. (2020). "Variability in the Analysis of a Single Neuroimaging Dataset by Many Teams". In: *Nature* 582.7810, pp. 84–88.

Breznau, Nate et al. (2021). "Observing Many Researchers Using the Same Data and Hypothesis Reveals a Hidden Universe of Data Analysis". In: *MetaArXiv Preprints*.

Broderick, Tamara, Ryan Giordano, and Rachael Meager (2020). "An Automatic Finite-sample Robustness Metric: When can Dropping a Little Data Make a Big Difference?" In: *arXiv preprint arXiv:2011.14999*.

Brodeur, Abel, Nikolai Cook, and Anthony Heyes (Nov. 2020). "Methods Matter: p-Hacking and Publication Bias in Causal Analysis in Economics". In: *American Economic Review* 110.11, pp. 3634–60. DOI: 10.1257/aer.20190687.

Brodeur, Abel, Mathias Lé, et al. (Jan. 2016). "Star Wars: The Empirics Strike Back". In: *American Economic Journal: Applied Economics* 8.1, pp. 1–32. DOI: 10.1257/app.20150044.

Bryan, Christopher J, David S Yeager, and Joseph M O'Brien (2019). "Replicator Degrees of Freedom Allow Publication of Misleading Failures to Replicate". In: *Proceedings of the National Academy of Sciences* 116.51, pp. 25535–25545.

Callaway, Brantly and Pedro HC Sant'Anna (2021). "Difference-in-differences with Multiple Time Periods". In: *Journal of Econometrics* 225.2, pp. 200–230.

Camerer, Colin F et al. (2016). "Evaluating Replicability of Laboratory Experiments in Economics". In: *Science* 351.6280, pp. 1433–1436.

Card, David et al. (2022). "Gender differences in Peer Recognition by Economists". In: *Econometrica* 90.5, pp. 1937–1971.

Chen, Wanyi and Mary Cummings (2024). "Subjectivity in Unsupervised Machine Learning Model Selection". In: *Proceedings of the AAAI Symposium Series*. Vol. 3. 1, pp. 22–29.

Fox, John and Sanford Weisberg (2019). *An R Companion to Applied Regression*. Third. Thousand Oaks CA: Sage. URL: https://socialsciences.mcmaster.ca/jfox/Books/Companion/.

Galiani, Sebastian, Paul Gertler, and Mauricio Romero (2017). *Incentives for Replication in Economics*. Tech. rep. National Bureau of Economic Research.

Giuntella, Osea and Jakub Lonsky (2020). "The effects of DACA on health insurance, access to care, and health outcomes". In: *Journal of Health Economics* 72, p. 102320. ISSN: 0167-6296. DOI: https://doi.org/10.1016/j.jhealeco.2020.102320.

Gould, Elliot et al. (2023). "Same Data, Different Analysts: Variation in Effect Sizes Due to Analytical Decisions in Ecology and Evolutionary Biology". In.

Herbert, Sylvérie et al. (Dec. 2021). *The Reproducibility of Economics Research: A Case Study*. Working Paper 853. Paris, France: Banque de France.

Holzmeister, Felix et al. (2023). *Heterogeneity in effect size estimates: Empirical evidence and practical implications*. Working Papers in Economics and Statistics 2023-17. Innsbruck: University of Innsbruck, Research Platform Empirical and Experimental Economics (eeecon).

Hoogeveen, Suzanne et al. (2023). "A Many-analysts Approach to the Relation between Religiosity and Well-being". In: *Religion, Brain & Behavior* 13.3, pp. 237–283.

Huntington-Klein, Nick (2021). *vtable: Variable Table for Variable Documentation*. R package version 1.4.6. URL: https://nickch-k.github.io/vtable/.

Huntington-Klein, Nick et al. (2021). "The Influence of Hidden Researcher Decisions in Applied Microeconomics". In: *Economic Inquiry* 59.3, pp. 944–960.

Jafari, Roy (2022). *Hands-On Data Preprocessing in Python: Learn how to effectively prepare data for successful data analytics*. Packt Publishing Ltd.

Jelveh, Zubin, Bruce Kogut, and Suresh Naidu (2024). "Political Language in Economics". In: *The Economic Journal*, ueae026.

Jones, Richard C. (2020). "A Time-Space Stream of DACA Benefits and Barriers Gleaned From the American Community Survey". In: *Hispanic Journal of Behavioral Sciences* 42.2, pp. 143–164. DOI: 10.1177/0739986320915849.

Klau, Simon et al. (2023). "Comparing the Vibration of Effects due to Model, Data Preprocessing and Sampling Uncertainty on a Large Data Set in Personality Psychology". In: *Meta-Psychology* 7.

Lang, Kevin (Forthcoming). "How Credible is the Credibility Revolution?" In: *Journal of Labor Economics*.

Leamer, Edward E (1983). "Let's Take the Con out of Econometrics". In: *The American Economic Review* 73.1, pp. 31–43.

Lundberg, Shelly and Jenna Stearns (2019). "Women in Economics: Stalled Progress". In: *Journal of Economic Perspectives* 33.1, pp. 3–22.

Magnus, Jan R. and Mary S. Morgan (1997). "Design of the Experiment". In: *Journal of Applied Econometrics* 12.5, pp. 459–465. (Visited on 12/17/2024).

Menkveld, Albert J. et al. (2024). "Nonstandard Errors". In: *The Journal of Finance* 79.3, pp. 2339–2390.

Open Science Collaboration (2015). "Estimating the Reproducibility of Psychological Science". In: *Science* 349.6251, aac4716.

Ortloff, Anna-Marie et al. (2023). "Different researchers, different results? analyzing the influence of researcher experience and data type during qualitative analysis of an interview and survey study on security advice". In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pp. 1–21.

Osborne, Jason W (2012). *Best Practices in Data Cleaning: A Complete Guide to Everything You Need to Do Before and After Collecting your Data*. Sage Publications.

Ostropolets, Anna et al. (2023). "Reproducible variability: assessing investigator discordance across 9 research teams attempting to reproduce the same observational study". In: *Journal of the American Medical Informatics Association* 30.5, pp. 859–868.

Pörtner, Claus C. and Nick Huntington-Klein (Oct. 2022). *Many Economists*. DOI: 10.17605/OSF.IO/CJ9YX. URL: osf.io/cj9yx.

Ruggles, Steven et al. (2024). *IPUMS USA: Version 15.0 [dataset]. Minneapolis, MN: IPUMS*.

Sarafoglou, Alexandra et al. (2024). "Subjective Evidence Evaluation Survey for Many-Analysts Studies". In: *Royal Society Open Science* 11.7, p. 240125. DOI: 10.1098/rsos.240125.

Schweinsberg, Martin et al. (2021). "Same Data, Different Conclusions: Radical Dispersion in Empirical Results when Independent Analysts Operationalize and Test the Same Hypothesis". In: *Organizational Behavior and Human Decision Processes* 165, pp. 228–249.

Silberzahn, Raphael et al. (2018). "Many Analysts, One Data Set: Making Transparent how Variations in Analytic Choices Affect Results". In: *Advances in Methods and Practices in Psychological Science* 1.3, pp. 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". In: *Psychological science* 22.11, pp. 1359–1366.

Stansbury, Anna and Robert Schultz (2023). "The Economics Profession's Socioeconomic Diversity Problem". In: *Journal of Economic Perspectives* 37.4, pp. 207–230.

Steegen, Sara et al. (2016). "Increasing Transparency Through a Multiverse Analysis". In: *Perspectives on Psychological Science* 11.5, pp. 702–712.

Sulik, Justin et al. (2023). "Why Do Scientists Disagree?" In.

U.S. Citizenship and Immigration Services (2016). *Number of I-821D, Consideration of Deferred Action for Childhood Arrivals by Fiscal Year, Quarter, Intake, Biometrics and Case Status: 2012-2016 (June 30)*.

Urban Institute (2022). *State Immigration Policy Resource*. URL: https://www.urban.org/data-tools/state-immigration-policy-resource.

Wickham, Hadley et al. (2019). "Welcome to the tidyverse". In: *Journal of Open Source Software* 4.43, p. 1686. DOI: 10.21105/joss.01686.

Appendix A: Research Task Description

This section outlines the detail of the research tasks and the differences between Tasks 1, 2, and 3. Full instructions are available in the online appendix/pre-registration at https://osf.io/9p7j6/.

In all research tasks, the specific goal given to researchers was:

> Among ethnically Hispanic-Mexican Mexican-born people living in the United States, what was the causal impact of eligibility for the Deferred Action for Childhood Arrivals (DACA) program (treatment) on the probability that the eligible person is employed full-time (outcome), defined as usually working 35 hours per week or more?
>
> DACA was implemented in 2012. Examine the effects on full-time employment in the years 2013-2016.

In simple terms, this asks researchers to estimate the impact of the DACA program on the probability that those eligible for the program usually work 35 hours per week or more in the years 2013-2016.

Researchers, many of whom are not from the United States and so may not be familiar with DACA, are given further background information about the DACA program:

- DACA allowed undocumented immigrants who were accepted into the program to have legal work authorization for two years without fear of deportation, and also allowed them to apply for drivers' licenses or other forms of identification. People could reapply after the two years expired, and many did.

- Applications for the program opened on August 15, 2012, and over the first four years of the program's existence, over 900,000 applications were received, about 90% of which were approved.(U.S. Citizenship and Immigration Services 2016)

- While the program was not specific to immigrants from any origin country, because of the structure of undocumented immigration to the United States, the great majority of eligible people were from Mexico.

Researchers were also given information on the eligibility criteria for DACA, which was intended to apply only to a specific subset of undocumented immigants who arrived in the United States as children, and not to all undocumented immigrants. Eligible people must:

- Have arrived in the United States before their 16th birthday.

- Not have had their 31st birthday as of June 15, 2012.

- Have lived continuously in the United States since June 15, 2007.

- Were present in the United States on June 15, 2012 and did not yet have legal status (either citizenship or legal residency) during that time.

An additional eligibility requirement was mistakenly omitted from the Task 1 instructions, but was included for Tasks 2 and 3:

- Eligible people must have completed at least high school (12th grade) or be a veteran of the military.

In addition to this information about the policy itself and the effect that researchers are supposed to identify, researchers were also given instructions about the data set to use and how to procure it, as well as some details on usage of the data:

- Data should come from the American Community Survey (ACS), using data no older than 2006, and no newer than 2016.

- In addition, a file of state/year-level data was provided including labor market data and the presence or absence of different immigration policies in different years. Immigration policy data comes from Urban Institute (2022).[23]

- ACS data should be procured from the IPUMS website (Ruggles et al. 2024), specifically selecting one-year ACS files and harmonized variables. Written and video instructions were included showing how to select data samples and variables on the IPUMS website.

- Researchers were not told which specific variables to use to determine eligibility status, but they were given guidance onto how to find relevant variables (like looking at the Person → Race, Ethnicity, and Nativity page to find variables relevant to ethnicity, birthplace, citizenship, and year of immigration).

- Several relevant features of the ACS that may affect analysis were emphasized: (a) ACS is a repeated cross-section, not a year-to-year panel data set, and (b) ACS does not list the month that data was collected in, so it is not possible to distinguish whether a given observation in 2012 is from before or after the policy was implemented, and (c) we do not actually observe in ACS whether a given person is enrolled in DACA, so we assume that all eligible people who are ethnically Mexican and Mexican-born are treated.

Finally, researchers were instructed to keep track of any variables used to limit their sample download on IPUMS, and to review the survey where they would be reporting their results before beginning their analysis.

[23]This file included the state/year-level unemployment rate and labor force participation rate. Immigration policy flags were for policies for undocumented immigrants to get state drivers' licenses, to get college financial aid, to be banned from state public colleges, or to follow Omnibus immigration legislation that serves to increase the surveillance of immigation documentation. Additional indicators were for participation in E-Verify laws that require employers to verify immigration authorization, to limit E-Verify participation, participation in Secure Communities, and for participation in task-force or jail based 287(g) policies.

From there, researchers were given free reign to complete the analysis as they thought most appropriate, including their own choice of statistical software, an instruction to use assistants for any work that they might normally use assistants for, and asking them to complete the analysis as they thought best, as though the research task had been their own idea, not trying to match or not-match other researchers or guess what analyses the project organizers wanted to see. Once finished, they uploaded all of their code and data to a Sharepoint website, wrote a short description and interpretation of their results focusing on a single "headline" result, and filled out the research survey to report their results.

For Task 2, all of the previous instructions remained in place, but several were added to further specify the research design:

- There is a "treated" group that is comprised of all ethnically Mexican and Mexican-born non-citizen individuals who are aged 26-30 on June 15, 2012 (recall that individuals must not have had their 31st birthday as of June 15, 2012 to be eligible for DACA).
- There is an "untreated" group that is comprised of people who would have been eligible for DACA, except that they were aged 31-35 on June 15, 2012.
- Researchers should estimate the effect of treatment by seeing how the 26-30 group changed from before treatment to after relative to how the 31-35 group changed (keeping in mind this is a repeated cross-section and not panel data).
- Researchers should attempt to estimate the effect for all individuals in the "treated" group and not, for example, estimate the effect only for men or only for women.
- The instructions specifically mention that researchers can, if they like, use covariates or account for differing trends to improve the comparability of the treated and untreated groups.

The task is otherwise unchanged for Task 2.

In Task 3, the instructions remain unchanged from Task 2, except that the data is provided directly instead of having researchers download data from IPUMS, omitting data from the year of 2012. In Task 3, project organizers cleaned the data, merged in the state policy data, created a variable indiciating whether a given individual was in the "treated" or "untreated" group, limited the sample only to individuals in "treated" or "untreated," and created simplified versions of variables like education. Researchers were instructed not to further limit the sample from this prepared data set, or to perform further extensive data cleaning.[24]

Appendix B: Hypotheses and Analysis Preregistration

This Appendix Section shows only the the preregistrated hypotheses and the associated analysis plan. Both are edited for clarity. For the full preregistration, see https://doi.org/10.17605/OSF.IO/CJ9YX.

B.1: Hypotheses

In each case, SD_i refers to the standard deviation of effect sizes across replicators reported in the ith round of the study:

- Round 1: Initial replication task
- Round 2: Results after peer review and revision
- Round 3: Second replication task
- Round 4: Results after second peer review and revision
- Round 5: Third replication task
- Round 6: Results after third peer review and revision

[24] There were three observations in the final cleaned data set that were missing values of the education variable. The final used sample in Task 3 sometimes differs by 3 across researchers, based on whether the analysis drops these individuals.

Hypotheses:

1. SD_1 will be greater than the mean reported standard error of effect sizes, among studies for which a standard error and single effect size can be derived.
2. SD_i will have a negative linear or quadratic relationship with i.
3. For i from 1 to 5, $SD_{i+1} < SD_i$.
4. Respondents assigned to the peer review condition in round i will have a smaller SD_i than respondents not assigned to peer review, where the round i results for anyone who does not submit a revision in round i is their round $i-1$ result.
5. Respondents assigned to the peer review condition in round i will have a smaller SD_{i+1} than respondents not assigned to peer review.

6. a. For a given peer review pair assigned to review each other in round i, the difference between their results in round i will be smaller than the difference between their results in round $i-1$.

   b. For a given peer review pair assigned to review each other in round i, the difference between their results in round i will be smaller than if they had not been assigned to each other.
7. The hypotheses 2 through 6 will also apply to the analytic sample size, using only rounds 1–4.
8. The hypotheses 2 through 6 will also apply to the number of observations determined to be eligible for DACA and in the analysis, using only rounds 1–4.
9. The hypotheses 2 through 6 will also apply to the number of observations determined to be ineligible for DACA and in the analysis, using only rounds 1–4.

B.2: Analysis Plan

1. SD_1 will be greater than the mean reported standard error of effect sizes, among studies for which a standard error and single effect size can be derived.

   We will calculate the standard deviation of the reported effect sizes in the first stage, and the mean of the reported standard errors in the first stage, and compare them.

2. SD_i will have a negative linear or quadratic relationship with i.

   We will take the reported effect sizes from the set of researchers who completed all stages of analysis, and subtract the mean. Then, we will use ordinary least squares to regress the square of this variable on a linear term representing the round of analysis, or a quadratic for round, depending on the apparent best-fit relationship in the data. If using a quadratic, the hypothesis is supported only if the effect is negative for all values of i in the data.

3. 3a-3e. For i from 1 to 5, $SD_{i+1} < SD_i$

   We will use Levene's test, with a median center, to compare the variance of the effect sizes in each round to the variance of effect sizes in the following round.

4. Respondents assigned to the peer review condition in round i will have a smaller SD_i than respondents not assigned to peer review, where the round i results for anyone who does not submit a revision in round i is their round $i-1$ result.

   We will use Levene's test, with a median center, to compare the variance of the effect sizes in round i among those who were assigned to peer review in round i against those who were not assigned to peer review in round i. This will be repeated for rounds 2, 4, and 6, and also for these three rounds pooled together.

5. Respondents assigned to the peer review condition in round i will have a smaller SD_{i+1} than respondents not assigned to peer review

   We will use Levene's test, with a median center, to compare the variance of the effect sizes in round $i+1$ among those who were assigned to peer review in round i against those who were not assigned to peer review in round i. This will be repeated for rounds 2, 4, and 6, and also for these three rounds pooled together.

6. a. For a given peer review pair assigned to review each other in round i, the difference between their results in round i will be smaller than the difference between their results in round $i-1$.

      We will calculate the absolute difference in effect sizes between each member of a peer review pair in their round i and round $i-1$ results. Then, we will compare the means of these absolute differences across rounds using a paired t-test. This analysis will be repeated in rounds 2, 4, and 6, and also for these three rounds pooled together.

   b. For a given peer review pair assigned to review each other in stage i, the difference between their results in stage i will be smaller than if they had not been assigned to each other.

      We will calculate the absolute difference in effect sizes between each member of a peer review pair in their round i results. Then, we will construct a null distribution of effect size differences by randomly assigning an equal number of fake peer review partnerships and calculating their average within-partnership differences in their round i results (where anyone who did not submit a revision in round i uses their round $i-1$ results). We will repeat this fake-partnership process 3,000 times and calculate the distribution of the mean absolute difference across these 3,000

iterations. We will then calculate the actual absolute difference's percentile pct of the null distribution. The p-value will be $2 * pct$ if $pct < 0.5$, and $2 * (1 - pct)$ if $pct >= 0.5$. This analysis will be repeated in rounds 2, 4, and 6, and also for all three rounds pooled together.

7. The hypotheses 2 through 6 will also apply to the analytic sample size, using only rounds 1-4.

8. The hypotheses 2 through 6 will also apply to the number of observations determined to be eligible for DACA and in the anlaysis, using only rounds 1-4.

9. The hypotheses 2 through 6 will also apply to the number of observations determined to be ineligible for DACA and in the analysis, using only rounds 1-4.

   The analysis will be the exact same as for analyses 2-6, except using overall analytic sample size, the number of observations included and eligible for DACA, and the number of observations included and ineligible for DACA instead of effect sizes, respectively, and limiting analysis only to rounds 1-4.

Appendix C: Additional Figures and Tables

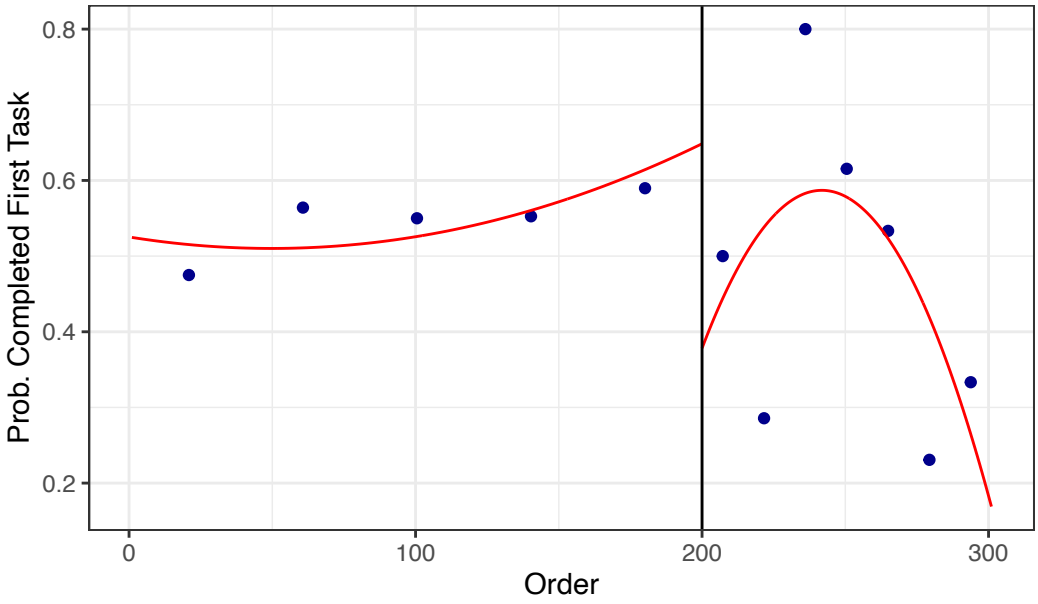


Figure C.1: Impact of Guaranteed Payment on Probability of Task 1 Completion

Table C.1: Linear and Quadratic Regression Discontinuity Estimates

	Linear	Quadratic
Intercept	0.543***	0.543***
	(0.036)	(0.035)
Order above 200	0.067	−0.245
	(0.116)	(0.170)
Linear x Above	−0.003	0.018**
	(0.002)	(0.009)
Squared x Above		0.000**
		(0.000)
Num.Obs.	282	282

Note: Slopes below 200 omitted since respondents below 200 did not know their order. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table C.2: Squared Difference to Round Mean
against Round number

Variable	Estimate	Std. Error	P-Value
Effect Size	0.0005	0.0015	0.7343
Sample Size (Total)	-3,498,932,503,410	2,381,655,690,344	0.1427
Sample Size (DACA)	-159,838,573,988	161,954,749,694	0.3244
Sample Size (Non-DACA)	-1,965,116,310,337	1,519,818,808,048	0.1969

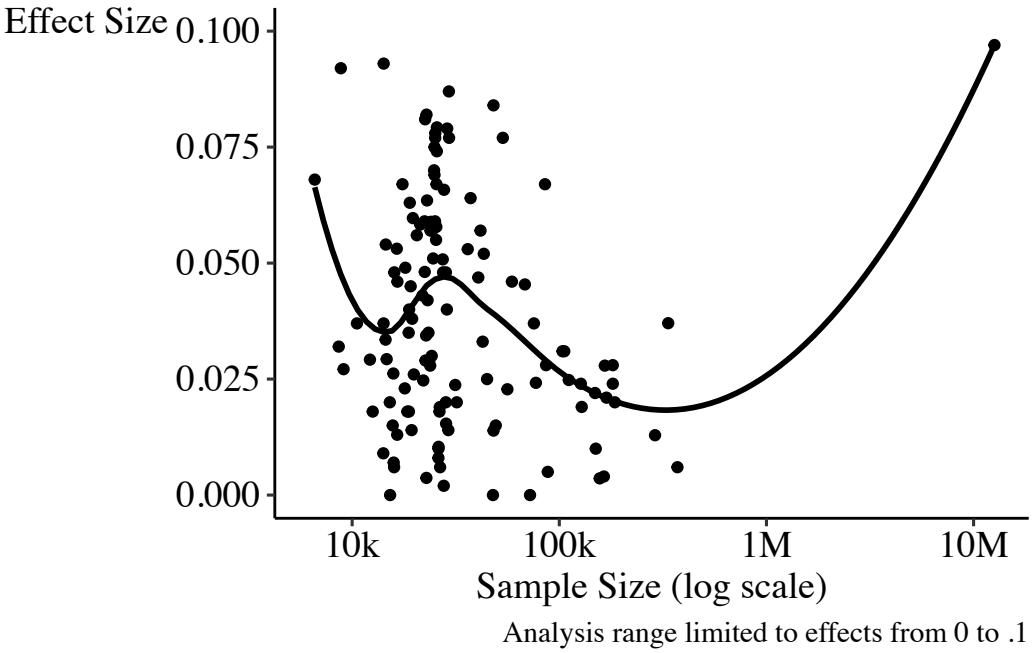


Figure C.2: Task 2 Effect Size and Sample Size

Table C.3: Covariate Inclusion Across Rounds and Estimated Effects

	Rate in			Effect		SE
Control	Task 1	Task 2	Task 3	Size	SD	Mean
Continuous Years in USA	0.13	0.13	0.12	0.054	0.123	0.035
Age	0.62	0.57	0.52	0.048	0.094	0.025
Year of Migration	0.14	0.13	0.11	0.048	0.112	0.033
Marital Status	0.10	0.13	0.12	0.047	0.071	0.016
Sex	0.63	0.64	0.72	0.046	0.101	0.027
Age at Migration	0.18	0.14	0.14	0.045	0.067	0.022
None	0.07	0.07	0.06	0.045	0.133	0.060
State	0.62	0.64	0.64	0.045	0.089	0.025
Year	0.68	0.60	0.57	0.045	0.094	0.026
Education	0.48	0.50	0.51	0.042	0.061	0.017
Other	0.66	0.63	0.63	0.040	0.086	0.038
Age in 2012	0.05	0.04	0.08	0.037	0.042	0.026
State Policy Variables	0.25	0.21	0.23	0.037	0.108	0.033
Unemployment Rate	0.32	0.27	0.30	0.036	0.097	0.033
Labor Force Participation Rate	0.22	0.17	0.20	0.035	0.115	0.041
English Speaker	0.17	0.17	0.23	0.034	0.100	0.046
Race	0.24	0.22	0.28	0.031	0.092	0.032

Notes: SD is the standard deviation of reported effect estimates among all estimates including the listed functional form. SE is the mean of all standard errors reported for those estimates.

Table C.4: Task 1 Effect and Samples by Sample Definitions, Full View

	Treated-Group Restrictions			All-Sample Restrictions					
	Effect Percentile			Effect Percentile			Sample Size Percentile		
Variable	25	50	75	25	50	75	25	50	75
Hispanic									
... Hispanic-Mexican	0.015	0.030	0.052	0.015	0.030	0.052	74,431	173,803	292,492
... Hispanic-Any	0.014	0.030	0.046	0.014	0.030	0.046	13,818	140,134	202,451
... Hispanic-Mex or Mex-Born	-0.042	-0.034	-0.026	-0.019	-0.019	-0.019	287,021	287,021	287,021
... None	0.020	0.032	0.052	0.018	0.026	0.052	61,225	403,130	1,582,703
Birthplace									
... Mexican-Born	0.017	0.030	0.051	0.017	0.030	0.051	58,150	141,847	277,277
... Hispanic-Mex or Mex-Born	-0.042	-0.034	-0.026	-0.009	0.001	0.010	326,913	366,804	406,696
... Non-US Born	-0.003	0.012	0.028	-0.001	0.013	0.028	131,426	171,812	247,497
... Central America-Born	0.057	0.057	0.057	0.057	0.057	0.057	9,711	9,711	9,711
... None	0.011	0.028	0.054	0.010	0.030	0.054	87,186	292,450	654,740
Citizenship									
... Non-Citizen	0.017	0.030	0.052	0.021	0.035	0.056	50,530	155,898	277,277
... Foreign-Born	0.021	0.026	0.032	0.021	0.026	0.032	228,357	360,308	492,260
... Non-Cit or Natlzd post-2012	0.015	0.027	0.037	0.016	0.030	0.045	88,848	159,122	214,588
... Other	0.011	0.023	0.052	0.012	0.027	0.035	15,216	61,225	112,780
... None	0.008	0.012	0.051	0.009	0.013	0.041	123,061	338,042	829,918
Age at Migration									
... < 16	0.017	0.030	0.053	0.017	0.030	0.045	10,973	44,073	127,504
... <= 16	0.010	0.027	0.051	0.009	0.042	0.051	117,536	172,149	204,920
... Other	0.014	0.030	0.041	0.017	0.029	0.051	45,945	112,918	163,604
... None	-0.005	-0.003	0.002	0.013	0.030	0.052	127,918	271,386	482,144
Age in June 2012									
... Year-Quarter Age	0.013	0.029	0.052	0.017	0.028	0.052	32,893	116,240	204,466
... Year-Only Age	0.018	0.030	0.050	0.018	0.039	0.051	48,132	111,882	281,340
... Other				0.014	0.019	0.023	90,418	95,154	99,891
... None	0.025	0.032	0.048	0.014	0.030	0.051	120,931	263,963	485,979
Year of Immigration									
... < 2007	0.016	0.030	0.052	0.013	0.028	0.042	13,222	31,878	57,192
... <= 2007	0.013	0.029	0.052	0.019	0.037	0.057	44,073	96,406	209,528
... < 2012	0.076	0.076	0.076	0.032	0.037	0.057	103,534	132,637	145,394
... <= 2012	0.032	0.150	0.270	0.029	0.092	0.155	263,220	471,364	679,507
... Any Year	0.014	0.024	0.035	0.015	0.030	0.036	82,855	140,134	274,695
... Other	0.008	0.035	0.051	0.028	0.029	0.059	85,681	104,628	123,061
... None	0.015	0.028	0.043	0.014	0.030	0.051	115,558	242,029	452,600
Education/Veteran									
... HS Grad or Veteran	0.190	0.270	0.305						
... HS Grad	0.016	0.022	0.040	0.016	0.039	0.052	62,631	127,504	155,898
... Other	0.016	0.028	0.050	0.016	0.027	0.042	42,071	74,431	139,789
... None	0.014	0.030	0.051	0.014	0.030	0.051	61,600	202,451	391,487
Years Continuous in USA									
... Used YRSUSA	0.017	0.030	0.053	0.017	0.026	0.045	41,450	141,847	367,300
... No YRSUSA	0.012	0.030	0.046	0.014	0.030	0.053	67,068	190,052	352,245

Table C.5: Task 2 Effect and Samples by Sample Definitions, Full View

	Treated-Group Restrictions			All-Sample Restrictions					
	Effect Percentile			Effect Percentile			Sample Size Percentile		
Variable	25	50	75	25	50	75	25	50	75
Hispanic									
... Hispanic-Mexican	0.014	0.029	0.057	0.015	0.029	0.057	19,074	25,176	43,558
... Hispanic-Any	0.018	0.037	0.058	0.018	0.037	0.058	18,803	23,133	25,649
... Hispanic-Mex or Mex-Born	0.048	0.048	0.048	0.048	0.048	0.048	22,416	22,416	22,416
... None	0.027	0.045	0.069	0.023	0.043	0.066	25,088	44,755	128,639
Birthplace									
... Mexican-Born	0.015	0.032	0.057	0.016	0.032	0.056	19,028	25,155	37,028
... Hispanic-Mex or Mex-Born	0.054	0.059	0.065	0.048	0.048	0.048	22,416	22,416	22,416
... Non-US Born	0.023	0.045	0.048	0.048	0.051	0.060	26,116	27,376	47,800
... Central America-Born	0.067	0.067	0.067	0.067	0.067	0.067	25,538	25,538	25,538
... None	0.018	0.029	0.068	0.009	0.027	0.072	18,750	47,848	128,343
Citizenship									
... Non-Citizen	0.018	0.031	0.057	0.018	0.034	0.058	19,174	25,176	42,172
... Foreign-Born	0.023	0.042	0.062	0.023	0.042	0.062	57,916	93,322	128,729
... Non-Cit or Natlzd post-2012	0.014	0.041	0.057	0.014	0.034	0.052	16,182	20,520	25,586
... Other	0.005	0.047	0.061	0.004	0.025	0.056	21,088	31,370	75,323
... None	0.000	0.028	0.048	0.020	0.029	0.048	20,065	37,677	92,659
Age at Migration									
... < 16	0.018	0.034	0.059	0.018	0.037	0.060	19,121	23,912	27,912
... <= 16	0.010	0.024	0.053	0.013	0.030	0.054	19,068	26,916	31,866
... Other	0.019	0.028	0.054	0.020	0.029	0.055	21,230	25,812	61,619
... None	-0.023	0.026	0.057	0.011	0.026	0.049	22,173	63,634	155,043
Age in June 2012									
... Year-Quarter Age	0.018	0.035	0.057	0.021	0.036	0.058	19,064	25,078	41,917
... Year-Only Age	0.017	0.021	0.059	0.018	0.031	0.059	22,061	25,418	48,125
... None	-0.192	0.006	0.042	0.006	0.022	0.046	18,892	27,376	138,560
Year of Immigration									
... < 2007	0.017	0.028	0.053	0.018	0.030	0.058	21,988	24,263	28,345
... <= 2007	0.018	0.034	0.056	0.022	0.038	0.060	22,398	25,588	32,630
... < 2012	0.029	0.029	0.029	0.034	0.038	0.042	21,170	27,663	34,156
... <= 2012	0.066	0.068	0.290	0.065	0.066	0.067	14,281	21,962	29,642
... Any Year	0.013	0.014	0.014	0.014	0.015	0.030	16,122	16,542	153,218
... Other	0.067	0.067	0.067						
... None	0.012	0.036	0.059	0.012	0.028	0.054	18,405	27,376	102,952
Education/Veteran									
... HS Grad or Veteran	0.015	0.032	0.058	0.015	0.035	0.059	19,121	24,787	28,783
... 12th Grade or Veteran	0.043	0.055	0.067	0.043	0.055	0.067	25,532	25,649	64,640
... HS Grad	0.015	0.019	0.042	0.015	0.019	0.035	18,944	20,342	26,829
... HS Grad or Non-Veteran	0.023	0.037	0.048	0.020	0.026	0.037	28,230	40,649	44,394
... Other	0.026	0.037	0.047	0.025	0.037	0.054	18,845	25,538	44,805
... None	0.018	0.029	0.066	0.017	0.028	0.054	19,562	56,976	164,874
Years Continuous in USA									
... Used YRSUSA	0.018	0.037	0.059	0.016	0.034	0.058	19,562	25,134	42,951
... No YRSUSA	0.015	0.029	0.057	0.016	0.031	0.057	18,750	25,639	49,356

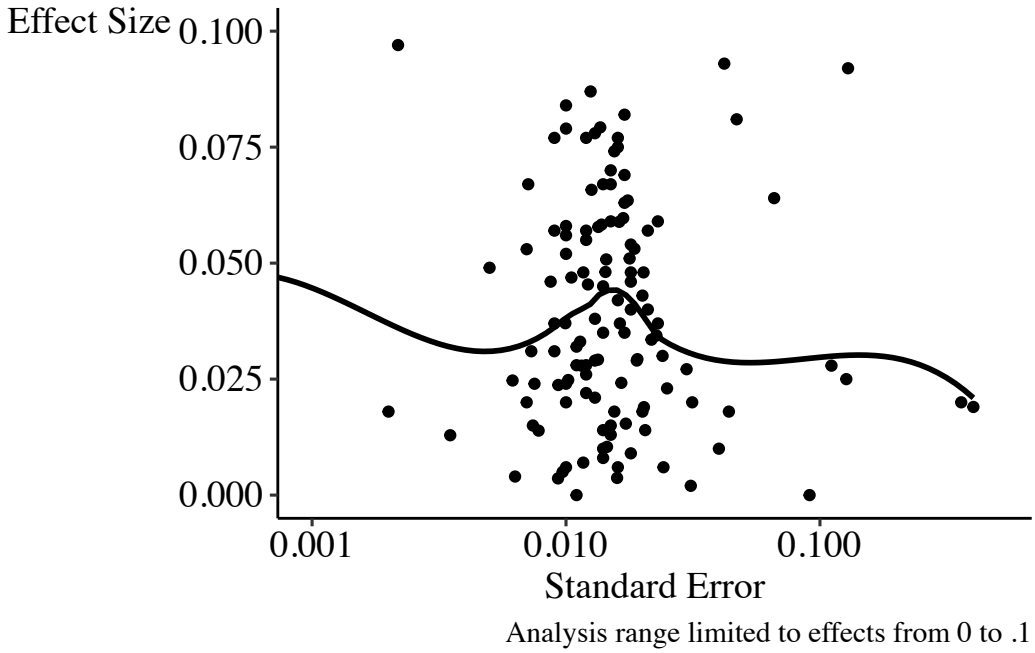


Figure C.3: Task 2 Effect Size and Standard Error

Appendix D: Peer Review

This section evaluates the impact of peer review on the later work performed by a researcher. The structure of peer review in this study is that, following each main task, 2/3 of the researchers are randomized into pairs that produce a peer review report of the other's work, while the remaining 1/3 do not receive or perform peer review. Then, researchers have an opportunity to revise their work.

Revision is optional, and relatively few researchers (fewer than 30 per task) chose to revise their work after receiving peer review. As such, we mostly look at the impact of peer review on the work performed in subsequent main tasks. The mechanisms by which peer review might be expected to change a researcher's work in normal journal submissions include both that researchers might find peer review comments helpful and incorporate them into their work, and that researchers are required by the journal submission process to incorporate most reviewer comments. In this study, our peer review process can only capture the first of these mechanisms, and in effect may be closer to comments received, for example, during seminar presentations.

In Appendix Table D.1, we incorporate revisions and show the variance of the entire sample of reported effects post-revision, replacing each researcher's reported task effect with its revision, if they revised their work. There is no statistically significant difference in variance between the reviewed and non-reviewed groups, nor is there a consistent effect in one direction. Similarly, as shown in Appendix Table D.2 there are no statistically significant differences in the variance of sample sizes between the peer-reviewed and non-peer-reviewed groups in the follow-up tasks.

Appendix Figure D.1 shows the distribution of effect sizes estimated by those who did, and did not, engage in peer review in each round. The left column of graphs shows the effects reported in each task before researchers were assigned to peer review, and the right column shows the

Table D.1: Post-Revision Variance in Effect Sizes by Peer Review

Task	Unreviewed Variance	Reviewed Variance	Levene Test p-value	Revised Variance
Task 1	0.002	0.012	0.173	0.001
Task 2	0.009	0.004	0.571	0.002
Task 3	0.001	0.008	0.210	0.015
Pooled	0.004	0.008	0.219	0.005

Table D.2: Post-Revision Variance in Sample Sizes by Peer Review

Task	Unreviewed Variance	Reviewed Variance	Levene Test p-value	Revised Variance
Overall Sample Size				
Task 1	3.396e+11	1.090e+13	0.239	2.369e+12
Task 2	2.179e+09	1.620e+12	0.479	1.649e+09
Pooled	1.886e+11	6.234e+12	0.163	1.074e+12
DACA Eligible Sample Size				
Task 1	7.785e+08	6.234e+11	0.450	1.722e+09
Task 2	9.849e+07	7.912e+10	0.383	2.761e+10
Pooled	6.738e+08	3.481e+11	0.315	1.537e+10
DACA Non-Eligible Sample Size				
Task 1	3.514e+11	6.044e+12	0.317	2.344e+12
Task 2	3.348e+12	8.872e+11	0.431	1.592e+09
Pooled	1.896e+12	3.495e+12	0.632	1.104e+12

effects reported in the follow-up task. As is expected given randomization, effect distributions are fairly similar pre-review between the review and non-review groups. No differences emerge between these groups in the follow-up task. Levene test p-values comparing effect size variance of peer-reviewed and non-peer-reviewed groups in follow-up tasks show p-values of 0.846 and 0.788 in Tasks 2 and 3, respectively, or 0.999 when pooling the two tasks. This is not strong evidence in favor of the idea that peer review might drive agreement between researchers due to the receipt of feedback. Similar results are found when comparing the distributions or variance of analytic, treatment, or control sample sizes between the peer-reviewed and non-peer-reviewed groups in the follow-up tasks.

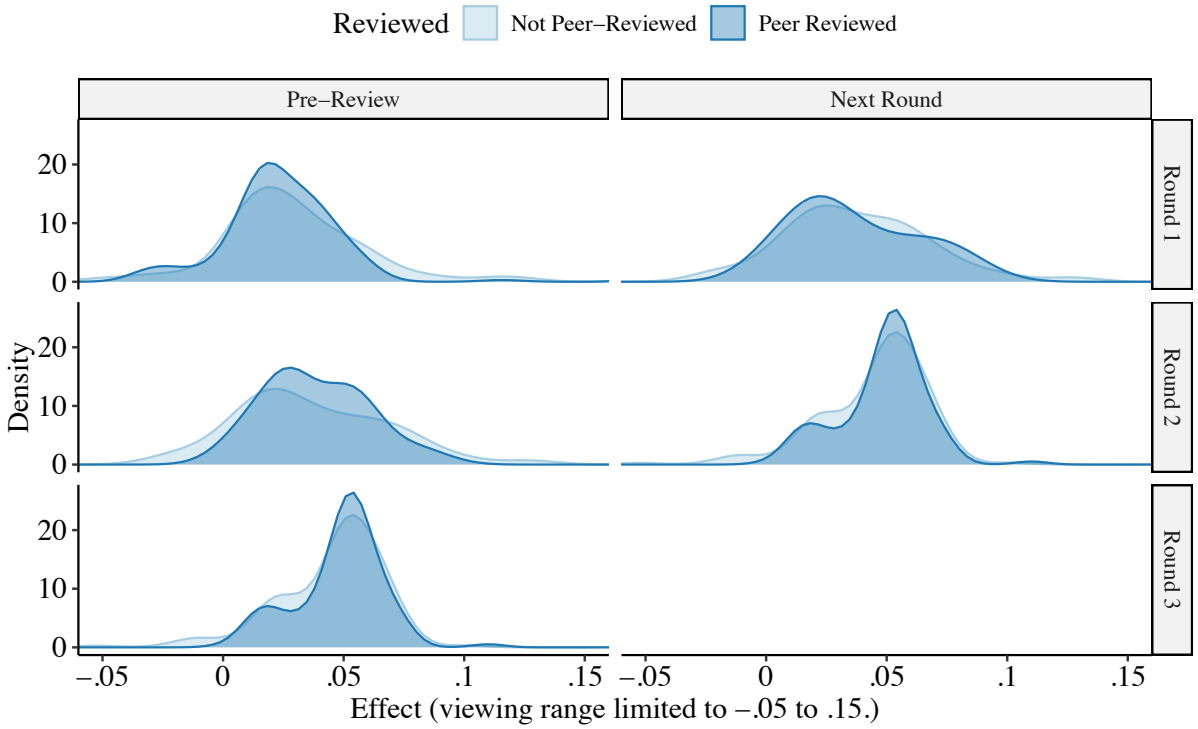


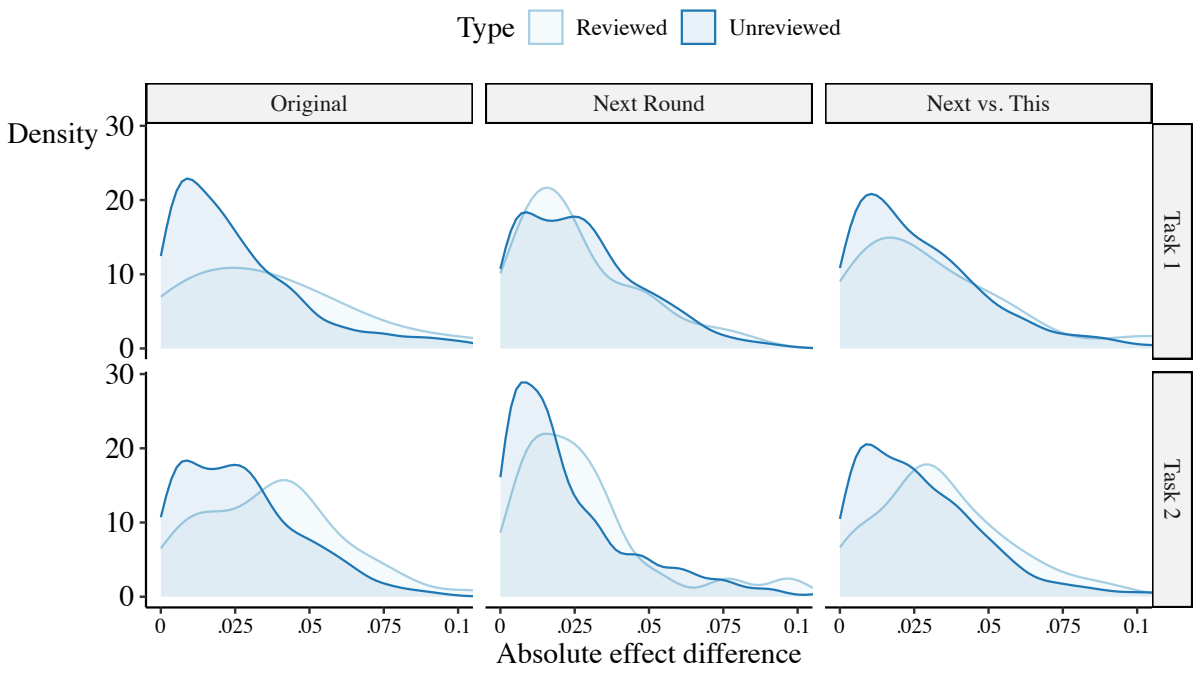
Figure D.1: Distributions of Reported Effect Sizes

Figure D.2 explores the possibility that peer review might not make the peer-reviewed group as a whole more similar, but rather just make someone more similar to their specific reviewer. We calculate the absolute difference in effects between each reviewer pair, in the task they

perform before reviewing (left column), in the follow-up task (middle column) and comparing your follow-up task against your reviewer's result this round (right column), with the right column representing the possibility that a researcher may select an analysis so as to produce a result more similar to the one they saw in the previous round.[25]

In Figure D.2 we see inconsistent evidence in favor of peer review. Task 1 review pairs became more similar in Task 2, while unreviewed pairs did not change. The change in average absolute effect differences from Task 1 to Task 2 was a statistically significant .051 greater for review pairs than non-review pairs (see Appendix Table D.3). However, this finding does not replicate in Task 2, where from Task 2 to Task 3, average absolute effect differences shrunk by a statistically significant .029 more for unreviewed than reviewed pairs. This is not consistent strong evidence of peer review making a researcher more like their reviewer as the result of feedback.

[25]The distributions of absolute differences for non-reviewed researchers are generated as a null distribution by matching every non-reviewed researcher to every other non-reviewed researcher and calculating all absolute differences. This null distribution represents the distribution of absolute differences among people who did not actually experience peer review. Notably, each non-reviewer is matched multiple times in this approach, instead of just once for reviewers. However, matching the non-reviewers only once to a single random pair just produces a noisier version of this all-matches null distribution. Averaging the single-random-match approach over many random single matches produces the same null distribution.



Original is this round vs. this round. Next round is next round vs. next round. Next vs. This is your next round vs. partner's this round. Values beyond .1 omitted for visibility. No weights applied.

Figure D.2: Comparisons of Effect Sizes vs. One's Reviewer

Table D.3: Paired Absolute Effect Differences and Peer Review

	Task 1	Task 2
Intercept	0.088***	0.065***
	(0.009)	(0.009)
Comparison: Next Round	−0.029**	−0.008
	(0.013)	(0.013)
Comparison: Next Round vs. This Round	−0.018	0.000
	(0.013)	(0.013)
Unreviewed	−0.048***	−0.003
	(0.009)	(0.009)
Next Round x Unreviewed	0.052***	−0.026**
	(0.013)	(0.013)
Next vs. This x Unreviewed	0.029**	−0.015
	(0.013)	(0.013)
Num.Obs.	7411	6970

Note: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Appendix E: Multi-Analyst Evaluation of Researcher Characteristics

E.1: Analysis by Project Organizer A

Table E.1 shows the F-statistic from a regression of the reported effect estimate on a set of indicators for that characteristic, as well as the associated p-value and R^2 from that regression. The indicators include each categorical researcher characteristic specified in Section 4.2, as well as an indicator for the use of R or Stata as a programming language. For all indicators, categories with 5 or fewer researchers in them were omitted before performing the analysis. This table allows us to see whether researchers with different characteristics reported different effect levels. Table E.2 does the same, but uses absolute deviation from the sample mean as the dependent variable, which allows us to see whether researchers with different characteristics showed greater agreement on effect levels with the group as a whole.

Tables E.1 and E.2 show that researcher characteristics hold basically no explanatory power for estimated effects either in level or deviation from the mean. Nearly all p-values are well above .05. In E.2, the p-value for race as an explanatory variable in Task 1 had a p-value below .1, but given how many comparisons there are in the table, this is likely to just be noise.

The only researcher characteristic that did seem to matter was the choice of programming language, which only weakly predicted effect level, but was a statistically significant predictor of being close to the mean effect in all three rounds.

Figure E.1 goes further into the split by language. We see that, of the two languages, Stata users were more likely to report effect estimates near the sample mean. 6.4%, 1.8%, and 0.9% of Stata users were more than .1 in absolute distance from the sample mean in Tasks 1, 2, and 3, respectively, while for R those values are 15.6%, 9.4%, and 12.5%. The number of R

Table E.1: Predicting Effect Level with Researcher Characteristics

	Task 1			Task 2			Task 3		
	F test			F test			F test		
	Stat.	p	R^2	Stat.	p	R^2	Stat.	p	R^2
Degree	0.929	0.337	0.007	0.122	0.727	0.001	0.085	0.771	0.001
Occupation	1.195	0.316	0.034	0.453	0.770	0.013	2.501	0.045	0.068
Research Experience	1.080	0.342	0.015	0.370	0.692	0.005	0.416	0.660	0.006
Gender	0.161	0.689	0.001	0.255	0.614	0.002	1.364	0.245	0.009
Race	1.026	0.383	0.022	1.306	0.275	0.028	0.342	0.795	0.007
LGBTQ+	0.426	0.654	0.006	0.183	0.833	0.003	0.045	0.956	0.001
Recruitment Source	0.360	0.698	0.005	1.661	0.194	0.024	1.400	0.250	0.020
Field	1.406	0.238	0.011	4.562	0.035	0.034	0.831	0.364	0.006
Coding Language	3.861	0.051	0.027	3.117	0.080	0.022	0.653	0.420	0.005

Note: Each line shows the results for a separate regression by task number. The dependent variable is the reported effect estimate and the independent variables are indicators capturing the researcher characteristics listed in the first column. The F-statistic and associated p-value are for a null hypothesis of no differences in effect size across indicators for the particular researcher characteristics. In addition, the R^2 value is reported for each regression.

Table E.2: Predicting Effect Deviation with Researcher Characteristics

	Task 1			Task 2			Task 3		
	F test			F test			F test		
	Stat.	p	R^2	Stat.	p	R^2	Stat.	p	R^2
Degree	1.915	0.169	0.013	0.740	0.391	0.005	0.630	0.429	0.004
Occupation	0.890	0.472	0.025	0.535	0.710	0.015	1.845	0.124	0.051
Research Experience	1.364	0.259	0.019	0.284	0.754	0.004	0.741	0.478	0.011
Gender	1.576	0.211	0.011	1.102	0.296	0.008	0.144	0.705	0.001
Race	2.180	0.093	0.045	0.129	0.943	0.003	0.762	0.517	0.016
LGBTQ+	0.202	0.817	0.003	0.515	0.599	0.007	0.253	0.776	0.004
Recruitment Source	2.064	0.131	0.030	0.197	0.822	0.003	0.552	0.577	0.008
Field	0.077	0.781	0.001	0.936	0.335	0.007	0.072	0.789	0.001
Coding Language	4.369	0.038	0.030	4.537	0.035	0.032	5.022	0.027	0.035

Note: Each line shows the results for a separate regression by task number. The dependent variable is the absolute deviation from the sample mean of the reported effect estimate and the independent variables are indicators capturing the researcher characteristics listed in the first column. The F-statistic and associated p-value are for a null hypothesis of no differences in deviation from the sample mean across indicators for the particular researcher characteristics. In addition, the R^2 value is reported for each regression.

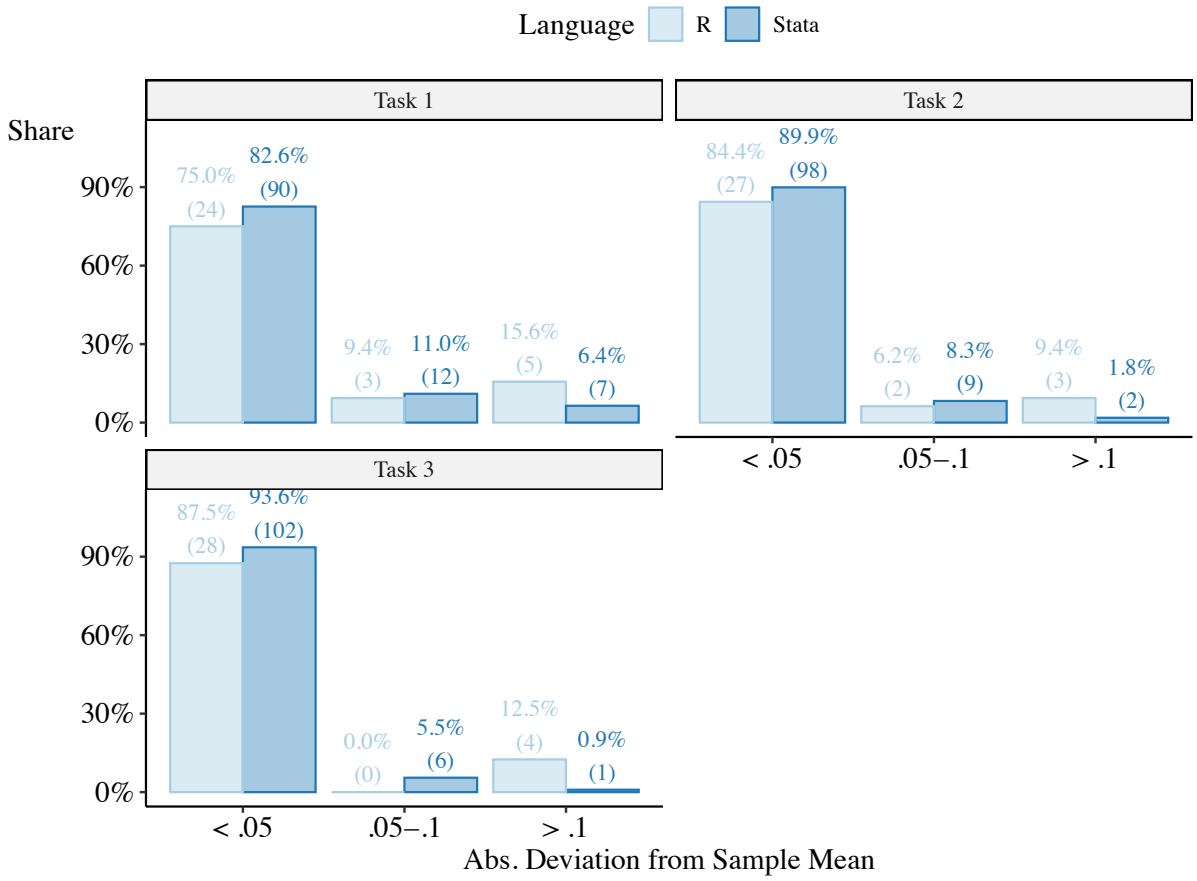


Figure E.1: Deviation from Sample Mean of Reported Effect by Language

users is relatively low at 32,[26] and so these numbers are sensitive to any researchers who were consistently outliers. There were two R users who had an absolute deviation from the mean of .1 or more every round, while all other R researchers with deviations of .1 or more only had deviations that large in a single round. If we omit those two consistently-high-deviation R users, the percentages are 10%, 3.3%, and 6.7% for R users, which are still higher than the percentages for Stata users.

Overall, there is little role for researcher professional or demographic characteristics in predicting either the level of the effects they reported, or the deviation of those effects from the mean. There is some explanatory power for the choice of programming language. R users were more likely than Stata users to report estimates far from average of what other users reported.

E.2: Analysis By Project Organizer B

Table E.3 looks at within-researcher variation in effect estimates across tasks. In the first three columns, the dependent variable is the absolute difference in effects for a given researcher across two tasks, while in the fourth column, the dependent variable is a researcher's maximum estimated effect minus their minimum.

Most researcher characteristics do not predict absolute within-researcher variation. Career stage, occupation, and number of published papers do not predict absolute differences in estimates across tasks to a statistically significant degree, with few exceptions.

One exception is that private researchers saw larger absolute changes between Task 1 and Task 3, and also more absolute variation overall, although the latter is only significant at the $\alpha = .1$ level. Probably the most interesting is that inexperience was related to smaller changes from Task 1 to Task 3: those who do not have a PhD showed a smaller change between Task 1

[26] This is one lower than the value reported in Section 4.2 because the researcher who was dropped from analysis, mentioned later in Section 4.2, was an R user.

Table E.3: Model Coefficients and Standard Errors for Task Comparisons

	Task 1 vs Task 2	Task 2 vs Task 3	Task 1 vs Task 3	Absolute Range
Intercept	0.053***	0.068***	0.053***	0.087***
	(0.015)	(0.018)	(0.017)	(0.022)
Grad student	0.090*	−0.015	0.099*	0.086
	(0.047)	(0.054)	(0.051)	(0.066)
Uni researcher	−0.001	−0.014	0.016	0.000
	(0.033)	(0.038)	(0.036)	(0.047)
Other	0.004	−0.013	0.032	0.011
	(0.070)	(0.081)	(0.077)	(0.099)
Private researcher	0.021	0.065	0.134***	0.110*
	(0.042)	(0.049)	(0.046)	(0.059)
Public researcher	0.047	−0.013	0.044	0.039
	(0.035)	(0.041)	(0.039)	(0.050)
Not PhD	−0.070*	0.003	−0.081*	−0.074
	(0.040)	(0.047)	(0.044)	(0.057)
1-5 papers	−0.007	−0.037	−0.011	−0.027
	(0.021)	(0.025)	(0.024)	(0.030)
0 papers	0.012	−0.047	−0.017	−0.026
	(0.032)	(0.037)	(0.035)	(0.045)
Num.Obs.	145	145	145	145
R2	0.046	0.040	0.088	0.047
R2 Adj.	−0.010	−0.017	0.034	−0.009

Note: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

and Task 3 (significant at $\alpha = .1$), and those with fewer papers also showed smaller absolute changes than those with 6+ papers (insignificant).