The Sources of Researcher Variation in Economics*

Nick Huntington-Klein Claus C. Pörtner

The Many-Economists Collaborative on Researcher Variation

*Corresponding author Nick Huntington-Klein, nhuntington-klein@seattleu.edu, +1 (206) 296-5815. Department of Economics, Seattle University, 901 12th Ave., Seattle, WA, 98122. 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. First and foremost, we are profoundly grateful to the researchers who participated in this project. Those contributors who wished to be recognized are listed in Appendix A as members of the "Many-Economists Collaborative on Researcher Variation"; some contributors chose not to be listed. Furthermore, 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, Ludwig-Maximilians-Universität München (LMU Munich), Western Washington University, and the 2024 Annual Meeting of WEAI for helpful comments and suggestions. Finally, the comments and suggestions from three anonymous reviewers and Editors David Romer and Erzo Luttmer have helped substantially improve the paper. Data and code for this project are available at https://github.com/many-economists/analysis. Preregistration for the project is available in Appendix C and at OSF: https://doi.org/10.17605/OSF.IO/CJ9YX.

Abstract

Disagreement among researchers is a central and productive feature of scientific progress. Yet researchers' incentives and behaviors, some only partly visible to readers, may affect empirical results. When these sources of variation affect estimates, standard errors may understate uncertainty. This paper synthesizes evidence on how researcher variation shapes empirical conclusions and presents new evidence on its sources and implications. We report results from a many-analyst study in which 146 research teams estimate the same causal effect under varying constraints. Although central estimates are broadly aligned, dispersion remains substantial. Imposing a shared research design reduced agreement, partly due to adherence errors. Further restricting data cleaning and preparation improved agreement. Variation across researchers is often comparable in magnitude to sampling uncertainty but is unevenly distributed across stages of the research process. Overall, the evidence suggests that conventional uncertainty summaries understate total uncertainty and that less visible research decisions warrant greater scrutiny.

1 Introduction

When economists read empirical results, they typically treat the reported standard errors as summarizing the relevant uncertainty around the estimated effects. However, this interpretation does not account for variation arising from researchers' incentives and behaviors, and a growing literature shows that these factors shape both the results themselves and which results enter the public domain (Stevenson and Fischman, Forthcoming). Issues such as p-hacking, selective reporting, intentional specification searching, and the inability to reproduce published results have attracted substantial attention (Simmons, Nelson and Simonsohn, 2011; Open Science Collaboration, 2015; Brodeur, Cook and Heyes, 2020). Yet the relevant mechanisms extend far beyond bad-faith conduct. Limited time and attention, publication incentives, differing ideas about best practices, and beliefs about which results are interesting or publishable can lead researchers to write up, emphasize, or continue pursuing some findings rather than others (Franco, Malhotra and Simonovits, 2014; Andrews and Kasy, 2019; Heckman and Moktan, 2020; Frankel and Kasy, 2022). These mechanisms make it important to understand not only whether results vary across researchers, but also where that variation arises and how visible it is to readers.

Here we focus on "researcher variation," defined as the variation in results that would arise from differences in decisions and behaviors among researchers conducting a common research task. This concept is closely related to "implementation variation" suggested in Stevenson and Fischman (Forthcoming). We take as our starting point the growing literature across economics and other fields showing that equally competent researchers can reach meaningfully different conclusions, even when addressing the same research question using the same underlying data (Silberzahn et al., 2018; Huntington-Klein et al., 2021; Menkveld et al., 2024).

Empirical research involves a series of choices, and disagreement among researchers is a natural

and productive feature, reflecting judgment calls and implementation decisions at multiple stages of the research process. However, not all research decisions are treated equally in how they are observed, evaluated, or debated within the profession.

Some choices—such as the identification strategy, estimator, or calculation of standard errors—are routinely scrutinized by referees and readers and are generally stated explicitly in published work, but other decisions are far less visible. Decisions about data cleaning, variable construction, sample restrictions, and handling missing or inconsistent observations are often described only briefly, if at all. As a result, variation arising from these less visible stages of the research process may be harder for readers to assess, even when those decisions meaningfully affect reported results. This raises several questions central to economics: how large is researcher variation, which stages of the research process contribute most to it, and how institutional features, such as research design constraints and peer review, shape the dispersion of results.

If researcher variation is large and concentrated in parts of the research process that are hard for readers to observe, standard modes of interpretation may understate the true uncertainty of the results. The key issue is not whether researchers disagree, but rather what kinds of disagreement arise, where in the research process they arise, and how visible they are to those evaluating empirical evidence. Variation in how research tasks are implemented changes not only the dispersion of possible estimates but also can shape the published record simply by expanding the set of estimates over which significance bias, file-drawer dynamics, sign selection, and ideological priors operate (Brodeur et al., 2016; Jelveh, Kogut and Naidu, 2024).

This paper first reviews the growing literature on researcher variation, which has sought to document and characterize variation arising from researcher decisions and behaviors. To organize this evidence and clarify what is known and unknown, we structure the review around a stylized representation of the empirical research process, distinguishing among choices regarding the research question, data and analytic sample construction, identification strategy, estimation

and model specification, coding and implementation, and inference choices and reporting. This framework enables us to synthesize evidence from multiverse and many-analyst studies, assess the magnitude of researcher variation, and evaluate which features of researchers, research environments, and research policies systematically shape empirical outcomes.

Second, we contribute new evidence from a large-scale, many-analyst study conducted in an applied economic setting. A large number of research teams independently addressed the same causal inference question using the same underlying data, first with substantial discretion and then under progressively tighter constraints. By structuring the study into multiple rounds, we can separate choices related to research design, data cleaning and preparation, and modeling decisions. In addition, randomized peer feedback is incorporated between rounds. Researchers were not required to satisfy peer feedback, which makes it unlike typical journal peer review, but still allows us to examine whether feedback disciplines subsequent choices. This design enables us to move beyond documenting researcher variation to examining where it arises and how it responds to institutional features common in applied economic research, at least in the context of one research question and policy effect estimation.

Our results indicate that researcher variation persists even in a familiar applied setting and that its sources are unevenly distributed across stages of the research process. In the first round, where researchers had substantial discretion, the middle half of reported estimates ranged from 1.4 to 5.1 percentage points, yielding an interquartile range of 3.7 percentage points. Imposing a shared research design in the second round did not increase agreement, partly because some researchers did not fully adhere to the specified design. Agreement improved only in round three, when researchers used a common pre-cleaned dataset; the interquartile range then fell to 2.7 percentage points.

Some disagreement among researchers is both unavoidable and benign, reflecting reasonable differences in judgment among competent analysts. At the same time, less visible parts of the

research process—particularly data preparation—can substantially shape empirical results, even when research design and modeling choices are relatively familiar and well understood. This does not imply that such variation can or should be eliminated, nor that conventional practices fail in general. Rather, the key task is to understand which sources of variation are most consequential and how visible those sources are to readers evaluating empirical evidence.

2 What Do We Know about Researcher Variation?

Economics has long recognized the potential for researcher variation (Cowles, 1933; Leamer, 1982), but systematic efforts to measure it have emerged only recently—primarily in response to broader concerns about empirical credibility outside economics (Simmons, Nelson and Simonsohn, 2011; Gelman and Loken, 2014; Open Science Collaboration, 2015; Nosek et al., 2022). In this section, we focus on how variation in empirical results can arise from discretionary research decisions, examining the magnitude of this variation, its causes, and whether it reflects substantive disagreement, arbitrary discretion, or error (Menkveld et al., 2024). We first describe the main empirical approaches used to study the degree of researcher variation and the estimated magnitudes. We then organize the literature around the stages of the empirical research process to clarify where variation arises, before turning to evidence on why it arises. Finally, we highlight the open questions we consider important.

The two central empirical challenges in studying researcher variation are how to observe it and how to measure its magnitude. Researcher variation is defined relative to alternative choices and therefore cannot be observed from a single published estimate. The two main designs used to address this problem are multiverse and many-analyst designs.

The multiverse design aims to answer a specific research question by conducting all reasonable analyses across the full set of data constructions and specification combinations, rather than relying on a small set of "preferred" analyses (Sala-i Martin, 1997; Steegen et al., 2016; Auspurg, 2025). This design is especially useful for assessing the sensitivity of conclusions to data, specification, and estimation choices and for answering the question of what researchers could plausibly have done differently (Heyman and Vanpaemel, 2022). However, it cannot tell us much about the effects of researcher characteristics, institutional incentives, or decisions outside the predefined choice set.

Multiverse designs have primarily been used to critique published studies, though they may eventually serve as an inferential tool or as a means to provide richer robustness evidence (Rohrer, Hullman and Gelman, 2026). What remains unclear is how to determine which choices are arbitrary and should be included in the choice set, and which choices are necessary to answer a given research question and should not be varied.[1] Compared with other social sciences, economics has the advantage of a stronger tradition of theoretical modeling, which provides limits, for example, on which variables can be included in an analysis and still yield a causal result.

In many-analyst designs, multiple independent researchers complete the same research task, allowing organizers to observe researcher behavior when analysts face a common empirical problem (Huntington-Klein et al., 2021; Menkveld et al., 2024). This design is especially useful for examining which analytic paths researchers choose and how those choices relate to researcher characteristics or features of the research environment. Many-analyst designs also more closely resemble the standard research process than multiverse designs because participants make connected decisions across the empirical workflow, rather than evaluating a set of specification alternatives. They have primarily been used to document the extent of

[1] In addition, it is possible that the number of analyses can create a false sense of rigor or certainty, that readers may mistakenly treat the frequency of results across specifications as probabilistic evidence, or that poorly justified specifications can exaggerate uncertainty.

variation across researchers, but they can also be used to study how institutional features such as peer review, task restrictions, or data standardization affect agreement.

What remains unclear is the external validity of these designs. The assigned task, incentives, review process, and time horizon necessarily differ from those faced by researchers pursuing their own projects for publication (Auspurg and Brüderl, 2024*b*). Moreover, ambiguity in the research question can be interpreted in two ways: as a design limitation if the goal is to isolate implementation variation, or as a substantive finding if the goal is to study how researchers translate a verbal question into empirical analyses. Hence, whether many-analyst designs understate or overstate researcher variation remains unresolved.

Even when research designs make researcher variation observable, its magnitude is difficult to summarize because there is no natural scale for measuring it (Coqueret and Pérignon, 2026). Existing studies therefore use several complementary benchmarks. One approach focuses on the dispersion of estimates across researchers or specifications, summarized by measures such as standard deviations or interquartile ranges (Menkveld et al., 2024, and our analyses below). A second asks whether researcher choices lead to qualitatively different conclusions, such as sign reversals or changes in statistical significance (Walter, Weber and Weiss, 2024; Weill et al., 2025). A third compares researcher variation with sampling uncertainty, often by relating the dispersion of estimates to reported standard errors (Huntington-Klein et al., 2021; Menkveld et al., 2024; Holzmeister et al., 2024). This comparison is intuitive because standard errors are the conventional measure of uncertainty in empirical work, but it is imperfect because sampling uncertainty depends on sample size, research design, and outcome variability, whereas researcher variation need not scale with those same factors. Hence, it remains an open question how to measure and interpret the magnitude of researcher variation.

With these caveats in mind, evidence from economics and finance suggests that researcher variation is often comparable to sampling uncertainty and, in some cases, substantially larger

(Magnus and Morgan, 1997; Huntington-Klein et al., 2021; Weill et al., 2025; Menkveld et al., 2024; Soebhag, Van Vliet and Verwijmeren, 2024). Direct comparisons of researcher variation across fields are inherently imprecise because studies differ in research tasks, estimands, sample sizes, and metrics. However, a synthesis of many-analyst studies finds that researcher variation in economics is relatively small compared with some other disciplines (Holzmeister et al., 2024), partly because other fields, such as sociology, exhibit extremely high dispersion among analysts (Silberzahn et al., 2018).[2]

2.1 A Research-Process Framework for Studying Researcher Variation

To understand where researcher variation arises, we introduce a stylized framework that divides the applied microeconomic research process into six stages, each involving discretionary choices and thus creating scope for variation among researchers. We adopt this framework as an organizing device rather than a rigid taxonomy to provide a common language for comparing findings across studies, clarify which decisions vary, and help distinguish intentional methodological disagreement from arbitrary, weakly reported, or erroneous choices. The literature on researcher variation does not always clearly distinguish among these stages, and individual studies often allow some stages to vary while holding others fixed.

First, researchers choose a *research question*, specifying the population of interest and the causal or descriptive parameters to be estimated. Second, researchers make decisions about *data and analytic sample construction*, including dataset selection, variable construction, sample restrictions, and handling missing data and outliers. Third, researchers select an *identification strategy*, which specifies the sources of identifying variation, the choice of designs (such as RCT, RDD, DiD, synthetic control, structural model, and so on), and the counterfactual

[2]For evidence on researcher variation in other fields, see, for example: religion (Hoogeveen et al., 2023), neuroimaging (Botvinik-Nezer et al., 2020), political science (Breznau et al., 2022), machine learning (Chen and Cummings, 2024), ecology and evolutionary biology (Gould et al., 2025), psychology (Boehm et al., 2018; Bastiaansen et al., 2020; Schweinsberg et al., 2021), and medical informatics (Ostropolets et al., 2023).

comparison that gives estimates a causal interpretation, if the research question is causal in nature. Fourth, researchers make *estimation and model specification choices*, including the choice of estimator (e.g., OLS, logit, Poisson, GMM), functional form, covariate inclusion, interactions, and variable transformations. Fifth, *coding and implementation* includes choices of software, numerical solvers, optimization routines, and default settings, as well as the potential for coding errors. Finally, researchers choose *inference and reporting conventions*, such as standard error adjustments, clustering, multiple-testing corrections, winsorization, and the presentation of results.

Clearly, some researcher choices and behaviors blur the boundary between these stages. For example, fixed effects can be viewed as model-specification choices when added to an existing identifying comparison, but as identification-strategy choices when they determine the comparison from which the causal effect is identified. Similarly, alternative difference-in-differences estimators may be viewed as specification choices when they provide different implementations of a common design and estimand, but they become identification or estimand choices when they alter the comparison groups, weighting structure, treatment-effect aggregation, or the assumptions under which the estimate is interpreted.

Research Question

Little is known about variation arising from choices of research question, since empirical studies of researcher variation necessarily treat the question as fixed. Yet a large sociology-of-science literature shows that incentives, norms, and evaluation systems shape which topics researchers pursue (e.g., Merton, 1973; Zuckerman, 1988; Lamont, 2009; Azoulay, Graff Zivin and Manso, 2011).[3] These upstream choices may affect researcher variation by altering both the feasible set of designs and the incentives researchers face.

[3]There may also be differences in which fields different researchers sort into, although it is unclear if this sorting affects researcher variation (Singhal and Sierminska, 2025).

Data and Analytic Sample Construction

After deciding on a topic, selecting data sources and constructing the analytic sample are often the next steps in the empirical research process. These decisions include how raw data are transformed into analysis-ready variables, which observations are included or excluded, and how missing values, outliers, and measurement ambiguities are addressed.

Data choices are ubiquitous in applied work, yet they are often weakly standardized and only partially documented. We still know little about the role this part of the research process plays in generating variation in results.[4] Many-analyst studies across economics and other fields have typically held data construction fixed by providing researchers with a common, analysis-ready dataset. More broadly, the scarcity of systematic evidence reflects both practical and conceptual challenges. Data preparation decisions are difficult to observe, hard to standardize, and often treated as ancillary to "analysis" proper.

There is some indirect evidence that data-related decisions can be a substantial source of researcher variation. Huntington-Klein et al. (2021) decomposes variation into components arising from downstream modeling choices and from data-preparation decisions embedded in the analytical workflow, and finds that the latter accounts for a substantial share of total variation. Similarly, differences in winsorization produced economically meaningful variation in estimated effects despite broad agreement on the underlying empirical framework in a many-analyst finance study (Menkveld et al., 2024). However, it remains unclear how much additional variation would arise if many-analyst studies did not condition on a fixed dataset. Below, we therefore extend the existing literature by explicitly allowing researchers to make their own sample-construction choices.

[4] Evidence from large-scale replication efforts indicates that the availability of pre-cleaned data does not necessarily increase replication success (Brodeur et al., 2024).

Identification Strategy

Evidence of researcher variation stemming from identification strategy choices predates the recent focus on researcher degrees of freedom, and it shows that even with a common data source and research objective, variation can be substantial. A prominent example is a coordinated set of studies published in the *Journal of Applied Econometrics* that estimated food demand and elasticities using the same data from the U.S. and the Netherlands (Magnus and Morgan, 1997). Researchers adopted different time-series modeling approaches, leading to substantial variation in the estimates.

Recent evidence from multiverse analyses points in a similar direction (Weill et al., 2025). For example, using heterogeneity-robust difference-in-differences estimators instead of standard two-way fixed-effects estimators changes the effective comparison groups and reduces, but does not eliminate, instability in point estimates of the effects of COVID-19 mobility-reducing policies across specifications (see also the discussion below). Similarly, many-analyst designs show that identification-level choices can generate policy-relevant variation (Huntington-Klein et al., 2021; Menkveld et al., 2024).[5]

Estimation and Model Specification

Early econometric work emphasized that model specification uncertainty could stem from both unavoidable researcher discretion and more problematic practices, such as data-driven model selection. Concerns about the sensitivity of results to specification choices motivated a large literature on data snooping and model uncertainty, highlighting that alternative specifications applied to the same data could yield materially different conclusions (Leamer, 2017; White,

[5]These examples are separate from cherry-picking or direct manipulations of results from using different identification strategies (Brodeur, Cook and Heyes, 2020; Stevenson and Fischman, Forthcoming)

2000; Lo and MacKinlay, 1990; Pötscher, 1991). A prominent illustration of this issue appears in growth regressions, where the choice of control variables varies widely across studies. Systematic exploration of alternative predictor combinations revealed that estimated relationships were highly sensitive to specification choices, underscoring the potential for researcher variation at this stage (Sala-i Martin, 1997).

When published studies examining the effects of social distancing policies during the early COVID-19 pandemic were reanalyzed under alternative modeling choices, relatively small differences—such as the choice of variable transformations—were sufficient to change the policy-relevant conclusions that could be drawn (Weill et al., 2025). In the same vein, across 110 papers in leading economic and political science journals, 74% of originally significant results retained their significance when the estimation method was changed as a robustness check during replication, for example, from linear probability model (LPM) to Probit (Brodeur et al., 2024).

Evidence from finance further illustrates the magnitude of specification-driven variation. Hypothetical multiverse analyses of factor models show that routine portfolio-construction decisions—such as breakpoint selection or rebalancing frequency—can produce large differences in estimated Sharpe ratios, often exceeding typical measures of sampling variation (Soebhag, Van Vliet and Verwijmeren, 2024). Related findings from portfolio-sorting exercises show that researcher variation in specification is comparable in magnitude to sampling variation, even when qualitative conclusions, such as the sign of estimated effects, remain relatively stable (Walter, Weber and Weiss, 2024). Analyses of financial anomalies likewise indicate that, for many persistent anomalies, variation induced by alternative analytic specifications exceeds sampling uncertainty when assessed with variance-decomposition approaches (Coqueret and Pérignon, 2026).

Coding and Implementation

Evidence from methodological audits underscores how sensitive results are to implementation choices. Comparisons of nonlinear solver performance across programming languages and software environments show that ostensibly equivalent estimation problems can yield meaningfully different numerical solutions, depending on algorithmic defaults and convergence criteria (McCullough and Vinod, 2003). By convention, economists generally do not report which programming language or estimation software they use, and so these sorts of differences are rarely visible in published work and cannot be scrutinized by a reader.

More recent work reinforces the relevance of implementation-related variation in applied research, as large-scale replication efforts document that coding errors are common in published economics and finance research, indicating that computational execution itself is a nontrivial source of researcher variation (Brodeur et al., 2024).[6] Taken together, this evidence suggests that computational implementation constitutes a distinct and often opaque source of researcher variation. While its effects are frequently observed in differences in reported results and inference, the underlying source lies in technical choices that are difficult to standardize, weakly documented, and rarely scrutinized.[7]

Inference Choices and Reporting

The final source of researcher variation stems from inference choices and reporting, where results are summarized and communicated. One set of mechanisms operates through reporting conventions and journal policies. For example, changes to norms governing how statistical

[6] For example, in recent work claiming that ideology affects empirical findings, a categorization error led to single-team categories that did not contribute to the estimate of the effect of ideology; fixing this error caused the ideology effect to disappear (Auspurg and Brüderl, 2026).

[7] Experimental evidence also suggests that responses to detected errors vary systematically across researchers: when errors occur, some researchers correct them while others do not, and the likelihood of correction depends in part on whether the error produces an unexpected or unfavorable estimate (Ferman and Finamor, 2025).

significance is communicated alter the informational environment in which results are evaluated. However, evidence from policy changes in economics journals indicates that modifying reporting practices, such as removing statistical significance stars from regression tables, does not necessarily reduce practices like p-hacking or publication bias, although it does change how results are framed and interpreted (Naguib, 2026). Related work on replication requirements suggests that policies affecting transparency and reporting may influence the visibility and dissemination of research, even when their effects on underlying analytic behavior are difficult to isolate (Höffler, 2017).

A second, closely related channel through which researcher variation arises is inaccuracies in reported inference. Structured evidence indicates that errors in reported statistics are not uncommon in published economics and finance research. Large-scale replication efforts document discrepancies between reported and recomputed results, while systematic examinations of published outputs reveal frequent, though often minor, inconsistencies in reported significance levels and summary statistics (Brodeur et al., 2024; Pütz and Bruns, 2020). Because such inaccuracies directly shape reported uncertainty and the interpretation of empirical findings, they constitute an additional source of researcher variation.

2.2 Why Does Researcher Variation Arise?

So far, we have focused on where researcher variation occurs in the research process. We now turn to three potential explanations for why researcher variation arises: researcher characteristics, political orientation, and features of the research environment.

Observable researcher characteristics appear to explain relatively little variation. Few studies examine this directly because doing so requires many researchers performing comparable tasks, but those that do generally find small effects of experience and other characteristics (Menkveld

et al., 2024). Evidence from large-scale replication efforts similarly finds little systematic relationship between author experience and replicability (Brodeur et al., 2024; Herbert et al., 2024), although the importance of researcher characteristics may depend on the type of task being evaluated (Song, Wu and Zhou, 2025).

Political orientation is a natural candidate for explaining researcher variation when empirical questions are politically salient. However, in a many-analyst study of immigration and support for social policies, researchers' political and policy attitudes explain little of the variation in estimated effects, as do their expectations, statistical experience, and other characteristics (Breznau et al., 2022).[8] Evidence from published work similarly suggests that ideological differences are detectable but modest in magnitude (Jelveh, Kogut and Naidu, 2024).

By contrast, features of the research environment appear more consequential. Peer review can reduce researcher variation by disciplining modeling, specification, and reporting choices (Menkveld et al., 2024).[9] Transparency policies and reproducible-workflow requirements may also matter, although their effects on underlying analytic behavior are difficult to isolate; journal requirements for replication packages, for example, may increase the visibility and dissemination of research without necessarily changing how analyses are conducted (Höffler, 2017).

More generally, researcher variation appears closely linked to the discretion afforded by the task or research environment. When permissible choices are narrowed by field-wide standards, task design, or explicit analytic restrictions, variation tends to decline. Limiting portfolio-construction decisions in financial analyses, for example, substantially reduces dispersion in

[8] A subsequent re-analysis of the same experiment argued that researchers' political attitudes toward immigration policy are strongly correlated with whether estimated effects are positive or negative, and that more extreme views, on either side, are associated with lower-quality analyses. However, this result is not stable after correcting a categorization error (Borjas and Breznau, 2024; Auspurg and Brüderl, 2026).

[9] Related evidence from an experimental study of policy professionals shows that enforced discussion among decision-makers reduces variation in policy judgments, even when underlying political preferences differ (Banuri, Dercon and Gauri, 2019).

estimated outcomes (Soebhag, Van Vliet and Verwijmeren, 2024, and see our analyses below). Similarly, evidence from economics and computer science indicates that more complex tasks, which allow more defensible analytic paths, are associated with greater researcher variation (Menkveld et al., 2024; Ortloff et al., 2023).

2.3 Open Questions

The evidence reviewed above shows that researcher variation is substantial, occurs across the different stages of the research process, and appears to be shaped more by research environments and institutional constraints than by observable researcher characteristics. Several important questions nevertheless remain unresolved.

First, there is no consensus on how to measure the magnitude or importance of researcher variation. Existing work uses dispersion measures, sign reversals, changes in statistical significance, and comparisons with sampling uncertainty, but each benchmark has limitations. Relatedly, the external validity of multiverse and many-analyst designs remains uncertain: do many-analyst studies understate or overstate variation in actual research, how should multiverse choice sets be defined, and should multiverse analyses be treated as inferential tools or primarily as critiques?

Second, we know relatively little about variation in data construction and analytic sample decisions. Most many-analyst studies provide researchers with cleaned data, leaving open how much variation arises earlier in the workflow. More generally, we are still in the early stages of understanding how much variation is driven by routine model choices and how that variation differs across identification strategies and model specifications, despite the profession's long-standing focus on these decisions.

Third, existing evidence offers limited guidance on how research norms and conventions evolve. Peer review, workflow standards, and reporting norms appear to shape researcher variation, but little is known about how these institutions emerge, diffuse, and interact, or when they constrain discretion without discouraging innovation. A long tradition in the history and sociology of science examines the evolution of scientific norms and standards of evidence (e.g., Kuhn, 1962; Shapin, 1994), but this work is largely qualitative and does not provide quantitative evidence on how norm evolution affects researcher variation in empirical work.

Finally, it remains unclear which forms of researcher variation are inherent to empirical research and which can be meaningfully shaped by institutional or policy interventions. Some variation persists even under tight constraints, reflecting unavoidable judgment calls, context-specific trade-offs, and imperfect execution. The central question is therefore which sources of variation are irreducible, which reflect valuable methodological disagreement, and which are most amenable to intervention without discouraging productive research practices.

These open questions motivate the analysis that follows. By extending the many-analyst design to allow researchers to make their own data construction and sample-definition choices, we provide evidence on the role of upstream discretion, how variation propagates across stages of the research process, and the extent to which common research environments discipline—or fail to discipline—researcher variation.

3 Empirical Strategy: Isolating Sources of Researcher Variation

To address the gaps identified above, we implement a many-analyst design in which the same set of researchers completes the same research task multiple times under systematically varied constraints. As discussed above, many-analyst designs involve trade-offs between experimental control and realism, and estimates of researcher variation obtained in such settings may not

fully generalize to typical research environments. The analysis that follows should therefore be understood as evidence on specific margins of researcher discretion under controlled conditions, rather than as a definitive measure of researcher variation across all empirical contexts.

In each round, researchers are asked to estimate the effect of the Deferred Action for Childhood Arrivals (DACA) program on the probability that affected individuals work full-time. The task is repeated three times—referred to as Task 1, Task 2, and Task 3—with variations in the information provided and the restrictions placed on researchers' choices. The intuition behind this design is straightforward: if restricting a particular margin of researcher discretion meaningfully reduces variation in results across researchers, then that margin is an important source of researcher variation.

After each task, a subset of researchers is randomly assigned to peer-feedback pairs, and all researchers are given the opportunity to revise their analyses. This design feature allows us to assess the extent to which deliberation and external scrutiny mitigate researcher variation, complementing the evidence reviewed above on the role of research environments and institutional constraints.

The following goals and instructions apply to all tasks (see Appendix B for more details):

- Estimate the causal effect of the DACA policy on the probability of working full-time among the group affected by the policy.
- Use American Community Survey (ACS) data to estimate the effect, using data from no older than 2006 and no newer than 2016.
- Obtain 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 statistical 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, and instructions to use assistants for any work they would normally use assistants for. They are told 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 from differences in research design choices. Instead of allowing any research design to identify the causal effect of interest, Task 2 provided specific definitions for which individuals comprised a "treated" group and which comprised an "untreated" comparison group.[10] Researchers were then instructed to estimate the effect by comparing how outcomes for the "treated" group changed from before DACA was implemented to after, against how outcomes 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

[10]Although eligibility criteria for DACA were explicitly stated in Task 1, Task 2 further narrows the treated group by limiting the acceptable age range. The Task 2 limitation was more significant for defining the untreated comparison group. Many researchers used a treated/untreated group approach in Task 1 before Task 2 specified it, but they defined the untreated group in highly diverse ways, as will be shown in the Results section.

instructions.

To illustrate the variation introduced by decisions 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.[11] 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 provided a pre-prepared treated/untreated-group indicator as specified in Task 2, limited the data set 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 not to further clean the data or limit the sample, although, as discussed below, some did anyway.

After each research task, 2/3 of researchers are randomly assigned to provide peer feedback, while 1/3 are not. Those assigned to peer feedback are randomly paired and receive the other member's work, including the research survey response and a brief write-up, usually with a regression table. Each member conducts a blind review of the other's work, which is then shared with the original researcher. Reviewers are instructed to write the review as though they were reviewing for a journal where a study of this kind "could be published if the work was of high quality." Examples of researcher submissions and peer feedback are provided in Appendix D.

[11] The Task 2 instructions leave some leeway in defining certain variables, particularly control variables such as education or race, for which a specific recoded version is available in Task 3 but 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.

After receiving feedback (or not receiving feedback), researchers may revise their work, but revision is not mandatory, and researchers are not required to satisfy their reviewer. This process therefore differs from typical journal peer review in several ways. Reviewers had completed the same task and therefore had extensive background information. Each reviewer also received feedback from the person they reviewed, and revisions were optional. Because feedback was paired, the design does not separately identify the effect of receiving comments from the effect of reviewing another researcher's work. Fewer than 20% of researchers submitted revisions: 26 in Task 1, 29 in Task 2, and 21 in Task 3. Researchers may still have incorporated comments in later tasks, which we examine in Appendix C.

After each research task and revision, researchers completed a survey about their work.[12] The survey asked them to report their findings, additional information such as sample size and standard errors, and choices made during the analysis, such as sample restrictions, treated-group definitions, estimator, and standard error adjustments. Researchers were also asked to justify these choices.

Several published papers use ACS data to study the effects of DACA, though to the best of our knowledge, none examines the effect on full-time employment. The design used in Tasks 2 and 3 was most directly inspired by Amuedo-Dorantes and Antman (2016), though the outcome and empirical design differ.[13] Researchers were told that prior studies existed and could be consulted for background, but no specific study was listed, and the instructions emphasized that prior work did not provide a "right answer" to match.

More detailed instructions for the research task and a description of the limitation between

[12]The design of this study and survey predates Sarafoglou et al. (2024). Therefore, we do not include questions about researchers' subjective assessments of topics such as methodology choices and consistency of results.

[13]We use a slightly wider time window and age range and do not limit the sample to household heads, which largely explains why the treated share differs from Amuedo-Dorantes and Antman (2016). Other studies examine related DACA outcomes, including educational and economic attainment, health care use and outcomes, and marriage-partner decisions (Amuedo-Dorantes and Antman, 2017; Jones, 2020; Giuntella and Lonsky, 2020; Amuedo-Dorantes and Wang, 2025).

task rounds are in Appendix B. Full instructions for each task, the post-task survey text, and the peer-reviewing instructions are available online at https://osf.io/9p7j6/, which also provides 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.[14]

4 Recruitment and Descriptive Statistics

Researcher recruitment criteria focused on identifying individuals who have produced applied microeconomic research, including non-academic applied microeconomic research. Researchers qualified for the project if they met any of the following criteria: they are academic faculty working in applied microeconomics; they are graduate students *and* have a published or forthcoming paper in applied microeconomics; or they hold a PhD *and* work in a role that involves writing non-academic reports using tools from applied microeconomics to estimate causal effects.[15] 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 (the variance of estimated effects across researchers). With 90 researchers

[14]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).

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

completing all tasks, we would have 90% power to detect this effect.[16] We further assumed attrition rates of roughly 50%, which would suggest recruiting 180 eligible researchers to achieve adequate power. We revised that goal to 200 to account for potentially optimistic assumptions and obtained funding to support payments to 200 researchers.

The project was advertised 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 compiled a list of all 286 economics departments listed in the U.S. News and World Report. We emailed the 264 departments for which we could find email addresses for a front desk or (preferably) the department chair, asking that the message be passed on to all relevant faculty. The recruitment message described the project and its goals, provided a link to a website (https://nickch-k.github.io/ManyEconomists/) with further details on project expectations and incentives for participation, and included 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 participants and authorship on the eventual paper, subject to journal acceptance and authorship policies.

4.1 Participation and Attrition

362 researchers applied for the project (Table 1 shows signups and attrition). Of those, 18.51% were ineligible, mostly because they were graduate students without a forthcoming paper. This left 295 eligible participants. Because this exceeded the budget of 200, we randomly ordered the 282 participants who had signed up by the 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

[16] For comparisons of only two 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.

Table 1: Participation and Attrition

Round	Participants	Attrition
Original Signup	362	18.51% (Ineligible)
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	

was given their place on 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 not completing all steps.

Our assumption that attrition rates would be near 50% was correct, with 49.49% of 295 eligible researchers completing all three stages. Nearly all attrition occurred because researchers failed to complete Task 1, with 141 failing to do so. After that, only 8 failed to complete Task 3. Hence, we have 146 researchers who completed all three research tasks, well above the goal of 90, and that we were able to pay all participants who completed all steps. Except for our analysis of our analyses of researcher characteristics, our base sample consists of the 145 participants who completed all three tasks and estimated the effect of DACA on the probability of employment.[17]

A potential concern with our incentive design is that payment and authorship were offered for completing all tasks, regardless of submission quality. The high recruitment numbers and the concentration of attrition before Task 1 allow us to provide suggestive evidence on whether guaranteed payment affected early participation.[18]

[17] 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.

[18] Seven researchers indicated on the consent form that they did not want payment; three completed the project.

We test this using a regression-discontinuity-style design around the payment cutoff, with results shown in Appendix Figure 2.1 and Appendix Table 2.1. We find no statistically significant effect of guaranteed payment on Task 1 completion, whether using a linear or quadratic specification above the cutoff, and the result is unchanged when late sign-ups are excluded.[19] This is not a perfect test because researchers beyond the cutoff may still have expected payment, especially if they were close to the cutoff. However, if researchers were primarily pursuing a payout, this is one place where we might expect strong evidence of that pattern, and we do not find it.

4.2 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 dropped out between Tasks 1 and 2. The majority of researchers were recruited via social media. Upon signup, researchers were about 90% confident in their ability to complete all three tasks. Average confidence was slightly higher among those who actually finished compared to the initial set of researchers (92% vs. 90%), indicating that those with higher initial confidence were slightly more likely to finish.

The majority of eligible researchers (83%) held PhDs. PhD holders were also more likely than other eligible researchers to complete all three tasks (51.8% vs. 38.0%). These PhDs were split between faculty (62%) and other non-faculty researchers (22%), both of which were more likely than graduate students to finish all three rounds. Most researchers had at least one published

[19] Researchers below the cutoff were not told their order, so we use a zero-order polynomial below the cutoff. Local-polynomial regression-discontinuity estimates are also insignificant. The quadratic estimate is negative and large, but appears to be driven by researchers farther from the cutoff and is offset by the positive linear estimate.

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 Task 2 are omitted because only four researchers dropped out between Task 1 and Task 2. Full results are available upon request.

paper.[20] About a third of initial researchers and 40% of the final set of researchers had worked in either immigration or labor economics, the fields closest to the research task at hand, with 5% having worked in both, although all researchers had worked in applied microeconomics generally.

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 share is similar to the share of faculty who are male at top economics departments in 2017 and among all actively publishing economists in 2019 (Lundberg and Stearns, 2019; Card et al., 2022). About half of the sample was in the United States, and about half was outside.[21] Aside from being skewed toward the United States, the sample largely reflects the group of people who publish work in applied microeconomics. The US overrepresentation is likely driven by emails sent to US economics departments, the project being advertised and carried out in English, and the project organizers being in the United States and advertising the project using their own social media.

5 Results

This section examines variation in effects, samples, and methods across researchers and conditions. We first establish that such variation exists and describe its distribution, then evaluate potential explanations for it based on our preregistered hypotheses. All results, except those

[20] Researchers in the "faculty" or "non-faculty researchers" categories who do not hold PhDs were either people hired into faculty roles without PhDs (such as ABDs or those in faculty positions 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.

[21] The representativeness of the racial mix 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.

that explicitly mention peer feedback, concern work submitted before peer feedback was received and revisions were made.

Our preregistered hypotheses are as follows: (1) the standard deviation of estimated effects, sample sizes, and treated and untreated group sample sizes will decrease from task to task; (2) peer feedback 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 Supplemental Appendix 1.

The results are based on survey responses from researchers about their findings and methodological decisions, and we did not cross-check these responses against the researchers' code. Hence, there is no guarantee of consistency between the two.[22] Consequently, the variation presented here reflects what readers might encounter in published study descriptions. Any discrepancies arising from errors in research reports are beyond the scope of this analysis but could be explored in future research.

5.1 Variation Across Researchers

Table 3 presents the distributions of reported effect estimates, standard errors, analytic sample sizes, and treated-group sample sizes across the three tasks. Because a small number of extreme reports have substantial effects on the means and standard deviations, we rely primarily on medians and interquartile ranges to characterize the central part of each distribution, while using the means and full ranges to describe the tails. The table reports both unweighted effect distributions and distributions using inverse-standard-error weights as a precision-weighted sensitivity analysis.[23]

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

[23]The weighted analysis was not preregistered. Following Auspurg and Brüderl (2024*a*) and Auspurg and Brüderl (2024*b*), weights equal the inverse of the reported standard error and are capped at the 95th

Table 3: Distribution of Reported Effect Estimates, Standard Errors, and Sample Sizes

Variable	N	Mean	SD	Min	Pctl. 25	Median	Pctl. 75	Max	IQR
Task 1									
Effect estimate (unweighted)	145	0.053	0.095	-0.049	0.014	0.030	0.051	0.660	0.036
Effect estimate (inverse-SE weighted)	138	0.044	0.094	-0.049	0.012	0.025	0.042	0.660	0.030
Reported standard error	139	0.019	0.055	0.000	0.005	0.007	0.013	0.460	0.008
Analytic sample size	145	828,318	3,056,037	681	61,600	179,960	356,787	29,536,580	295,187
Treated-group sample size	141	96,395	648,493	270	17,950	34,435	52,581	7,727,201	34,631
Task 2									
Effect estimate (unweighted)	145	0.044	0.100	-0.390	0.015	0.032	0.058	0.850	0.042
Effect estimate (inverse-SE weighted)	141	0.045	0.067	-0.090	0.018	0.033	0.058	0.850	0.040
Reported standard error	141	0.031	0.078	0.001	0.010	0.014	0.020	0.744	0.010
Analytic sample size	144	157,006	1,065,593	6,196	18,981	25,414	48,125	12,609,847	29,144
Treated-group sample size	140	31,948	221,175	3,519	5,953	11,157	15,832	2,627,183	9,879
Task 3									
Effect estimate (unweighted)	145	0.045	0.101	-0.810	0.031	0.050	0.058	0.650	0.027
Effect estimate (inverse-SE weighted)	142	0.064	0.106	-0.810	0.036	0.052	0.061	0.650	0.025
Reported standard error	144	0.059	0.268	0.000	0.015	0.018	0.026	2.747	0.011
Analytic sample size	145	16,904	1,756	7,833	17,379	17,382	17,382	17,832	3
Treated-group sample size	129	9,433	3,008	11	5,149	11,382	11,382	17,383	6,233

Note: Effect-size estimates and standard errors are expressed in proportion units; multiplying by 100 gives percentage-point units. In the unweighted rows, each submission receives equal weight. In the weighted rows, the mean, standard deviation, and quantiles use weights equal to the inverse of each submission's reported standard error. We cap weights at the 95th percentile of the weights, equivalent to a reported standard error of 0.005, to prevent submissions with extremely small reported standard errors from dominating the distribution. Weighted rows exclude submissions with missing or zero reported standard errors. All 145 researchers reported an effect estimate in each task; smaller values of N reflect skipped survey items. The lower N for the Task 3 treated-group sample size largely reflects researchers who regarded that quantity as implied by the provided data.

The unweighted means vary relatively little across tasks, from 0.044 to 0.053, but these means obscure important differences in the shapes of the distributions, illustrating why measures of the center and the tails are best considered separately. In Task 1, the mean estimate of 0.053 lies above the 75th percentile because several large positive estimates pull the mean upward; the median estimate is only 0.030. In Task 2, the mean is 0.044 and the median is 0.032. In Task 3, by contrast, the mean of 0.045 lies below the median of 0.050 because of estimates in the negative tail.

Agreement in the central part of the effect distribution changes non-monotonically as researcher discretion is restricted. In Task 1, where researchers had the greatest freedom over research design and data preparation, the middle half of unweighted estimates ranges from 0.014 to 0.051, producing an IQR of 0.037. Task 2 imposes a common research design but does not

percentile of the positive, finite inverse-standard-error weights calculated across the three tasks. The construction of the weights, exclusions from the weighted sample, and reasons for differences in sample sizes across rows are described in the table note.

increase agreement: the IQR rises by approximately 16 percent, to 0.043, with estimates ranging from 0.015 to 0.058. Agreement is greatest in Task 3, when researchers receive pre-cleaned data in addition to the specified design. The IQR falls to 0.027—37 percent below its Task 2 value and 27 percent below its Task 1 value—with the middle half of estimates ranging from 0.031 to 0.058.

These distributions indicate moderate disagreement about the magnitude of the effect but little disagreement about its sign, since both endpoints of the interquartile range are positive in all three tasks. The increased agreement in Task 3 reflects primarily an increase in the lower end of the central distribution, from 1.4 percentage points in Task 1 to 3.1 percentage points in Task 3, while the upper end changes only from 5.1 to 5.8 percentage points. Thus, specifying a common research design does not by itself narrow the central distribution, whereas combining a common design with common data preparation produces the most concentrated estimates.

This concentration should not be interpreted as agreement throughout the distribution. The full range is from -0.049 to 0.660 in Task 1, from -0.390 to 0.850 in Task 2, and from -0.810 to 0.650 in Task 3. Every task therefore contains estimates that differ sharply in both sign and magnitude from the central results. The main descriptive finding is not complete convergence, but greater concentration in the middle of the distribution once researchers work from the same prepared data, alongside substantial disagreement in the tails.

Inverse-standard-error weighting modestly narrows the central distribution in every task but does not alter this qualitative pattern. The weighted IQR is 0.030 in Task 1, increases to 0.040 in Task 2, and then falls to 0.025 in Task 3. The corresponding weighted medians are 0.025, 0.033, and 0.052. Researchers reporting smaller standard errors therefore tend to report somewhat more similar estimates, but giving those estimates greater weight neither eliminates the increase in dispersion between Tasks 1 and 2 nor changes the conclusion that Task 3 has the narrowest central distribution.

Reported standard errors increase as the research tasks become more restrictive. The median reported standard error doubles from 0.007 in Task 1 to 0.014 in Task 2 and rises further to 0.018 in Task 3. The 25th-to-75th percentile ranges similarly shift upward, from 0.005–0.013 in Task 1 to 0.010–0.020 in Task 2 and 0.015–0.026 in Task 3. This change is therefore not driven solely by a few extreme standard errors; it occurs throughout the center of the distribution and is consistent with the much smaller analytic samples used in the later tasks. At the same time, the means and standard deviations are heavily affected by the right tails, particularly in Task 3, where the mean reported standard error is 0.059, but the maximum is 2.747.

The relationship between researcher variation and reported sampling uncertainty consequently depends on which task and which summaries are compared. For example, the effect-estimate IQR exceeds the median reported standard error in each task, but the gap becomes smaller across tasks as effect estimates become more concentrated and reported standard errors increase. This comparison indicates that between-researcher variation is not negligible relative to the sampling uncertainty reported by a typical researcher. It should not, however, be interpreted as a stable ratio or as a method for combining researcher and sampling uncertainty: the two quantities measure different concepts, and the standard errors rise partly because the later tasks impose much smaller samples.[24]

Restrictions on researcher discretion have a substantially larger effect on reported sample sizes than on effect estimates. In Task 1, the median analytic sample contains 179,960 observations, and the middle half ranges from 61,600 to 356,787, an IQR of 295,187. Some researchers use extremely broad comparison groups, producing samples of several million observations and a maximum exceeding 29 million. Specifying the treated and comparison groups in Task 2 reduces the median to 25,414 and the IQR by approximately 90 percent, to 29,144, although

[24]The ratio of the standard deviation of estimates to the mean reported standard error is especially sensitive to outliers in both distributions and to changes in the sample sizes and designs used across tasks. We therefore use the comparison only to provide scale within this setting, not to compare the amount of researcher variation across tasks, research questions, or fields.

extreme sample sizes remain. In Task 3, when researchers receive prepared data, the median is 17,382 and the IQR is only 3 observations. Thus, specifying the design substantially increases agreement about the analytic sample, while providing the prepared data produces nearly complete agreement among the central half of submissions.

Treated-group sizes show a similar, though less complete, convergence. The median falls from 34,435 in Task 1 to 11,157 in Task 2, close to the Task 3 median of 11,382, while the IQR declines from 34,631 to 9,879. The similarity of the Task 2 and Task 3 medians suggests that the typical researcher comes reasonably close to the intended treated-group definition once it is specified, but the remaining Task 2 dispersion shows that researchers implement the same eligibility rules differently, including through apparent errors discussed in Section 5.3. The reported Task 3 IQR of 6,233 primarily reflects different interpretations of the survey question rather than differences in the data researchers used.[25]

Figure 1 and Figure 2 show the full distributions of effects and sample sizes across tasks.[26] The Task 2 distributions are somewhat bimodal, especially for weighted effect estimates: one mode resembles Task 1, while the other resembles what later appears in Task 3. The bimodality persists in Task 3, but with greater density at the higher mode and more overall agreement. We return to the Task 2 bimodality in Section 5.5.

Reported standard errors increase across tasks, from 0.019 in Task 1 to 0.031 in Task 2 and 0.059 in Task 3, despite relatively stable average effects. This increase primarily reflects the design-induced narrowing of samples, so our later analysis focuses less on the magnitude of reported standard errors than on the adjustments researchers used to calculate them (Section 5.3).

[25] Researchers were instructed not to count DACA-eligible individuals as treated during pre-DACA years when reporting the treated-group sample size. Many instead reported the number eligible for DACA across all years. These responses differ even though the researchers used the same underlying eligibility indicator.

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

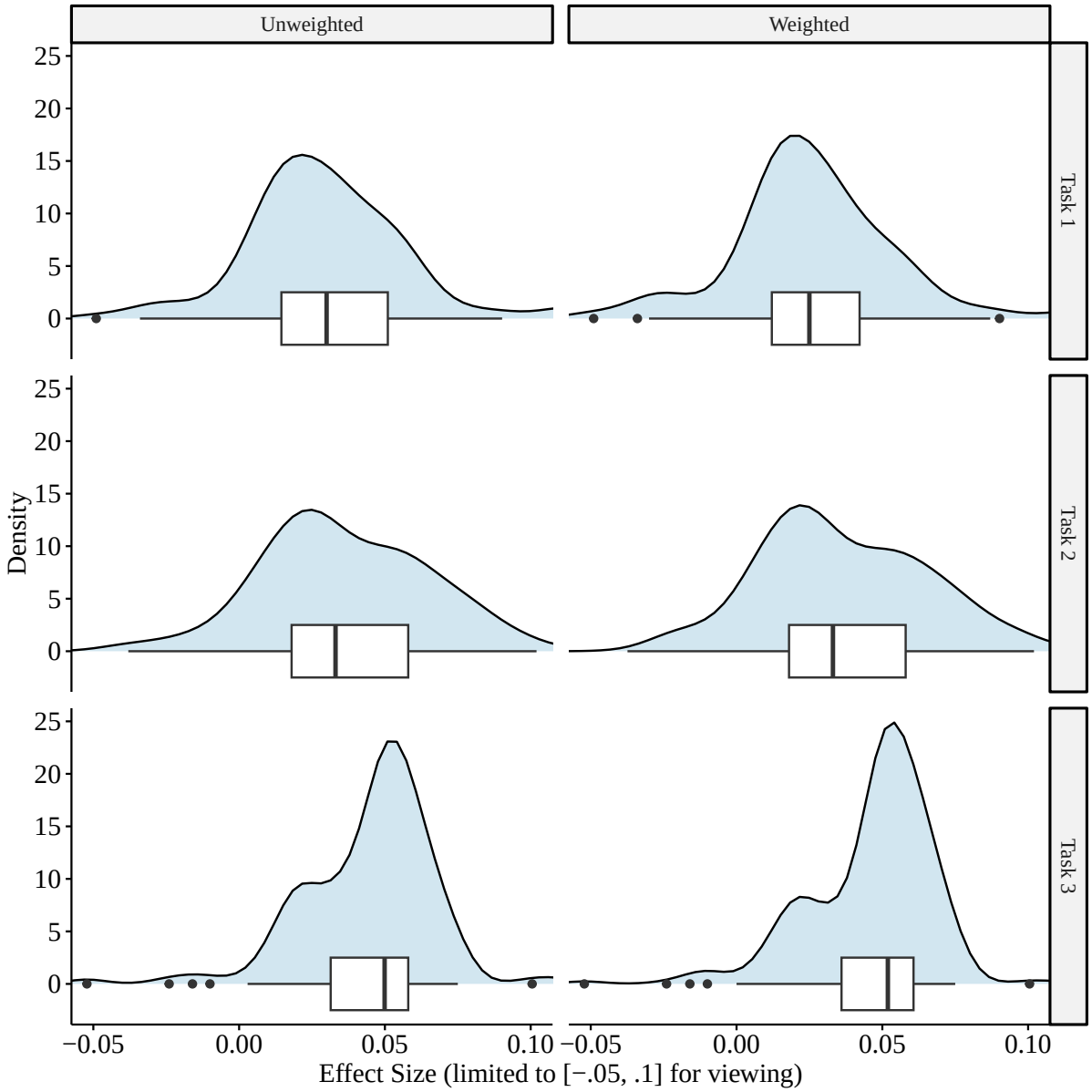


Figure 1: Distributions of Reported Effect Sizes by Task With the Weighted Distributions Using Inverse-Standard-Error Weights

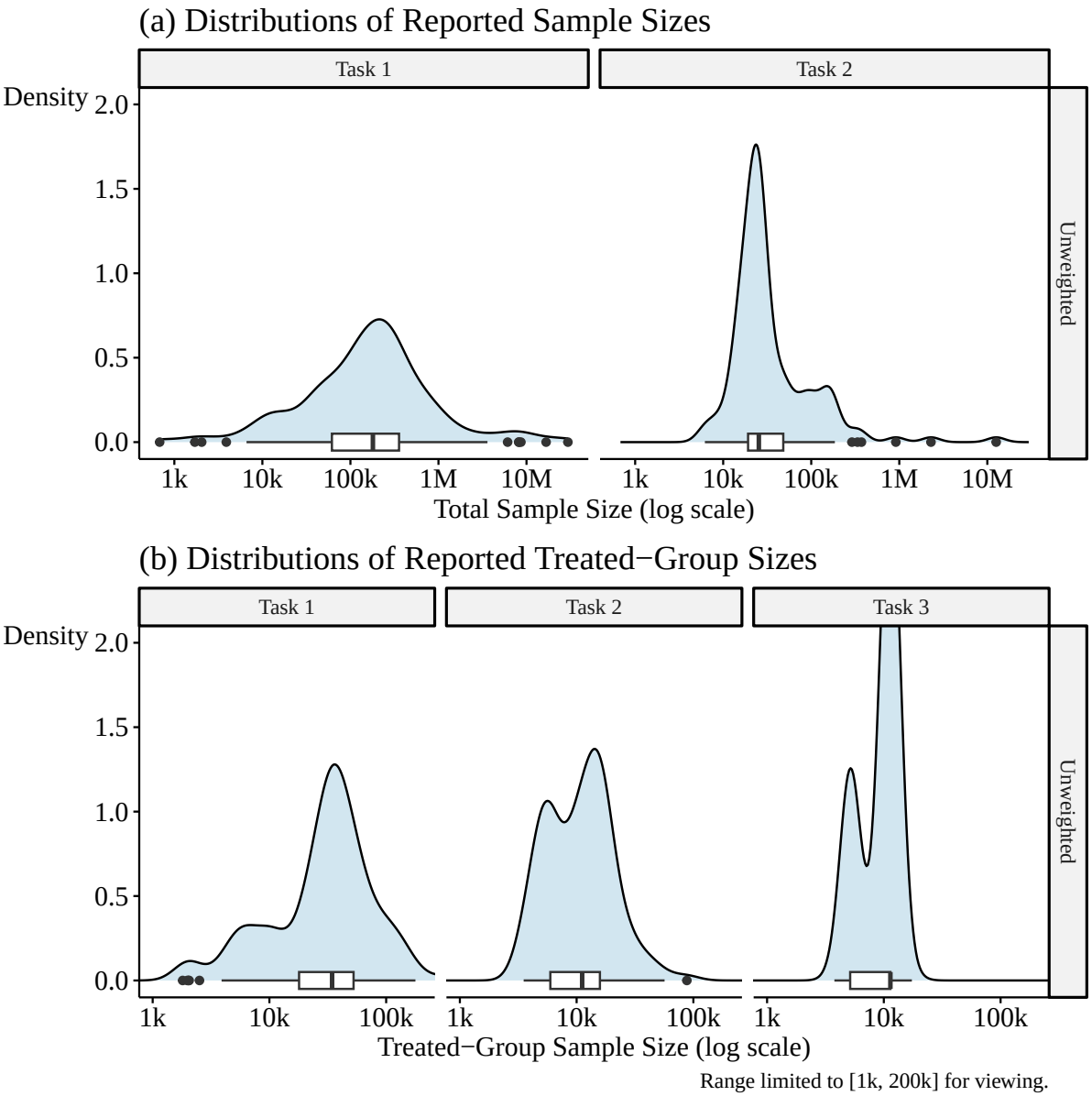


Figure 2: Distributions of Reported Sample Sizes and Treated-Group Sizes

Figure 3 shows each reported effect, ranked within each task, along with confidence intervals reconstructed from the reported standard errors. The center of the distribution shows broad agreement on positive effects, but researchers disagree on whether those effects are statistically significant: 78%, 60%, and 64% of the weighted estimates are statistically significant at the 95% level in Tasks 1, 2, and 3, respectively. Estimates with very large confidence intervals tend to appear in the tails of the distribution.

Researchers' estimates in one task are only weakly predictive of their estimates in later tasks. As Figure 4 shows, there is little visible or linear relationship between a researcher's reported effects across tasks. The only statistically significant correlation is between Tasks 1 and 2, and it is small, at 0.197.

The preregistered variance tests provide limited statistical evidence of changes in dispersion across tasks. For effect sizes, we do not reject any of the preregistered Levene tests for equal variance at the 95% level; the lowest p-value is 0.197, from the comparison of Task 1 to Task 1 Revision. However, for sample sizes, we reject equal variance between Tasks 1 and 2 at the 0.05 level, consistent with the narrowing expected when the treated and comparison groups are specified.

The preregistered regression of squared deviations from round means on round number yields similar conclusions. None of the round coefficients are statistically significant at the 95% level, although the sample-size coefficients are negative as expected (Appendix Table 2.2). Finally, peer feedback is not associated with a statistically significant or consistent reduction in the variance of effects or sample sizes in subsequent tasks (Appendix C).

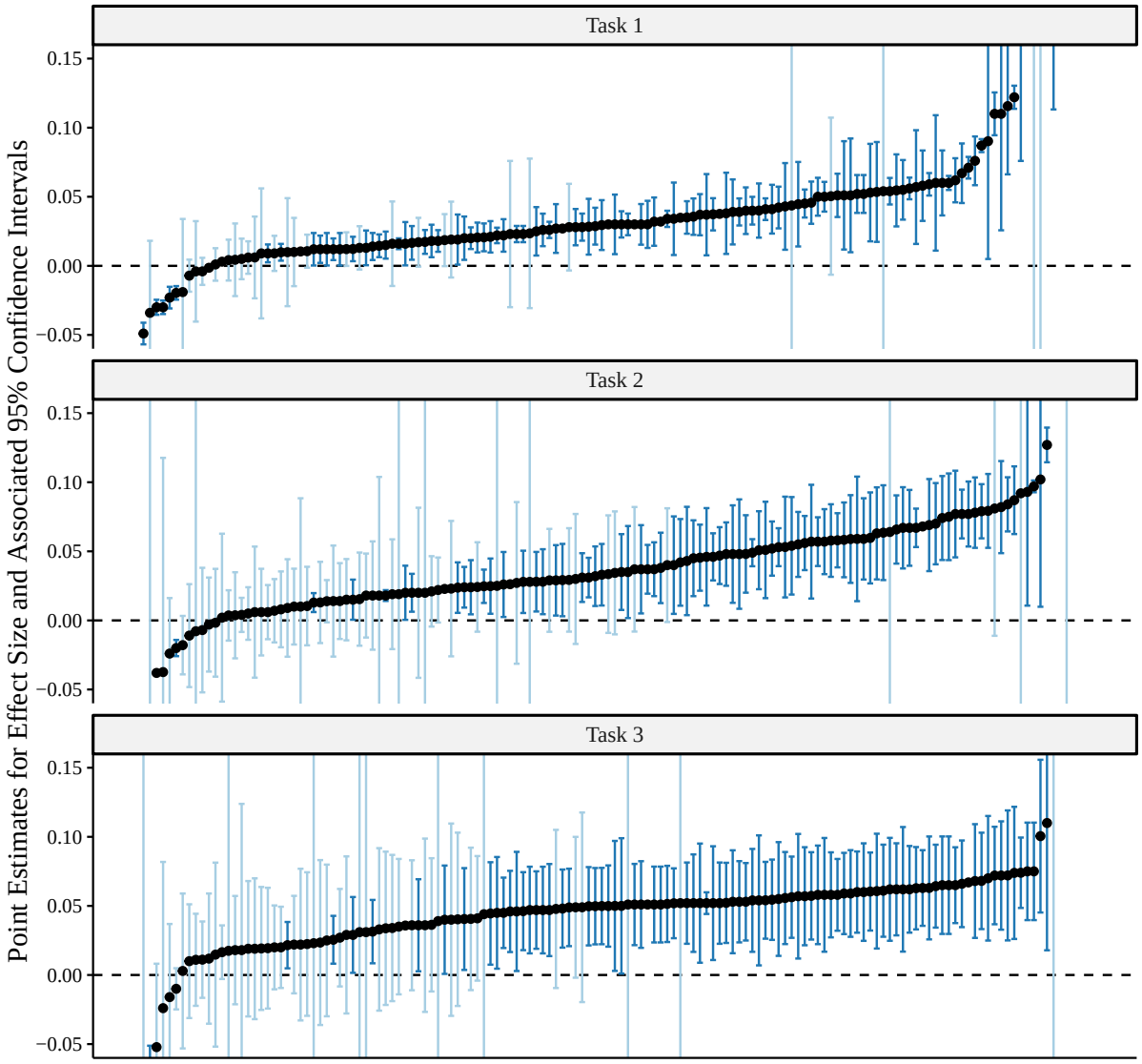
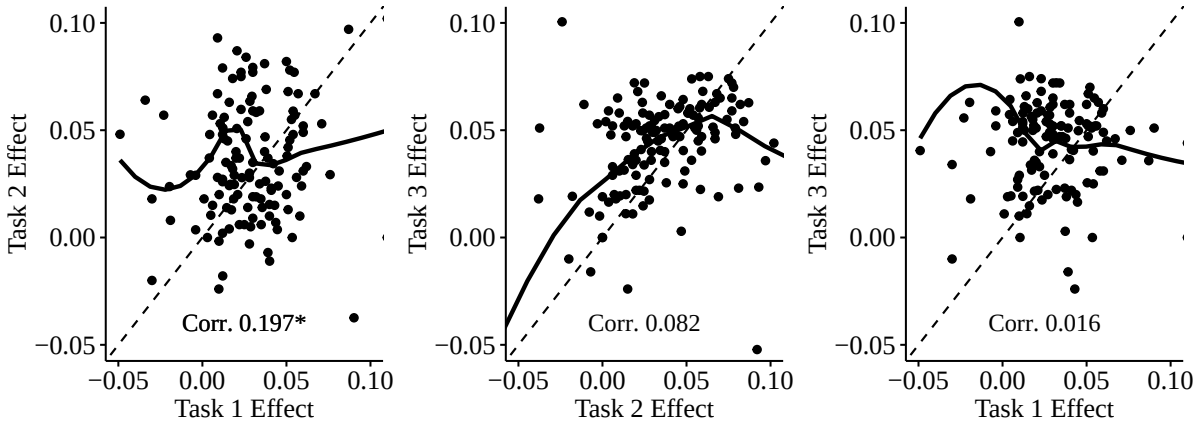


Figure 3: Specification Curve for All Reported Estimates by Task with Estimates Ordered From Smallest to Largest



Visible range limited to effects from 0 to .1. Dashed lines indicate that the researcher reported the exact same estimate in both rounds.
*/**/***: p < .1/.05/.01.

Figure 4: Same-Researcher Effect Sizes Across Tasks

5.2 Assessment of Research Quality

One possible explanation for the findings in Section 5.1 is that variation stems from low-quality research submissions. Compensation in the Many-Economists project depended on completing all three tasks, rather than on the quality of work, so some researchers may have devoted limited effort or submitted analyses with fundamental flaws, skewing the results.

To assess this possibility, we used Google Gemini 2.5 Pro to infer the quality of each reviewed submission from the substantive concerns raised in its peer feedback. The model assigned scores from 1, indicating fundamental flaws, to 5, indicating that no substantive revisions were needed. A full description of the prompting process, as well as what each score means, is in Supplemental Appendix 4. Because the scores are inferred from peer feedback, we use them as relative measures of quality within the study rather than as absolute assessments of publishability.

The distribution of scores rounded to the closest integer by task is shown in Table 4. Scores are concentrated in the middle categories. The mean changes little between Tasks 1 and 2,

Table 4: LLM Assessment of Peer Feedback Positivity

Round	Quality score 1	Quality score 2	Quality score 3	Quality score 4	Quality score 5	N
Task 1	5.1%	36.4%	40.4%	16.2%	2.0%	99
Task 2	4.1%	32.0%	43.3%	19.6%	1.0%	97
Task 3	1.0%	25.3%	33.3%	33.3%	7.1%	99

Note: Entries are the percentage of reviewed submissions receiving each score. Higher scores indicate a more favorable assessment of the underlying submission inferred from the substantive concerns raised in the peer feedback: 1 indicates fundamental flaws; 2, a need for major reanalysis or redesign; 3, meaningful but non-fundamental revisions; 4, minor revisions or polish; and 5, no substantive revisions. Gemini 2.5 Pro scored each peer report ten times; the ten scores were averaged and rounded to the nearest integer. Scores are available only for submissions assigned to peer review and are intended for comparisons within this study, not as absolute assessments of submission quality.

from 2.74 to 2.81, before increasing to 3.20 in Task 3. The Task 3 shift is consistent with improved submission quality as the research process became more structured, although it may reflect both better execution and the fact that the prepared data limited the scope for serious errors.

Table 5 examines within-researcher variation among researchers with scored peer feedback in at least two tasks. Twenty-seven percent received the same rounded score in every scored task, and 73 percent had scores differing by no more than one point. Thus, large changes in assessed quality across a researcher's submissions were uncommon, although this descriptive pattern should be interpreted cautiously given the coarse scale and small number of observations per researcher.

Table 6 mirrors Table 3, but the sample is limited to research tasks that received peer feedback and had a rounded peer feedback score of 3 or higher (58 entries in Task 1, 62 in Task 2, and 73 in Task 3).[27] Higher-quality research shows an almost identical distribution to the full sample, with unweighted effect-size IQRs of 0.035 in Task 1 (vs. 0.036 in the full sample), 0.042 in Task 2 (vs. 0.042), and 0.025 in Task 3 (vs. 0.027). Weighted-effect IQR comparisons are wider for

[27]The results in this section are nearly identical if, instead of limiting the analysis to tasks that received peer feedback, we impute un-reviewed scores using a researcher's peer feedback scores on other tasks (for example, if someone received a score of 3 on Task 1 and 4 on Task 2, but no score on Task 3, we impute the Task 3 score to be 3.5).

Table 5: Minimum and Maximum Quality Scores for Researchers with 2+ Reviews

	Maximum Score				
Minimum Score	1	2	3	4	5
1	0	6	2	0	0
2	0	7	23	17	5
3	0	0	15	20	4
4	0	0	0	7	1
5	0	0	0	0	0

Note: The sample includes researchers that received scored peer feedback in at least two tasks. Rows give each researcher's minimum rounded quality score and columns give the maximum. Diagonal cells therefore indicate no change across tasks; cells immediately above the diagonal indicate a one-point range. Score definitions are provided in Table 4.

Task 1 (0.030 vs. 0.040), close for Task 2 (0.039 vs 0.040), and narrowed for Task 3 (0.014 vs 0.025). For sample size, the quality-restricted sample has a much smaller range than in the full sample, with IQRs of 251k (vs. 300k) and 10k (vs. 30k) in Tasks 1 and 2, respectively. The especially tight bounds on treated-group sample size indicate a higher ability to adhere to task instructions in this higher-quality research.

Figure 2.4 through Figure 2.6 in Supplemental Appendix 2 reproduce the results from Section 5.1 using the quality-controlled sample. We observe some differences compared with the main analysis. In particular, we see less bimodality in Task 2 (consistent with another analysis of Task 2 bimodality below in Section 5.5). There are fewer outliers overall and more agreement.

Overall, perceived research quality accounts for some of the variation across researchers, with higher-rated analyses yielding fewer outlier estimates and greater agreement. This suggests that some dispersion in our main results reflects analyses unlikely to survive a quality screen like that associated with review for publication in a journal. The relevant comparison, however,

Table 6: Distribution of Reported Effect Estimates, Standard Errors, and Sample Sizes among Submissions Scored 3 or Higher

Variable	N	Mean	SD	Min	Pctl. 25	Median	Pctl. 75	Max	IQR
Task 1									
Effect estimate (unweighted)	58	0.057	0.108	-0.030	0.019	0.030	0.053	0.660	0.035
Effect estimate (inverse-SE weighted)	57	0.061	0.124	-0.030	0.014	0.030	0.054	0.660	0.040
Reported standard error	57	0.014	0.037	0.002	0.005	0.007	0.011	0.283	0.006
Analytic sample size	58	537,205	2,180,505	2,045	48,311	143,738	299,359	16,695,554	251,048
Treated-group sample size	56	44,333	48,817	1,981	16,328	34,276	49,478	309,478	33,150
Task 2									
Effect estimate (unweighted)	62	0.036	0.031	-0.037	0.015	0.034	0.057	0.127	0.042
Effect estimate (inverse-SE weighted)	61	0.040	0.031	-0.037	0.019	0.037	0.058	0.127	0.039
Reported standard error	61	0.023	0.045	0.006	0.011	0.015	0.020	0.360	0.009
Analytic sample size	61	38,767	39,456	6,212	19,367	24,595	29,375	181,299	10,008
Treated-group sample size	59	10,776	5,820	3,579	5,407	9,664	15,254	27,530	9,847
Task 3									
Effect estimate (unweighted)	73	0.052	0.062	-0.016	0.033	0.051	0.058	0.550	0.025
Effect estimate (inverse-SE weighted)	71	0.051	0.039	-0.016	0.046	0.052	0.060	0.550	0.014
Reported standard error	73	0.025	0.029	0.000	0.014	0.017	0.024	0.179	0.010
Analytic sample size	73	17,222	1,003	10,205	17,380	17,382	17,382	17,832	2
Treated-group sample size	64	9,987	2,749	5,149	8,313	11,382	11,382	17,383	3,069

Note: The sample is limited to submissions assigned to peer review that received a rounded LLM-based quality score of 3 or higher. A score of 3 indicates a submission requiring meaningful but non-fundamental revisions; scores of 4 and 5 indicate minor revisions or no substantive revisions, respectively. Table 4 describes the complete scoring procedure. Effect estimates and standard errors are reported as proportions. Unweighted rows give each submission equal weight. Weighted rows use the We cap weights at the 95th percentile of the weights, equivalent to a reported standard error of 0.005 to prevent submissions with extremely small reported standard errors from dominating the distribution. Weighted rows exclude submissions with missing or zero reported standard errors. Other differences in N reflect skipped survey questions. Variation in the reported Task 3 treated-group size partly reflects different interpretations of the survey question despite use of the same provided eligibility indicator.

is not simply with published papers. Publication selects not only on technical quality but also on novelty, contribution, and positioning. Several researchers could produce similar estimates using similar designs, but only one such paper would likely be published.

Researchers estimated an effect using a design closely related to an existing published paper (Amuedo-Dorantes and Antman, 2016), and our peer-feedback assessment removes analyses with more obvious problems. We interpret the quality-filtered results as representing what remains after removing these more obvious problems, although they still receive less scrutiny than an actual publication. Published papers may therefore show less dispersion than even our high-quality subset, but the exercise is intended to make the remaining variation informative about disagreement among credible empirical analyses.

While this section finds that higher-quality work shows greater agreement in the aggregate,

perceived quality does not change the overall pattern of how agreement varies across the tasks (which is relevant to the question of how these results might apply to actual publications of still-higher quality). Going from Task 1 to Task 2, the IQR widens from 0.036 to 0.039 unweighted or 0.035 to 0.039 weighted. Then, from Task 2 to Task 3, the IQR shrinks as agreement improves, from 0.039 to 0.019 unweighted or 0.039 to 0.014 weighted. Although there is more agreement among this group, meaningful variation remains in both estimates and sample sizes. For the remainder of the study, we return to using the much larger full sample of data.

5.3 Researchers' Analytic and Sample Choices

This section examines to what extent researchers' analytic and sample choices explain variation in reported effects and sample sizes. These are the choices most visible to readers and peer reviewers and include the estimator, use of survey weights, standard-error adjustment, controls, functional forms, and sample restrictions.

For most researchers, the estimator, use of ACS sampling weights, and standard-error adjustment did not vary across tasks, so Table 7 pools choices across tasks. The choices generally align with common applied econometric practice. Although the dependent variable is binary, linear regression was used in 82% of entries, while 13% used logit or probit. Many of the linear regressions used a fully or nearly fully saturated difference-in-differences design, which mitigates the main concerns with linear probability models. Other researchers used matching estimators or recently introduced difference-in-differences estimators such as Callaway and Sant'Anna (2021). Despite the IPUMS recommendation, only 25% used survey weights.

Standard-error adjustments varied more than estimators. Fifty-five percent of entries clustered standard errors at some level, though the clustering level differed, and 17% used

Table 7: Reported Estimation and Inference Choices

Variable	N	Percent	Variable	N	Percent
Method	435		S.E. Adjustment	435	
... Linear Regression	357	82%	... Cluster (State)	118	27%
... Logit/Probit	56	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	435		... None	95	22%
... No Sample Weights	326	75%			
... Sample Weights	109	25%			

Note: The unit of observation is a researcher-task submission. The table pools the 145 submissions from each of Tasks 1–3, for $N = 435$. Within each section, categories are mutually exclusive and percentages sum to 100, apart from rounding. "Method" refers to the primary estimator rather than the research design; for example, a difference-in-differences design estimated using ordinary least squares is classified as linear regression. "Weights" indicates whether ACS person-level sampling weights were used. "Het-Robust" denotes heteroskedasticity-robust standard errors without clustering; the remaining categories describe the reported clustering or other standard-error adjustment.

heteroskedasticity-robust but not cluster-robust standard errors. Clustering at the state or state-year level is likely difficult to justify in this setting because treatment is not assigned at the state level (Abadie et al., 2023). At the same time, clustering at the group level is a common convention in panel-data work (Bertrand, Duflo and Mullainathan, 2004), and some software aimed at economists uses group-level clustering as the default (StataCorp, 2024).

The choices in Table 7 account for only a modest share of the variation in estimated effects. Regressing effect sizes on indicators for estimator, survey weights, and standard-error adjustment (a non-preregistered analysis) yields R^2 values of 0.162, 0.073, and 0.023 in Tasks 1–3, respectively. These choices explain more of the variation in statistical significance in Tasks 2 and 3, with corresponding R^2 values of 0.026, 0.227, and 0.338. Estimation method explains most of this variation in significance, followed by standard-error adjustment. However, the signs of the estimated associations between specific standard-error adjustments and statistical significance are not consistent across tasks.

The researchers disagreed substantially on the appropriate controls.[28] Across the 435 submissions, there were 333 unique sets of included covariates, and 64% chose 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 substantial variation in the sets of included covariates. In Task 1, for example, there were ten covariates with inclusion rates between 0.2 and 0.8, 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 affect the effect estimates: the mean reported effects differ by 2.3 percentage points when we compare 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 the difference toward a larger value than noise alone would produce. Appendix Table 2.3 shows that there do not appear to be major differences in the average reported standard errors as a function of which covariates are included (from 0.016 at the minimum to 0.060 at the maximum, noting that 0.060 is an outlier and the second-highest is only 0.046), or in the standard deviation of the effect distribution among researchers (from 0.042 to 0.133).

Functional-form choices explain more variation than the choice of which covariates to include. As Table 8 shows, the difference between the highest- and lowest-average-effect variants was 3.4 percentage points for age, 0.7 percentage points for education, and 2.4 percentage points for state/year controls. In this setting, therefore, how researchers included common controls mattered more than whether they included them.

[28] Appendix Table 2.3 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 8: 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

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.

Sample construction varied substantially, including in Task 2, where the instructions specified the treated-group definition. Table 9 reports our hand-coded sample restrictions based on researchers' code rather than from survey responses.[29] For most eligibility components, the option matching the Task 2 instructions (which is given as the first row in each section of the table) was the most common treated-group definition, but adherence was far from complete. Birthplace, citizenship, age at migration, year-quarter age, and Hispanic ethnicity matched the instructions in 77–81% of Task 2 treated-group definitions; education and veteran status matched in 74%. Year of immigration showed more disagreement, with the matching share depending on whether "<= 2007" is treated as equivalent to "< 2007." Years in the USA is the only case where the correct option was unambiguously not the most common: DACA

[29]This is the only part of the paper that does not rely on researcher survey responses. For each researcher's Task 1 and Task 2 code, organizers read the code and recorded aspects of the sample definitions used for the overall analytic sample and for the treated group, including definitions that appeared to result from coding errors. For example, some researchers likely used non-veterans rather than veterans because they coded VETSTAT == 1, which indicates non-veteran status in IPUMS. Earlier versions of the analysis based on self-reported sample limitations found similar mismatch rates, so coding errors alone do not account for these results.

eligibility is based on continuous residence in the United States. However, most researchers used only year of immigration rather than the YRSUSA variables that directly track continuous residence.

We also see meaningful variation between researchers when there is not a clear "correct" option. In the analytic sample definition, no specific usage of any one variable was used by more than 84% of the sample.

The effect of any single sample-construction choice is difficult to estimate because most alternatives were used by few researchers. Table 10 therefore focuses on two comparisons with enough observations to be informative: whether researchers used YRSUSA and whether they used "< 2007" or "<= 2007" for the year of migration. These choices had modest effects on median estimates in Task 1 but were more consequential for sample size: the less restrictive year-of-migration rule produced median sample sizes roughly three times as large. In Task 2, effect-size differences across these treated-group definitions were larger but still moderate. Other sample restrictions are sometimes associated with large differences in effects and sample sizes, though most such comparisons are based on small cell sizes (Appendix Tables 2.4 and 2.5).

Overall, the visible analytic choices most commonly scrutinized in peer review explain only part of the variation among researchers. Researchers differed substantially in covariate sets, functional forms, and sample construction, yet estimator choice, weighting, standard-error adjustment, and covariate inclusion do not fully account for the dispersion in reported effects. The strongest evidence of consequential variation in this section comes from sample construction and implementation, which is consistent with the larger changes in agreement observed when the data were pre-cleaned in Task 3.

Table 9: 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%

Notes: 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 10: Effect Estimates and Sample Sizes by Sample-Construction Choices

	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

Notes: The unit of observation is a researcher-task submission. The first set of effect percentiles groups submissions according to the rule used to define the treated group; the remaining columns group submissions according to restrictions applied to the full analytic sample. "< 2007" includes immigrants arriving before 2007, whereas "≤ 2007" also includes those arriving during 2007. "Used YRSUSA" indicates that the researcher used the IPUMS measure of years lived continuously in the United States; "No YRSUSA" indicates that this measure was not used. Effect estimates are reported as proportions, so 0.01 corresponds to one percentage point. Sample-size percentiles refer to the total analytic sample. The table reports the 25th, 50th, and 75th percentiles within each category.

5.4 Researcher Characteristics and Effects

As listed in our preregistration, the two project organizers independently evaluated the relationship between researcher characteristics and the effects they reported using a many-analyst approach, with both organizers analyzing the same data and research question in separate analyses. The preregistration called for independent analyses 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 attrition from the study after Task 1 was minimal, part (b) was dropped from the analysis. Full results from each project organizer can be found in Supplemental Appendix 3.

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 the level of the effect estimate reported, the deviation of the estimate from the mean, or changes in the estimate from task to task. The only relevant difference we found was that a minority of researchers who used the R programming language were more likely to report outlier estimates than those who used Stata. In Task 3, for example, 0.9% of Stata users were more than 0.1 in absolute distance from the effect estimate sample mean, while 12.5% of R users were. Our analysis is limited by the range of researcher characteristics we collected; in particular, we do not have information on researcher personality or political leanings, as in Borjas and Breznau (2024), Jelveh, Kogut and Naidu (2024), or Sulik et al. (2023).

5.5 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. While we see this pattern for sample sizes and some researcher choices, the distribution of effects actually widened between Tasks 1 and 2. There is also an emerging bimodality in both effects and sample sizes, with most researchers reporting estimates that reflected the distribution of effects seen in Task 1, while a smaller group reported larger effects more like those in Task 3. This unpreregistered analysis examines potential explanations for these unexpected findings.

A major contributing factor to Task 2's bimodality is whether the treated-group definition in the instructions was implemented. Task 2 provided a precise definition of who should be included in the treated group, and we examine whether each researcher followed the full set of treated-group definition instructions exactly. Any mismatch could be small, such as using "<= 16" instead of "< 16" for age at migration, or large, such as omitting the requirement that eligible people be non-citizens. Figure 5 shows the distribution of effects for researchers who followed the definition precisely compared with those who had any mismatch in their criteria. 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. Many other decisions are made, and the treated-group implementation captures only one angle. Furthermore, the share of researchers matching exactly across all fields is fairly low at 32.4%, as shown in Table 11. Recall that even very minor mismatches are counted as mismatches. Perfect-match rates were slightly below average among researchers who mostly work in labor (31.9% for

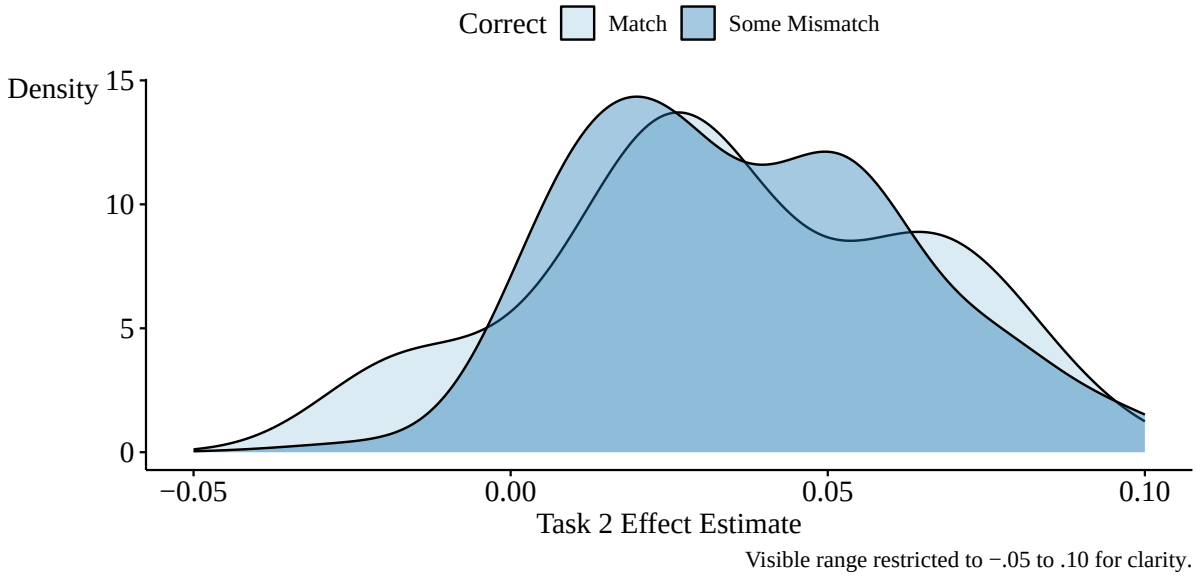


Figure 5: Task 2 Effect Distributions Among Those with Exact Treated-Group Definition Matches vs. Those with Some Mismatch

Table 11: 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

Notes: Any deviation from treated-group definition is treated as a mismatch, even if minor.

labor-only and 0% for labor-and-immigration).[30]

Several other correlates did not explain the bimodal outcomes in Task 2. The reported sample sizes and standard errors for Task 2 do not strongly explain the reported effects (see Appendix Figure 2.2 and Appendix Figure 2.3). The bimodality is also not a feature of researchers trying to make their Task 2 results consistent with their Task 1, as Figure 4 in Section 5.1 shows.

[30]Researchers working in immigration but not labor had a 100% match rate, but these are small samples.

6 Discussion

We examine how researcher variation arises across stages of an applied empirical workflow by having 146 research teams complete the same causal inference task under progressively tighter constraints. The central pattern is not merely that researchers disagree, but that disagreement is concentrated in specific stages of the research process. Observed variation in researchers' choices is especially pronounced in data preparation and sample construction, less so in modeling choices, where widely shared conventions exist. Choices were not responsive to peer feedback or researcher background.

Despite substantial heterogeneity in data cleaning, variable construction, and modeling decisions, reported policy effects were relatively similar across researchers near the center of the distribution. Even in the least constrained task, most estimates clustered in a narrow range, with disagreement driven primarily by outliers. Constraining the research design alone did not reduce dispersion in estimated effects and, in some cases, increased it due to imperfect adherence to the specified protocol. In contrast, although meaningful variation remained, the closest agreement in reported effects occurred when pre-cleaned data were provided. These patterns highlight that differences in data preparation, not just analytical discretion, can substantially contribute to researcher variation.

At the same time, several important limitations shape how these findings should be interpreted. First, the results are specific to the empirical task studied here: estimating the effect of DACA eligibility on full-time employment using repeated cross-sectional survey data. This research question lends itself naturally to a familiar before–after comparison with a clear treatment definition, which likely limits the scope for disagreement relative to more complex settings. As a result, the magnitude of variation observed here should not be taken as representative of applied economics as a whole. Rather, the contribution of this study lies in illustrating

the relative importance of different sources of variation—particularly the prominence of data preparation choices—even in a comparatively structured research environment.

Second, the incentive structure in this project differs from that of conventional academic research. Participants were compensated for completing the tasks, but compensation was not tied to effort or quality, and peer feedback carried no formal consequences. This design choice limits how much of the observed variation can be attributed solely to principled methodological disagreement. Some variation may reflect differences in effort or attention rather than genuine differences in judgment. We address this concern by examining higher-quality submissions, as assessed by peer feedback, and find that focusing on these submissions reduces dispersion but does not eliminate it or alter the relative agreement across the three research tasks. Nonetheless, this study was not designed to replicate the incentive environment of journal publication, and conclusions about how researcher variation operates under full publication incentives should therefore be drawn with caution.

Taken together, these findings suggest that researcher variation in applied economics is not uniformly distributed across research decisions. When conventions are well established, such as the use of linear models in difference-in-differences settings, researchers tend to converge. When conventions are weaker, less visible, or less routinely scrutinized—most notably in data cleaning and preprocessing—researchers diverge, sometimes with meaningful consequences for reported results. Standard error adjustments were an illustrative case of the contrast between strong and weak conventions: most researchers adjusted their standard errors, a strong convention, with a majority applying clustering to this difference-in-differences-style setting, another fairly strong convention, but there was little agreement on the level at which to cluster, a weak convention. The most common clustering choices, where clustering occurred at the state or state-year level, followed common conventions for difference-in-differences, even though that convention does not apply to this setting where treatment is not assigned at the state level and

the conventional choice is not best-practice (Abadie et al., 2023). Understanding this pattern is essential for interpreting empirical findings, even when headline estimates appear stable.

7 Reading Research with Researcher Variation in Mind

Disagreement among researchers is a central feature of scientific progress rather than a pathology to be eliminated. Differences in how researchers approach the same empirical question can reflect informed methodological disagreement, domain knowledge, or alternative modeling perspectives. Such diversity is often productive in advancing understanding. Evidence from many-analyst studies across economics and related fields shows that substantial dispersion in results can arise even when researchers are competent and well-intentioned (Silberzahn et al., 2018; Huntington-Klein et al., 2021; Menkveld et al., 2024).

It is more concerning when researcher variation reflects errors or choices that are largely invisible to readers and therefore escape scrutiny. Understanding where researcher variation arises and how large it is relative to other sources of uncertainty is essential for interpreting empirical findings.

In our study, researcher variation in estimated policy effects was meaningful yet limited in scope. Across tasks, the interquartile range of estimated effects spanned roughly 2 to 4 percentage points. Importantly, in this setting, this dispersion is comparable in magnitude to the uncertainty conveyed by reported standard errors, underscoring that conventional uncertainty summaries can understate total uncertainty even when estimates appear relatively stable. This aligns with a growing body of literature showing that researcher-induced variation can be comparable in magnitude to, or larger than, sampling uncertainty, implying that conventional uncertainty summaries may understate total uncertainty around empirical estimates

(Huntington-Klein et al., 2021; Holzmeister et al., 2024; Menkveld et al., 2024; Stevenson and Fischman, Forthcoming).

Researchers and readers are accustomed to scrutinizing identification strategies and modeling choices, and these decisions are typically visible in published work. By contrast, some of the most consequential sources of variation identified here—data cleaning, variable construction, and sample definition—are often only partially documented or omitted entirely from research discussions. This study is not sufficient evidence to prove that data preparation is the most important source of researcher variation, but it does establish that there is variation in data preparation choices and that, in our case, pre-cleaning the data led to the most meaningful reduction in disagreement.

In our study, where commonly accepted practices existed, such as linear modeling in a difference-in-differences framework or the use of adjusted standard errors, researchers largely converged. Where such conventions were weaker or less explicit, researchers diverged, sometimes substantially. This pattern is consistent with findings from many-analyst and multiverse studies, which document greater dispersion when analytic choices are weakly standardized or poorly reported (Steegen et al., 2016; Menkveld et al., 2024).

A central insight from this study is not that researcher variation is ubiquitous, but that it is unevenly distributed across stages of the research process. The absence of shared conventions and transparent reporting in data preparation creates scope for variation that is both difficult for readers to detect and potentially consequential for results. This helps explain why agreement on headline estimates can coexist with substantial heterogeneity in underlying research decisions, a phenomenon also documented in earlier many-analyst work (Silberzahn et al., 2018; Breznau et al., 2022).

It is neither feasible nor desirable to eliminate researcher variation entirely. Empirical research

inevitably involves judgment, and rigid standardization risks discouraging useful exploration or entrenching flawed practices. The relevant distinction is between variation that reflects informed methodological disagreement and variation that arises when there are no shared expectations about how key steps—such as data cleaning—should be conducted or reported. Reducing the latter does not require uniform answers but clearer norms for process and transparency.[31]

Several directions for future research emerge from the unresolved questions identified in the review in Section 2 and from the patterns documented in our many-analyst study. One underexplored area concerns incentives and effort in many-analyst designs. As Auspurg and Brüderl (2024*b*) emphasize, many-analyst studies necessarily abstract from the full incentive structure of academic publishing. Understanding how researcher variation changes when incentives more closely resemble those in conventional research—such as career stakes, publication pressure, or reputational concerns—remains an open question. Designing studies that vary incentives in a controlled way while preserving feasibility would substantially improve our ability to interpret observed variation.

A second open question concerns context dependence. Researcher variation likely varies across empirical settings, depending on data complexity, identification strategy, and the availability of established conventions. Our study focuses on a familiar difference-in-differences setting and widely used survey data. Future work comparing researcher variation across research designs, subfields, and data environments would help identify which features of empirical work are most conducive to agreement and which are most prone to arbitrary dispersion.

Third, progress in this area is constrained by the absence of a common metric for comparing

[31] For data cleaning in particular, neither formal training nor reporting standards are well established in economics. Even when journals require replication packages, these often start with already-cleaned datasets, limiting scrutiny of earlier preprocessing decisions (American Economic Association, 2024). Other disciplines have developed explicit guidance on data preprocessing and documentation (Osborne, 2012; Jafari, 2022), and recent economics textbooks have begun to treat data cleaning as a core research skill rather than a purely mechanical step (Békés and Kézdi, 2021).

researcher variation across studies and contexts. Although recent work proposes ways to incorporate researcher variation into uncertainty assessments (e.g., Huntington-Klein et al., 2021; Menkveld et al., 2024; Coqueret, 2024; Stevenson and Fischman, Forthcoming), we lack tools that enable meaningful comparisons of the magnitude of researcher variation across settings. Developing such metrics would support both meta-scientific research and the interpretation of empirical findings.

Fourth, a question that was not relevant when the study was designed: to what extent will human researchers make research decisions in the future, and how will variation change if computers make those decisions instead? There are already two studies that use agentic AI to complete the research task in this paper (Grundl, 2026; McCully, 2026). While neither allows the full range of freedom that the human researchers in this study had (in particular, agents were not given full flexibility in defining their IPUMS download), both studies find that the AI agents produce average results similar to those of human researchers, with a narrower range of variation. By some metrics, the agents produce work of higher quality. Examining the sources of researcher variation among AI agents (as well as which forms of human variation are not repeated by AI agents and whether that is desirable) is a key next step, given the likelihood that AI agents will make an increasing share of research choices in the future.

Finally, future research could more directly investigate the boundary between productive disagreement and error. Distinguishing variation that reflects defensible alternative judgments from variation driven by mistakes, misunderstandings, or low-effort execution is critical for determining when variation should be viewed as informative rather than problematic. A single researcher systematically reporting many potential research avenues can improve communication about this disagreement and perhaps address some of the ways in which publication incentives intersect with research choices (Stevenson and Fischman, Forthcoming). Combining many-analyst designs with systematic quality assessments, replication checks, or targeted

constraints may offer a promising path forward.

Taken together, these directions underscore that researcher variation is not a nuisance to be eliminated but a feature of empirical research that must be understood. By documenting where variation arises in a controlled setting, this study contributes to a growing literature clarifying how empirical uncertainty should be assessed and how empirical findings should be interpreted.

References

Abadie, Alberto, Susan Athey, Guido W Imbens, and Jeffrey M Wooldridge. 2023. "When Should You Adjust Standard Errors for Clustering?" *The Quarterly Journal of Economics*, 138(1): 1–35.

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

Amuedo-Dorantes, Catalina, and Chunbei Wang. 2025. "Intermarriage amid immigration status uncertainty: Evidence from DACA." *Journal of Policy Analysis and Management*, 44(4): 1317–1346.

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

Amuedo-Dorantes, Catalina, and Francisca Antman. 2017. "Schooling and labor market effects of temporary authorization: evidence from DACA." *Journal of Population Economics*, 30(1): 339–373.

Andrews, Isaiah, and Maximilian Kasy. 2019. "Identification of and Correction for Publication Bias." *American Economic Review*, 109(8): 2766–94.

Arel-Bundock, Vincent. 2022. "modelsummary: Data and Model Summaries in R." *Journal of Statistical Software*, 103(1): 1–23.

Auspurg, Katrin. 2025. "Robustness is better assessed with a few thoughtful models than with billions of regressions." *Proceedings of the National Academy of Sciences*, 122(43).

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

Auspurg, Katrin, and Josef Brüderl. 2024*a*. "Claims about "a hidden universe of uncertainty" in social research are exaggerated: How many-analyst studies risk overestimating uncertainty." MetaArXiv Preprint.

Auspurg, Katrin, and Josef Brüderl. 2024*b*. "Toward a more credible assessment of the credibility of science by many-analyst studies." *Proceedings of the National Academy of Sciences*, 121(38): e2404035121.

Auspurg, Katrin, and Josef Brüderl. 2026. "Fragile Evidence for an Ideological Bias in the Production of Research Findings: Comment on Borjas and Breznau." MetaArXiv.

Azoulay, Pierre, Joshua S. Graff Zivin, and Gustavo Manso. 2011. "Incentives and Creativity: Evidence From the Academic Life Sciences." *The RAND Journal of Economics*, 42(3): 527–554.

Banuri, Sheheryar, Stefan Dercon, and Varun Gauri. 2019. "Biased Policy Professionals." *The World Bank Economic Review*, 33(2): 310–327.

Barrett, Tyson, Matt Dowle, Arun Srinivasan, Jan Gorecki, Michael Chirico, and Toby Hocking. 2024. "data.table: Extension of data.frame." R package version 1.15.4.

Bastiaansen, Jojanneke A, Yoram K Kunkels, Frank J Blaauw, Steven M Boker, Eva Ceulemans, Meng Chen, Sy-Miin Chow, Peter de Jonge, Ando C Emerencia, Sacha Epskamp, et al. 2020. "Time to get Personal? The Impact of Researchers Choices on the Selection of Treatment Targets Using the Experience Sampling Msethodology." *Journal of Psychosomatic Research*, 137: 110211.

Becker, Jason, Chung hong Chan, David Schoch, and Thomas J. Leeper. 2023. "rio: A Swiss-Army Knife for Data I/O." R package version 1.0.1.

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." *CREA Discussion Papers*, , (13).

Bertrand, M., E. Duflo, and S. Mullainathan. 2004. "How Much Should We Trust Differences-In-Differences Estimates?" *The Quarterly Journal of Economics*, 119(1): 249–275.

Boehm, Udo, Jeffrey Annis, Michael J Frank, Guy E Hawkins, Andrew Heathcote, David Kellen, Angelos-Miltiadis Krypotos, Veronika Lerche, Gordon D Logan, Thomas J Palmeri, et al. 2018. "Estimating Across-trial Variability Parameters of the Diffusion Decision Model: Expert Advice and Recommendations." *Journal of Mathematical Psychology*, 87: 46–75.

Borjas, George J, and Nate Breznau. 2024. "Ideological Bias in Estimates of the Impact of Immigration." National Bureau of Economic Research Working Paper 33274.

Botvinik-Nezer, Rotem, Felix Holzmeister, Colin F Camerer, Anna Dreber, Juergen Huber, Magnus Johannesson, Michael Kirchler, Roni Iwanir, Jeanette A

Mumford, R Alison Adcock, et al. 2020. "Variability in the Analysis of a Single Neuroimaging Dataset by Many Teams." *Nature*, 582(7810): 84–88.

Breznau, Nate, Eike Mark Rinke, Alexander Wuttke, Hung H. V. Nguyen, Muna Adem, Jule Adriaans, Amalia Alvarez-Benjumea, Henrik K. Andersen, Daniel Auer, Flavio Azevedo, Oke Bahnsen, Dave Balzer, Gerrit Bauer, Paul C. Bauer, Markus Baumann, Sharon Baute, Verena Benoit, Julian Bernauer, Carl Berning, Anna Berthold, Felix S. Bethke, Thomas Biegert, Katharina Blinzler, Johannes N. Blumenberg, Licia Bobzien, Andrea Bohman, Thijs Bol, Amie Bostic, Zuzanna Brzozowska, Katharina Burgdorf, Kaspar Burger, Kathrin B. Busch, Juan Carlos-Castillo, Nathan Chan, Pablo Christmann, Roxanne Connelly, Christian S. Czymara, Elena Damian, Alejandro Ecker, Achim Edelmann, Maureen A. Eger, Simon Ellerbrock, Anna Forke, Andrea Forster, Chris Gaasendam, Konstantin Gavras, Vernon Gayle, Theresa Gessler, Timo Gnambs, Amélie Godefroidt, Max Grömping, Martin Groß, Stefan Gruber, Tobias Gummer, Andreas Hadjar, Jan Paul Heisig, Sebastian Hellmeier, Stefanie Heyne, Magdalena Hirsch, Mikael Hjerm, Oshrat Hochman, Andreas Hövermann, Sophia Hunger, Christian Hunkler, Nora Huth, Zsófia S. Ignácz, Laura Jacobs, Jannes Jacobsen, Bastian Jaeger, Sebastian Jungkunz, Nils Jungmann, Mathias Kauff, Manuel Kleinert, Julia Klinger, Jan-Philipp Kolb, Marta Kołczyńska, John Kuk, Katharina Kunißen, Dafina Kurti Sinatra, Alexander Langenkamp, Philipp M. Lersch, Lea-Maria Löbel, Philipp Lutscher, Matthias Mader, Joan E. Madia, Natalia Malancu, Luis Maldonado, Helge Marahrens, Nicole Martin, Paul Martinez, Jochen Mayerl, Oscar J. Mayorga, Patricia McManus, Kyle McWagner, Cecil Meeusen, Daniel Meierrieks, Jonathan Mellon, Friedolin Merhout, Samuel Merk, Daniel Meyer, Leticia Micheli, Jonathan Mijs, Cristóbal Moya, Marcel Neunhoeffer, Daniel Nüst, Olav Nygård, Fabian Ochsenfeld, Gunnar

Otte, Anna O. Pechenkina, Christopher Prosser, Louis Raes, Kevin Ralston, Miguel R. Ramos, Arne Roets, Jonathan Rogers, Guido Ropers, Robin Samuel, Gregor Sand, Ariela Schachter, Merlin Schaeffer, David Schieferdecker, Elmar Schlueter, Regine Schmidt, Katja M. Schmidt, Alexander Schmidt-Catran, Claudia Schmiedeberg, Jürgen Schneider, Martijn Schoonvelde, Julia Schulte-Cloos, Sandy Schumann, Reinhard Schunck, Jürgen Schupp, Julian Seuring, Henning Silber, Willem Sleegers, Nico Sonntag, Alexander Staudt, Nadia Steiber, Nils Steiner, Sebastian Sternberg, Dieter Stiers, Dragana Stojmenovska, Nora Storz, Erich Striessnig, Anne-Kathrin Stroppe, Janna Teltemann, Andrey Tibajev, Brian Tung, Giacomo Vagni, Jasper Van Assche, Meta van der Linden, Jolanda van der Noll, Arno Van Hootegem, Stefan Vogtenhuber, Bogdan Voicu, Fieke Wagemans, Nadja Wehl, Hannah Werner, Brenton M. Wiernik, Fabian Winter, Christof Wolf, Yuki Yamada, Nan Zhang, Conrad Ziller, Stefan Zins, and Tomasz Żółtak. 2022. "Observing many researchers using the same data and hypothesis reveals a hidden universe of uncertainty." *Proceedings of the National Academy of Sciences*, 119(44): e2203150119.

Brodeur, Abel, Derek Mikola, Nikolai Cook, et al. 2024. "Mass Reproducibility and Replicability: A New Hope." IZA - Institute of Labor Economics Discussion Paper Series 16912, Bonn, Germany.

Brodeur, Abel, Mathias Lé, Marc Sangnier, and Yanos Zylberberg. 2016. "Star Wars: The Empirics Strike Back." *American Economic Journal: Applied Economics*, 8(1): 1–32.

Brodeur, Abel, Nikolai Cook, and Anthony Heyes. 2020. "Methods Matter: p-Hacking and Publication Bias in Causal Analysis in Economics." *American Economic Review*, 110(11): 3634–60.

Callaway, Brantly, and Pedro HC Sant'Anna. 2021. "Difference-in-differences with Multiple Time Periods." *Journal of Econometrics*, 225(2): 200–230.

Card, David, Stefano DellaVigna, Patricia Funk, and Nagore Iriberri. 2022. "Gender differences in Peer Recognition by Economists." *Econometrica*, 90(5): 1937–1971.

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

Coqueret, Guillaume. 2024. "Forking paths in financial economics." Emlyon Business School Working Paper, Ecully, France.

Coqueret, Guillaume, and Christophe Pérignon. 2026. "Persistent Anomalies and Non-standard Sharpe Ratios." HEC Paris Research Paper FIN-2025-1578, Paris, France.

Cowles, Alfred. 1933. "Can Stock Market Forecasters Forecast?" *Econometrica*, 1(3): 309.

Ferman, Bruno, and Lucas Finamor. 2025. "There must be an error here! Experimental evidence on coding errors' biases." Sao Paulo School of Economics Working Paper, Sao Paulo, Brazil.

Fox, John, and Sanford Weisberg. 2019. *An R Companion to Applied Regression.* . Third ed., Thousand Oaks CA:Sage.

Franco, Annie, Neil Malhotra, and Gabor Simonovits. 2014. "Publication bias in the social sciences: Unlocking the file drawer." *Science*, 345(6203): 1502–1505.

Frankel, Alexander, and Maximilian Kasy. 2022. "Which Findings Should Be Published?" *American Economic Journal: Microeconomics*, 14(1): 1–38.

Gelman, Andrew, and Eric Loken. 2014. "The Statistical Crisis in Science." *American Scientist*, 102(6): 460–465.

Giuntella, Osea, and Jakub Lonsky. 2020. "The effects of DACA on health insurance, access to care, and health outcomes." *Journal of Health Economics*, 72: 102320.

Gould, Elliot, Hannah S. Fraser, Timothy H. Parker, Shinichi Nakagawa, Simon C. Griffith, Peter A. Vesk, Fiona Fidler, Daniel G. Hamilton, Robin N. Abbey-Lee, Jessica K. Abbott, Luis A. Aguirre, Carles Alcaraz, Irith Aloni, Drew Altschul, Kunal Arekar, Jeff W. Atkins, Joe Atkinson, Christopher M. Baker, Meghan Barrett, Kristian Bell, Suleiman Kehinde Bello, Iván Beltrán, Bernd J. Berauer, Michael Grant Bertram, Peter D. Billman, Charlie K. Blake, Shannon Blake, Louis Bliard, Andrea Bonisoli-Alquati, Timothée Bonnet, Camille Nina Marion Bordes, Aneesh P. H. Bose, Thomas Botterill-James, Melissa Anna Boyd, Sarah A. Boyle, Tom Bradfer-Lawrence, Jennifer Bradham, Jack A. Brand, Martin I. Brengdahl, Martin Bulla, Luc Bussière, Ettore Camerlenghi, Sara E. Campbell, Leonardo L. F. Campos, Anthony Caravaggi, Pedro Cardoso, Charles J. W. Carroll, Therese A. Catanach, Xuan Chen, Heung Ying Janet Chik, Emily Sarah Choy, Alec Philip Christie, Angela Chuang, Amanda J. Chunco, Bethany L. Clark, Andrea Contina, Garth A. Covernton, Murray P. Cox, Kimberly A. Cressman, Marco Crotti, Connor Davidson Crouch, Pietro B. D'Amelio, Alexandra Allison de Sousa, Timm Fabian Döbert, Ralph Dobler, Adam J. Dobson, Tim S. Doherty, Szymon Marian Drobniak, Alexandra Grace Duffy, Alison B. Duncan, Robert P. Dunn, Jamie Dunning, Trishna Dutta, Luke Eberhart-Hertel, Jared Alan Elmore, Mahmoud Medhat Elsherif, Holly M. English, David C. Ensminger, Ulrich Rainer Ernst, Stephen M. Ferguson, Esteban Fernandez-Juricic, Thalita Ferreira-Arruda, John Fieberg, Elizabeth A. Finch, Evan A. Fiorenza, David N. Fisher, Amélie Fontaine, Wolfgang Forstmeier, Yoan Fourcade, Graham S. Frank, Cathryn A. Freund, Eduardo Fuentes-Lillo, Sara L. Gandy, Dustin G. Gannon, Ana I. García-Cervigón, Alexis C. Garret-

son, Xuezhen Ge, William L. Geary, Charly Géron, Marc Gilles, Antje Girndt, Daniel Gliksman, Harrison B. Goldspiel, Dylan G. E. Gomes, Megan Kate Good, Sarah C. Goslee, J. Stephen Gosnell, Eliza M. Grames, Paolo Gratton, Nicholas M. Grebe, Skye M. Greenler, Maaike Griffioen, Daniel M. Griffith, Frances J. Griffith, Jake J. Grossman, Ali Güncan, Stef Haesen, James G. Hagan, Heather A. Hager, Jonathan Philo Harris, Natasha Dean Harrison, Sarah Syedia Hasnain, Justin Chase Havird, Andrew J. Heaton, María Laura Herrera-Chaustre, Tanner J. Howard, Bin-Yan Hsu, Fabiola Iannarilli, Esperanza C. Iranzo, Erik N. K. Iverson, Saheed Olaide Jimoh, Douglas H. Johnson, Martin Johnsson, Jesse Jorna, Tommaso Jucker, Martin Jung, Ineta Kačergytė, Oliver Kaltz, Alison Ke, Clint D. Kelly, Katharine Keogan, Friedrich Wolfgang Keppeler, Alexander K. Killion, Dongmin Kim, David P. Kochan, Peter Korsten, Shan Kothari, Jonas Kuppler, Jillian M. Kusch, Malgorzata Lagisz, Kristen Marianne Lalla, Daniel J. Larkin, Courtney L. Larson, Katherine S. Lauck, M. Elise Lauterbur, Alan Law, Don-Jean Léandri-Breton, Jonas J. Lembrechts, Kiara L'Herpiniere, Eva J. P. Lievens, Daniela Oliveira de Lima, Shane Lindsay, Martin Luquet, Ross MacLeod, Kirsty H. Macphie, Kit Magellan, Magdalena M. Mair, Lisa E. Malm, Stefano Mammola, Caitlin P. Mandeville, Michael Manhart, Laura Milena Manrique-Garzon, Elina Mäntylä, Philippe Marchand, Benjamin Michael Marshall, Charles A. Martin, Dominic Andreas Martin, Jake Mitchell Martin, April Robin Martinig, Erin S. McCallum, Mark McCauley, Sabrina M. McNew, Scott J. Meiners, Thomas Merkling, Marcus Michelangeli, Maria Moiron, Bruno Moreira, Jennifer Mortensen, Benjamin Mos, Taofeek Olatunbosun Muraina, Penelope Wrenn Murphy, Luca Nelli, Petri Niemelä, Josh Nightingale, Gustav Nilsonne, Sergio Nolazco, Sabine S. Nooten, Jessie Lanterman Novotny, Agnes Birgitta Olin, Chris L. Organ, Kate L.

Ostevik, Facundo Xavier Palacio, Matthieu Paquet, Darren James Parker, David J. Pascall, Valerie J. Pasquarella, John Harold Paterson, Ana Payo-Payo, Karen Marie Pedersen, Grégoire Perez, Kayla I. Perry, Patrice Pottier, Michael J. Proulx, Raphaël Proulx, Jessica L. Pruett, Veronarindra Ramananjato, Finaritra Tolotra Randimbiarison, Onja H. Razafindratsima, Diana J. Rennison, Federico Riva, Sepand Riyahi, Michael James Roast, Felipe Pereira Rocha, Dominique G. Roche, Cristian Román-Palacios, Michael S. Rosenberg, Jessica Ross, Freya E. Rowland, Deusdedith Rugemalila, Avery L. Russell, Suvi Ruuskanen, Patrick Saccone, Asaf Sadeh, Stephen M. Salazar, Kris Sales, Pablo Salmón, Alfredo Sánchez-Tójar, Leticia Pereira Santos, Francesca Santostefano, Hayden T. Schilling, Marcus Schmidt, Tim Schmoll, Adam C. Schneider, Allie E. Schrock, Julia Schroeder, Nicolas Schtickzelle, Nick L. Schultz, Drew A. Scott, Michael Peter Scroggie, Julie Teresa Shapiro, Nitika Sharma, Caroline L. Shearer, Diego Simón, Michael I. Sitvarin, Fabrício Luiz Skupien, Heather Lea Slinn, Grania Polly Smith, Jeremy A. Smith, Rahel Sollmann, Kaitlin Stack Whitney, Shannon Michael Still, Erica F. Stuber, Guy F. Sutton, Ben Swallow, Conor Claverie Taff, Elina Takola, Andrew J. Tanentzap, Rocío Tarjuelo, Richard J. Telford, Christopher J. Thawley, Hugo Thierry, Jacqueline Thomson, Svenja Tidau, Emily M. Tompkins, Claire Marie Tortorelli, Andrew Trlica, Biz R. Turnell, Lara Urban, Stijn Van de Vondel, Jessica Eva Megan van der Wal, Jens Van Eeckhoven, Francis van Oordt, K. Michelle Vanderwel, Mark C. Vanderwel, Karen J. Vanderwolf, Juliana Vélez, Diana Carolina Vergara-Florez, Brian C. Verrelli, Marcus Vinícius Vieira, Nora Villamil, Valerio Vitali, Julien Vollering, Jeffrey Walker, Xanthe J. Walker, Jonathan A. Walter, Pawel Waryszak, Ryan J. Weaver, Ronja E. M. Wedegärtner, Daniel L. Weller, and Shannon Whelan. 2025. "Same data, different analysts: variation in effect sizes due

to analytical decisions in ecology and evolutionary biology." *BMC Biology*, 23(1): 35.

Grundl, Serafin. 2026. "Claude Code as an Empirical Economist: Like Humans but Without the Tails." Board of Governors of the Federal Reserve System Working Paper, Washington, DC.

Heckman, James J., and Sidharth Moktan. 2020. "Publishing and Promotion in Economics: The Tyranny of the Top Five." *Journal of Economic Literature*, 58(2): 419–70.

Herbert, Sylvérie, Hautahi Kingi, Flavio Stanchi, and Lars Vilhuber. 2024. "Reproduce to validate: A comprehensive study on the reproducibility of economics research." *Canadian Journal of Economics/Revue canadienne d'économique*, 57(3): 961–988.

Heyman, Tom, and Wolf Vanpaemel. 2022. "Multiverse analyses in the classroom." *MetaPsychology*, 6.

Höffler, Jan H. 2017. "Replication and Economics Journal Policies." *American Economic Review*, 107(5): 52–55.

Holzmeister, Felix, Magnus Johannesson, Robert Böhm, Anna Dreber, Jürgen Huber, and Michael Kirchler. 2024. "Heterogeneity in effect size estimates." *Proceedings of the National Academy of Sciences*, 121(32): e2403490121.

Hoogeveen, Suzanne, Alexandra Sarafoglou, Balazs Aczel, Yonathan Aditya, Alexandra J Alayan, Peter J Allen, Sacha Altay, Shilaan Alzahawi, Yulmaida Amir, Francis-Vincent Anthony, et al. 2023. "A Many-analysts Approach to the Relation between Religiosity and Well-being." *Religion, Brain & Behavior*, 13(3): 237–283.

Huntington-Klein, Nick. 2021. "vtable: Variable Table for Variable Documentation." R package version 1.4.6.

Huntington-Klein, Nick, Andreu Arenas, Emily Beam, Marco Bertoni, Jeffrey R Bloem, Pralhad Burli, Naibin Chen, Paul Grieco, Godwin Ekpe, Todd Pugatch,

Martin Saavedra, and Yaniv Stopnitzky. 2021. "The Influence of Hidden Researcher Decisions in Applied Microeconomics." *Economic Inquiry*, 59(3): 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." *The Economic Journal*, 134(662): 2439–2469.

Jones, Richard C. 2020. "A Time-Space Stream of DACA Benefits and Barriers Gleaned From the American Community Survey." *Hispanic Journal of Behavioral Sciences*, 42(2): 143–164.

Kuhn, Thomas S. 1962. *The Structure of Scientific Revolutions.* Chicago, IL.:University of Chicago Press.

Lamont, Michèle. 2009. *How Professors Think: Inside the Curious World of Academic Judgment.* Cambridge, MA:Harvard University Press.

Leamer, Edward E. 1982. "Sets of Posterior Means with Bounded Variance Priors." *Econometrica*, 50(3): 725.

Leamer, Edward E. 2017. "Specification Problems in Econometrics." *The New Palgrave Dictionary of Economics*, 1–7. London:Palgrave Macmillan UK.

Lo, Andrew W., and A. Craig MacKinlay. 1990. "Data-Snooping Biases in Tests of Financial Asset Pricing Models." *Review of Financial Studies*, 3(3): 431–467.

Lundberg, Shelly, and Jenna Stearns. 2019. "Women in Economics: Stalled Progress." *Journal of Economic Perspectives*, 33(1): 3–22.

Magnus, Jan R., and Mary S. Morgan. 1997. "Design of the Experiment." *Journal of Applied Econometrics*, 12(5): 459–465.

McCullough, B. D, and H. D Vinod. 2003. "Verifying the Solution from a Nonlinear Solver: A Case Study." *American Economic Review*, 93(3): 873–892.

McCully, Brett. 2026. "Researcher-Induced Estimation Uncertainty at Scale Using Agentic AI." University of Turin Working Paper, Turin, Italy.

Menkveld, Albert J., Anna Dreber, Felix Holzmeister, Juergen Huber, Magnus Johannesson, Michael Kirchler, Sebastian Neusüss, Michael Razen, Utz Weitzel, et al. 2024. "Nonstandard Errors." *The Journal of Finance*, 79(3): 2339–2390.

Merton, Robert K. 1973. *The Sociology of Science: Theorectical and Empirical Investigations.* Chicago, IL:University of Chicago Press.

Naguib, Costanza. 2026. "P-hacking and Significance Stars." *The Economic Journal*, 136(675): 1140–1154.

Nosek, Brian A., Tom E. Hardwicke, Hannah Moshontz, Aurélien Allard, Katherine S. Corker, Anna Dreber, Fiona Fidler, Joe Hilgard, Melissa Kline Struhl, Michèle B. Nuijten, Julia M. Rohrer, Felipe Romero, Anne M. Scheel, Laura D. Scherer, Felix D. Schönbrodt, and Simine Vazire. 2022. "Replicability, Robustness, and Reproducibility in Psychological Science." *Annual Review of Psychology*, 73(1): 719–748.

Open Science Collaboration. 2015. "Estimating the reproducibility of psychological science." *Science*, 349(6251): aac4716.

Ortloff, Anna-Marie, Matthias Fassl, Alexander Ponticello, Florin Martius, Anne Mertens, Katharina Krombholz, and Matthew Smith. 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." *CHI' 23: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 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, Yasser Albogami, Mitchell Conover, Juan M Banda, William A Baumgartner Jr, Clair Blacketer, Priyamvada Desai, Scott L DuVall, Stephen Fortin, James P Gilbert, et al. 2023. "Reproducible variability: assessing investigator discordance across 9 research teams attempting to reproduce the same observational study." *Journal of the American Medical Informatics Association*, 30(5): 859–868.

Pörtner, Claus C., and Nick Huntington-Klein. 2022. "Many Economists." https://osf.io/cj9yx.

Pötscher, B.M. 1991. "Effects of Model Selection on Inference." *Econometric Theory*, 7(2): 163–185.

Pütz, Peter, and Stephan B. Bruns. 2020. "The (Non-)Significance of Reporting Errors in Economics: Evidence from Three Top Journals." *Journal of Economic Surveys*, 35(1): 348–373.

Rohrer, Julia M, Jessica Hullman, and Andrew Gelman. 2026. "What's a multiverse good for anyway?" Wilhelm Wundt Institute for Psychology, Leipzig University Working Paper, Leipzig, Germany.

Ruggles, Steven, Sarah Flood, Matthew Sobek, Daniel Backman, Annie Chen, Grace Cooper, Stephanie Richards, Renae Rodgers, and Megan Schouweiler. 2024. "IPUMS USA: Version 15.0 [dataset]. Minneapolis, MN: IPUMS."

Sala-i Martin, Xavier X. 1997. "I Just Ran Four Million Regressions." National Bureau of Economic Research NBER Working Paper 6252.

Sarafoglou, Alexandra, Suzanne Hoogeveen, Don van den Bergh, Balazs Aczel, Casper J. Albers, Tim Althoff, Rotem Botvinik-Nezer, Niko A. Busch, An-

drea M. Cataldo, Berna Devezer, Noah N. N. van Dongen, Anna Dreber, Eiko I. Fried, Rink Hoekstra, Sabine Hoffman, Felix Holzmeister, Jürgen Huber, Nick Huntington-Klein, John Ioannidis, Magnus Johannesson, Michael Kirchler, Eric Loken, Jan-Francois Mangin, Dora Matzke, Albert J. Menkveld, Gustav Nilsonne, Don van Ravenzwaaij, Martin Schweinsberg, Hannah Schulz-Kuempel, David R. Shanks, Daniel J. Simons, Barbara A. Spellman, Andrea H. Stoevenbelt, Barnabas Szaszi, Darinka Trübutschek, Francis Tuerlinckx, Eric L. Uhlmann, Wolf Vanpaemel, Jelte Wicherts, and Eric-Jan Wagenmakers. 2024. "Subjective Evidence Evaluation Survey for Many-Analysts Studies." *Royal Society Open Science*, 11(7): 240125.

Schweinsberg, Martin, Michael Feldman, Nicola Staub, Olmo R van den Akker, Robbie CM van Aert, Marcel ALM Van Assen, Yang Liu, Tim Althoff, Jeffrey Heer, Alex Kale, et al. 2021. "Same Data, Different Conclusions: Radical Dispersion in Empirical Results when Independent Analysts Operationalize and Test the Same Hypothesis." *Organizational Behavior and Human Decision Processes*, 165: 228–249.

Shapin, Steven. 1994. *A Social History of Truth: Civility and Science in Seventeenth-Centry England.* Chicago, IL.:University of Chicago Press.

Silberzahn, Raphael, Eric L Uhlmann, Daniel P Martin, Pasquale Anselmi, Frederik Aust, Eli Awtrey, Štěpán Bahník, Feng Bai, Colin Bannard, Evelina Bonnier, et al. 2018. "Many Analysts, One Data Set: Making Transparent how Variations in Analytic Choices Affect Results." *Advances in Methods and Practices in Psychological Science*, 1(3): 337–356.

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

Singhal, Karan, and Eva Sierminska. 2025. "Who Studies What? Country of Origin, Gender, and Field Specialization among Economics PhDs." IZA – Institute of Labor Economics Discussion Paper Series 18348, Bonn, Germany.

Soebhag, Amar, Bart Van Vliet, and Patrick Verwijmeren. 2024. "Non-standard errors in asset pricing: Mind your sorts." *Journal of Empirical Finance*, 78: 101517.

Song, Ziyu, Shan Wu, and Yuqin Zhou. 2025. "Code like an economist: Analyzing LLMs' code generation capabilities in economics and finance." *Finance Research Letters*, 86: 108540.

Stansbury, Anna, and Robert Schultz. 2023. "The Economics Profession's Socioeconomic Diversity Problem." *Journal of Economic Perspectives*, 37(4): 207–230.

StataCorp. 2024. "Difference-in-differences (DID) and DDD models."

Steegen, Sara, Francis Tuerlinckx, Andrew Gelman, and Wolf Vanpaemel. 2016. "Increasing Transparency Through a Multiverse Analysis." *Perspectives on Psychological Science*, 11(5): 702–712.

Stevenson, Megan T, and Joshua B Fischman. Forthcoming. "Humans in the Loop: The Next Frontier in the Credibility Revolution." *Journal of Economic Literature*.

Sulik, Justin, Nakwon Rim, Elizabeth Pontikes, James Evans, and Gary Lupyan. 2023. "Why Do Scientists Disagree?" PsyArXiv.

Urban Institute. 2022. "State Immigration Policy Resource."

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)."

Walter, Dominik, Rüdiger Weber, and Patrick Weiss. 2024. "Methodological uncertainty in portfolio sorts." University of Konstanz Working Paper.

Weill, Joakim A., Matthieu Stigler, Olivier Deschenes, and Michael R. Springborn. 2025. "Researchers' degrees of flexibility: Revisiting COVID-19 policy evaluations." *Economic Inquiry*, 63(2): 441–462.

White, Halbert. 2000. "A Reality Check for Data Snooping." *Econometrica*, 68(5): 1097–1126.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, Alex Hayes, Lionel Henry, Jim Hester, Max Kuhn, Thomas Lin Pedersen, Evan Miller, Stephan Milton Bache, Kirill Müller, Jeroen Ooms, David Robinson, Dana Paige Seidel, Vitalie Spinu, Kohske Takahashi, Davis Vaughan, Claus Wilke, Kara Woo, and Hiroaki Yutani. 2019. "Welcome to the tidyverse." *Journal of Open Source Software*, 4(43): 1686.

Zuckerman, Harriet. 1988. "Handbook of Sociology." , ed. Neil J. Smelser, Chapter The Sociology of Science, 511–574. Thousand Oaks, CA, US:Sage Publications, Inc.

Appendix A: Project Contributors

- Yubraj Acharya, Department of Health Policy and Administration, The Pennsylvania State University
- Matus Adamkovic, Centre of Social and Psychological Sciences, Slovak Academy of Sciences
- Joop Adema, University of Innsbruck
- Lameck Ondieki Agasa, Department of Mathematics and Actuarial Sciences, Kisii University
- Imtiaz Ahmad, Department of Economics, National University of Sciences and Technology
- Mevlude Akbulut-Yuksel, Department of Economics, Dalhousie University; IZA
- Martin Eckhoff Andresen, Department of Economics, University of Oslo
- David Angenendt, TUM School of Management, Technical University of Munich
- José-Ignacio Antón, Department of Applied Economics, University of Salamanca
- Andreu Arenas, Economics Department and Institut d'Economia de Barcelona, University of Barcelona
- Erkmen Giray Aslim, Department of Economics, University of Vermont
- Stanislav Avdeev, Amsterdam School of Economics, University of Amsterdam
- Andrew Bacher-Hicks, Arnold Ventures
- Bradley J. Baker, Department of Sport, Tourism and Hospitality Management, Temple University
- Imesh Nuwan Bandara, Independent
- Avijit Bansal, Finance & Control Area, Indian Institute of Management Calcutta
- David Bartram, Department of Sociology, University of Leicester
- Katarzyna Bech-Wysocka, Institute of Econometrics, Warsaw School of Economics
- Christopher Troy Bennett, RTI International

- Andu Berha, University of Alberta
- Inés Berniell, Department of Economics, Universidad Nacional de La Plata
- Moiz Bhai, Department of Accounting, Economics, and Finance, University of Arkansas at Little Rock
- Shreya Bhattacharya, Centre for Advanced Research and Learning (CAFRAL)
- Jeffrey R. Bloem, International Food Policy Research Institute
- Margaret E Brehm, Department of Economics, Oberlin College
- Martín Brun, Finnish Centre of Excellence in Tax Systems Research (FIT), Tampere University
- Florent Buisson, Independent
- Pralhad Burli, Independent
- Andrew M. Camp, Annenberg Institute, Brown University
- Nicola Cerutti, Bank of England
- Weiwei Chen, Department of Economics, Finance and Quantitative Analysis, Kennesaw State University
- Jeffrey Clement, School of Business and Economics, Augsburg University
- Matthew Collins, J.E. Cairnes School of Business and Economics, University of Galway
- Lee Crawfurd, Center for Global Development
- John Cullinan, School of Business & Economics, University of Galway
- Lachlan Deer, Department of Management and Marketing, University of Melbourne
- Reid Dorsey-Palmateer, Department of Economics, Western Washington University
- Nicolas J. Duquette, Sol Price School of Public Policy, University of Southern California
- Diego Marino Fages, Department of Economics, Durham University
- Grace Falken, Center for Education Data and Research, University of Washington
- Christine Farquharson, Institute for Fiscal Studies
- Jan Feld, School of Economics and Finance, Victoria University of Wellington

- Yevgeniy Feyman, Department of Health and Human Services
- Nathan Fiala, Agricultural and Resource Economics, University of Connecticut
- Anne Fitzpatrick, Department of Agricultural, Environmental, and Development Economics, The Ohio State University
- Andrey Fradkin, Marketing, Boston University, Questrom School of Business
- Evaewero French, School of Public Policy, Oregon State University
- Wei Fu, Health Management and Systems Sciences Department, University of Louisville
- Luca Fumarco, Department of Economics, Masaryk University
- Sebastian Gallegos, Business School, Universidad Adolfo Ibanez
- Julio Galárraga, Department of Economics, Universidad de las Américas Ecuador
- Aaron M. Gamino, Department of Economics and Finance, Middle Tennessee State University
- Romain Gauriot, Department of Economics, Deakin University
- Victor Gay, Department of Social and Behavioral Sciences, Toulouse School of Economics
- Savas Gayaker, Department of Econometrics, Ankara Hacı Bayram Veli University
- Jules Gazeaud, CNRS
- Alexandra de Gendre, Department of Economics, The University of Melbourne
- Gregory Gilpin, Department of Agricultural Economics and Economics, Montana State University
- Daniele Girardi, Department of Political Economy, King's College London
- Dan Goldhaber, Center for Education Data and Research (University of Washington), National Center for Analysis of Longitudinal Data in Education Research (AIR), University of Washington, American Institutes for Research
- Mark N. Harris, School of Accounting, Economics and Finance, Curtin University
- Blake H. Heller, Hobby School of Public Affairs, University of Houston
- Daniel J. Henderson, Department of Economics, Finance and Legal Studies, University

of Alabama

- Arne Henningsen, Department of Food and Resource Economics, University of Copenhagen
- Junita Henry, Department of Global Health, Harvard University
- Clément Herman, Department of Economics, Princeton University
- Øystein Hernæs, The Ragnar Frisch Centre for Economic Research
- Andrew J. Hill, Department of Agricultural Economics and Economics, Montana State University
- Felix Holzmeister, Department of Economics, University of Innsbruck
- Martijn Huysmans, School of Economics, Utrecht University
- M. Saad Imtiaz, Department of Humanities and Social Sciences, Lahore University of Management Sciences
- Anil K. Jain, International Finance, Federal Reserve Board of Governors
- Niklas Jakobsson, Karlstad Business School, Karlstad University
- José Kaire, School of Politics and Global Studies, Arizona State University
- Kalyan Kumar Kameshwara, Department of Education, University of Bath
- Daniel H Karney, Department of Economics, Ohio University
- Sie Won Kim, Department of Economics, Texas Tech University
- Valentin Klotzbücher, Department of Economics, University of Freiburg
- Christoph Kronenberg, CINCH, University Duisburg-Essen
- Daniel LaFave, Department of Economics, Colby College
- David Lang, Department of Political Science, Stanford University
- Ryan Lee, College of Business, University of La Verne
- Maxime Liégey, FSEG, Strasbourg University
- Dede Long, Department of Humanities, Social Sciences, and the Arts, Harvey Mudd College

- Jan Marcus, School of Business and Economics, Freie Universität Berlin
- Gabriele Mari, Department of Public Administration and Sociology, Erasmus University Rotterdam
- Ian McCarthy, Department of Economics, Emory University
- Laura Meinzen-Dick, Department of Economics, Villanova University
- Erik Merkus, Independent
- Klaus M. Miller, Department of Marketing, HEC Paris
- Lukas Mogge, Economic Policy Lab Climate, Migration and Development, RWI – Leibniz Institute for Economic Research
- S. M. Woahid Murad, School of Accounting, Economics and Finance, Curtin University
- Rafiuddin Najam, Department of Public Policy, University of North Carolina at Chapel Hill
- Elias Naumann, GESIS Leibniz Institute for the Social Sciences
- Job Nda Nmadu, Department of Agricultural Economics and Farm Management, Federal University of Technology, Minna, Nigeria
- Gorkem Turgut Ozer, Paul College of Business and Economics, University of New Hampshire
- Jayash Paudel, Department of Economics, University of Oklahoma
- Filippos Petroulakis, Bank of Greece and University of Piraeus
- Christian Peukert, Department of Strategy, Globalization and Society, University of Lausanne, Faculty of Business and Economics (HEC)
- Visa Pitkänen, Finnish Competition and Consumer Authority
- Simon Porcher, Department of Management, Université Paris Panthéon-Assas
- Manab Prakash, Central Department of Economics, Tribhuvan University
- Andrew Adrian Yu Pua, School of Economics, De La Salle University - Manila
- Todd Pugatch, Department of Economics, University at Buffalo

- Daniel S. Putman, Center for Social Norms and Behavioral Dynamics, University of Pennsylvania
- Veeshan Rayamajhee, Department of Economics, Applied Statistics, and International Business, New Mexico State University
- Obeid Ur Rehman, Department of Economics, Toronto Metropolitan University
- Maira Emy Reimão, Department of Economics, Villanova University
- Anna Reuter, Heidelberg Institute of Global Health, Heidelberg University
- Michael David Ricks, Department of Economics, University of Nebraska - Lincoln
- Fernando Rios-Avila, Levy Economics Institute
- Julian Roeckert, Policy Lab - Climate Change, Development and Migration, RWI - Leibniz Institute for Economic Research
- Ivan Ropovik, Faculty of Education, Charles University
- Jayjit Roy, Department of Economics, Appalachian State University
- Nicolas Salamanca, Melbourne Institute, Applied Economic & Social Research, The University of Melbourne
- Margaret Samahita, School of Economics, University College Dublin
- Aparna Samudra, Department of Economics, RTM Nagpur University
- Vassiki Sanogo, Contract Worker, Pharmacoevidence, Otsuka Pharmaceutical Development & Commercialization, Inc.
- Orkhan Sariyev, Department of Rural Development Theory and Policy, University of Hohenheim
- Henning Schaak, Department of Economics and Social Sciences, University of Natural Resources and Life Sciences, Vienna
- Joel E. Segel, Department of Health Policy and Administration, Penn State University
- Hans Henrik Sievertsen, School of Economics, VIVE and University of Bristol
- Mike Smet, Department of Work and Organisation Studies, KU Leuven

- Brock Smith, Department of Agricultural Economics and Economics, Montana State University
- Lucy C. Sorensen, Department of Public Administration and Policy, University at Albany, SUNY
- Lisa Spantig, School of Business and Economics, RWTH Aachen University
- Krzysztof Szczygielski, Faculty of Economics Sciences, University of Warsaw
- Anirudh Tagat, Department of Economics, Monk Prayogshala
- Huseyin Tastan, Department of Economics, Yildiz Technical University
- Martin Trombetta, Área de Economía, Universidad Nacional de General Sarmiento
- Madhavi Venkatesan, Department of Economics, Northeastern University
- Antoine Vernet, The Bartlett School of Sustainable Construction, University College London
- Eden Volkov, Office of the Assistant Secretary for Policy and Evaluation, Department of Health and Human Services
- Gary A. Wagner, Department of Economics & Finance, University of Louisiana at Lafayette
- Yue Wang, Department of Economics and Finance, University of Canterbury
- Zachary Ward, Department of Economics, Baylor University
- Tom Waters, Institute for Fiscal Studies
- Ellerie Weber, Department of Population Health Science and Policy, Icahn School of Medicine at Mount Sinai
- Stephen E Weinberg, Health Research, Inc
- Kristina S. Weißmüller, Department of Political Science and Public Administration, Vrije Universiteit Amsterdam
- Christian Westheide, Stockholm Business School, Stockholm University, and Leibniz Institute for Financial Research SAFE

- Kevin M. Williams, Department of Economics, Occidental College
- Xiaoyang Ye, Annenberg Institute for School Reform, Brown University
- Jisang Yu, Department of Agricultural Economics, Kansas State University
- Muhammad Umer Zahid, Agricultural and Resource Economics, University of Connecticut
- Raffaele Zanoli, Department of Agricultural, Food & Environmental Sciences, Università Politecnica delle Marche (Marche Polytechnic University)

Appendix B: 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 immigrants 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).[32]

- 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

[32]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 immigration 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.

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.

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 indicating 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.[33]

Appendix C: Peer Feedback

This section evaluates the impact of peer feedback on the later work performed by a researcher. The structure of peer feedback in this study is that, following each main task, 2/3 of the researchers are randomized into pairs that produce a peer feedback report of the other's work,

[33]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.

Table C.1: Post-Revision Variance in Effect Sizes by Peer Feedback

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

while the remaining 1/3 do not receive or perform peer feedback. Then, researchers have an opportunity to revise their work.

Revision is optional, and relatively few researchers chose to revise their work after receiving peer feedback (26 in Task 1, 29 in Task 2, and 21 in Task 3). As such, we mostly look at the impact of peer feedback 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 feedback process can only capture the first of these mechanisms, and in effect may be closer to comments received, for example, during seminar presentations, which is why we refer in this paper to "peer feedback" instead of "peer review".

In Appendix Table C.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 C.2 there are no statistically significant differences in the variance of sample sizes between the groups receiving or not receiving peer feedback in the follow-up tasks.

Table C.2: Post-Revision Variance in Sample Sizes by Peer Feedback

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

Appendix Figure C.1 shows the distribution of effect sizes estimated by those who did, and did not, engage in peer feedback 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 effects reported in the follow-up task. As is expected given randomization, effect distributions are fairly similar pre-feedback between the feedback and non-feedback groups. No differences emerge between these groups in the follow-up task. Levene test p-values comparing effect size variance of groups receiving or not receiving peer feedback 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 feedback might drive agreement between researchers. Similar results are found when comparing the distributions or variance of analytic, treatment, or control sample sizes between the groups receiving or not receiving peer feedback in the follow-up tasks.

Figure C.2 explores the possibility that peer feedback might not make the group receiving peer feedback 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 receiving feedback (left column), in the follow-up task (middle column)

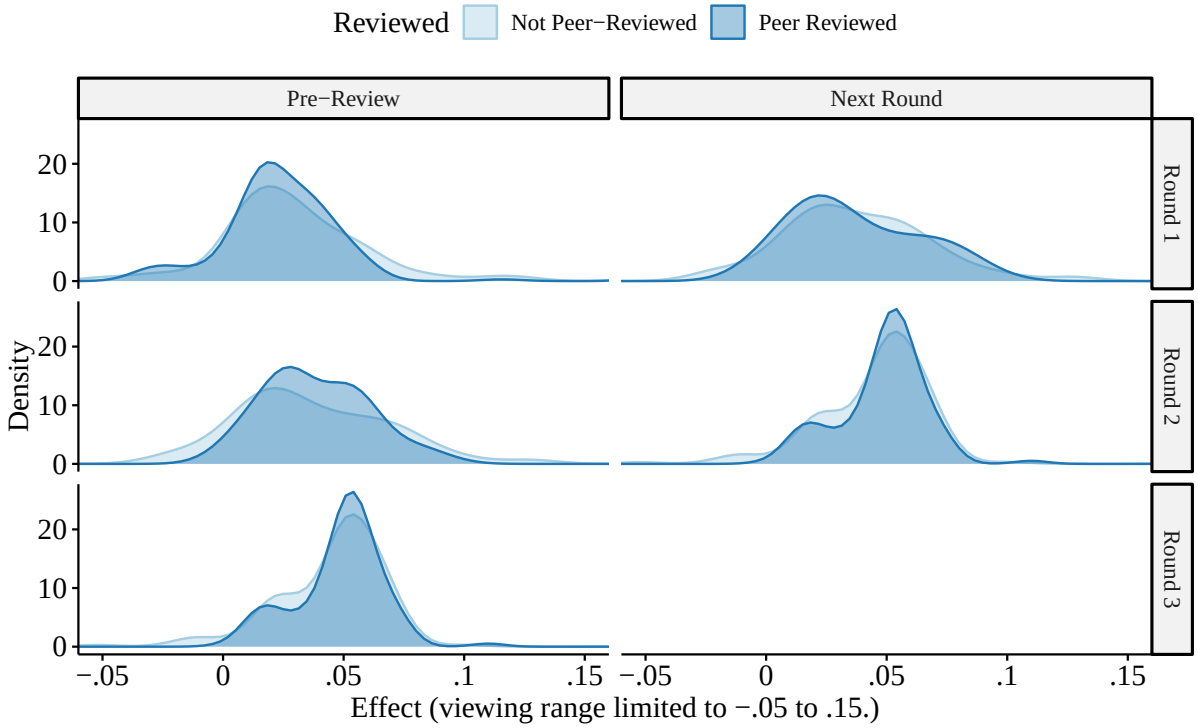


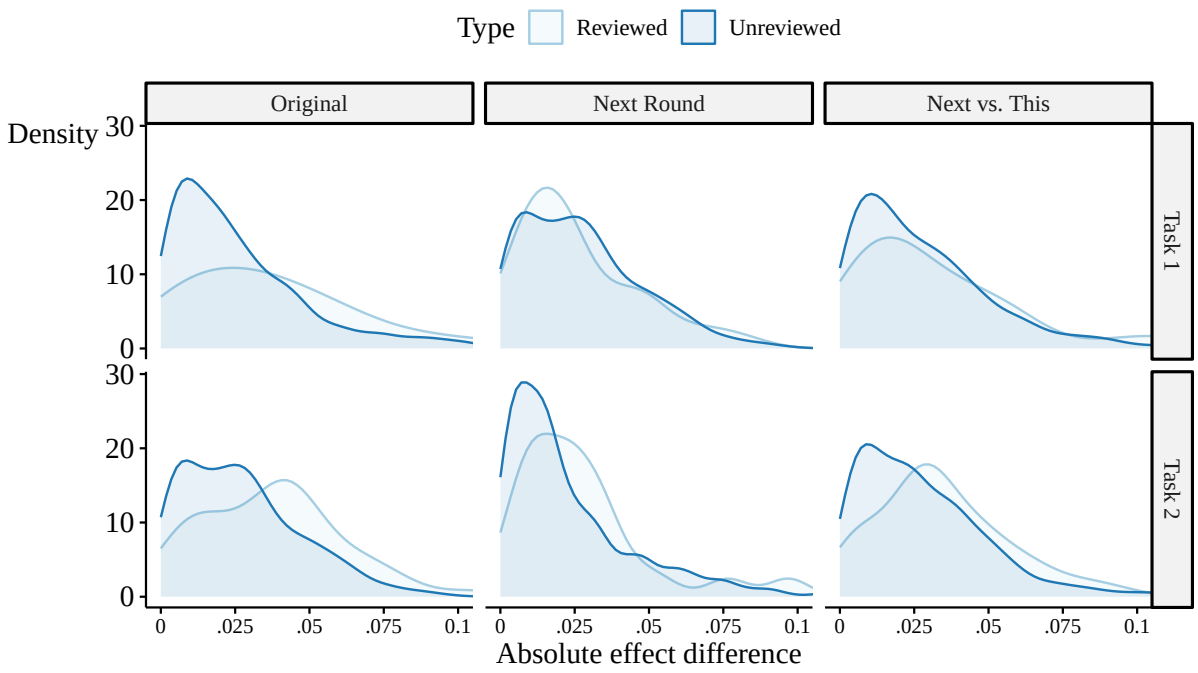
Figure C.1: Distributions of Reported Effect Sizes

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.[34]

In Figure C.2 we see inconsistent evidence in favor of peer feedback. Task 1 feedback pairs became more similar in Task 2, while non-feedback pairs did not change. The change in average absolute effect differences from Task 1 to Task 2 was a statistically significant 0.051 greater for feedback pairs than non-feedback pairs (see Appendix Table C.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 0.029 more for non-feedback pairs than for feedback pairs.

This collection of findings is not consistent strong evidence of peer feedback making a researcher more like their reviewer. When interpreting this result, it is important to keep in mind that the form of peer feedback used in this study is less intensive than is received during a typical publication peer-review process, since authors are not required to satisfy reviewers.

[34]The distributions of absolute differences for researchers not receiving feedback are generated as a null distribution by matching every researcher not receiving feedback to every other researcher not receiving feedback and calculating all absolute differences. This null distribution represents the distribution of absolute differences among people who did not actually experience peer feedback. Notably, each non-reviewer is matched multiple times in this approach, instead of just once for people who actually did receive feedback. However, matching the non-feedback group 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 C.2: Comparisons of Effect Sizes vs. One's Reviewer

Table C.3: Paired Absolute Effect Differences and Peer Feedback

	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 D: Example Submissions

Below are an example submitted piece of research, and an example submitted piece of peer feedback. These are two separate pieces; this particular piece of feedback is not being directed at this particular research. All typos are preserved, but formatting of Stata output has been mildly changed. Both are picked randomly from among the submitted work and peer feedback. All submitted work and peer feedback are available for download with the project data.

D.1: Example Submission of Research

We explore a difference-in-differences estimation strategy to estimate the causal association between the implementation of the DACA program and full-time employment of the affected immigrants. More specifically, following the guidelines to apply to the DACA program, we categorize individuals who were born in Mexico and identify as Hispanics, who were not US citizens, who arrived in the U.S. before the age of 16 and who were younger than 31 in June 2012 as the potential affected group. On the other hand, our control group includes individuals who identify as Hispanics and Mexicans, who are U.S. citizens, who were 31 years old in June 2012 when the DACA instituted. In our empirical design, our main variable interest is the interaction between the affected /DACA group and an indicator for years after the implementation of DACA which corresponds to years 2013-2016. We control for being in DACA group, age and year fixed effects, current state of residence fixed effects, language proficiency, sex and other time-variant state variables such as annual labor force participation and unemployment rate in our estimation framework. We exclusively focus on the sample who are not currently in school in our analysis. We further cluster the standard errors by state of residence.

The main variable of interest in our analysis is full-time employment, which takes a value of 1 if the individual is employed and worked at least 35 hours or more per week. In our difference-in-

differences analysis, we find that the estimated effects of the DACA on full-time employment is not statistically distinguishable from zero. More specifically, The point estimate on the interaction between the DACA group and Post dummy is very close to zero and insignificant. This null effect of the DACA program on full-time employment remains virtually unchanged when we include aforementioned control variables gradually moving from more parsimonious to the more comprehensive specification. This robust finding suggests that the estimated null effect is indeed stable across specifications and are not an artifact of any control variable. Thus, our preferred specification is the more rigorous specification which incorporates the state-level time-variant controls in addition to the state-level fixed effects, age and year fixed effects.

We further the potential pathways to unmask the estimated null effect of the DACA program on immigrants' full-time employment. We first estimate the difference-in-differences for the entire sample regardless of the individuals' school attendance status. This analysis yields a negative and statistically significant point estimate of the DACA program, suggesting that the potentially affected immigrants are 3 percentage points less likely to be fully employed after the implementation of the DACA program. This corresponds to a 5 percent decline in full-time employment among the affected group compared to the mean of 65.3 percent. These contrasting results between the entire sample and the sample exclusively focusing on individuals who are no longer in school postulates that the significantly increased access to the academic opportunities with the DACA program could be one of the potential mediators of the estimated results on full-time employment. Thus, we investigated whether DACA led to changes in schooling decision of the affected individuals as well to better understand the potential drivers behind our results. In this analysis, our outcome of interest is whether an individual is still in school or not. Our investigation shows that the potentially affected immigrants indeed are more likely to be in school after the implementation of the policy suggesting that the DACA program allowed them to pursue further education. We also note that these estimated results on schooling is even stronger when we exclusively focus on the

individuals with non-missing full-time employment indicator, which corresponds to the sample we estimate the full-time employment analysis.

Main Specification Results

```
reg FullTimeWork DACA TreatedGroup sex speakeng lfpr unemp i.age ///
  i.year i.statefip if school==1, cluster(statefip)
Linear regression Number of obs = 204,239
 F(31, 50) = .
 Prob > F = .
 R-squared = 0.0647
Root MSE = .40011
 (Std. Err. adjusted for 51 clusters in statefip)
------------------------------------------------------------------------
              |               Robust
 FullTimeWork | Coef.      Std. Err.  t       P>|t| [95% Conf. Interval]
-------------------------------------------------------------------------
 DACA         |  .0004365  .0036768     0.12  0.906 -.0069486  .0078215
 TreatedGroup |  .004317   .0051256     0.84  0.404 -.0059782  .0146121
 sex          | -.1264222  .0030355   -41.65  0.000 -.1325193 -.1203251
 speakeng     |  .0065437  .0019601     3.34  0.002  .0026067  .0104807
 lfpr         |  .0060043  .0017137     3.50  0.001  .0025622  .0094464
 unemp        | -.0051915  .0010394    -4.99  0.000 -.0072792 -.0031037
```

Entire Sample

```
reg FullTimeWork DACA TreatedGroup sex speakeng lfpr unemp i.age ///
  i.year i.statefip , cluster(statefip)
Linear regression Number of obs = 288,494
 F(31, 50) = .
Prob > F = .
R-squared = 0.1970
Root MSE = .4266
 (Std. Err. adjusted for 51 clusters in statefip)
------------------------------------------------------------------------
             |               Robust
 FullTimeWork | Coef.      Std. Err.  t      P>|t| [95% Conf. Interval]
-------------------------------------------------------------------------
 DACA         | -.0335954  .0038919   -8.63  0.000 -.0414125 -.0257783
 TreatedGroup |  .0698207  .0057753   12.09  0.000  .0582205  .0814208
 sex          | -.1335975  .0032543  -41.05  0.000 -.1401339 -.127061
 speakeng     |  .0107906  .0021813    4.95  0.000  .0064094  .0151718
 lfpr         |  .0083131  .001506     5.52  0.000  .0052882  .0113381
 unemp        | -.0038236  .0009477   -4.03  0.000 -.005727  -.0019202
```

Schooling for Entire Sample

```
reg school DACA TreatedGroup sex speakeng lfpr unemp i.age ///
  i.year i.statefip , cluster(statefip)
Linear regression Number of obs = 594,533
 F(32, 50) = .
 Prob > F = .
 R-squared = 0.4455
```

```
Root MSE = .37211
 (Std. Err. adjusted for 51 clusters in statefip)
------------------------------------------------------------------------------
             |               Robust
 school      | Coef.      Std. Err. t     P>|t| [95% Conf. Interval]
-------------+----------------------------------------------------------------
 DACA        |  .0647602  .005124   12.64 0.000  .0544683  .075052
 TreatedGroup | -.1326744  .0057917 -22.91 0.000 -.1443074 -.1210415
 sex         |  .0496567  .0032621  15.22 0.000  .0431046  .0562087
 speakeng    | -.0015924  .0016217  -0.98 0.331 -.0048496  .0016648
 lfpr        | -.0027813  .0013596  -2.05 0.046 -.0055121 -.0000506
 unemp       | -.0013587  .001655   -0.82 0.416 -.0046829  .0019
```

Schooling for Individuals Only With Non-Missing Full-Time Work Information

```
reg school DACA TreatedGroup sex speakeng lfpr unemp i.age ///
  i.year i.statefip if FullTimeWork!=. , cluster(statefip)
Linear regression Number of obs = 288,494
 F(31, 50) = .
Prob > F = .
R-squared = 0.2251
Root MSE = .40032
 (Std. Err. adjusted for 51 clusters in statefip)
------------------------------------------------------------------------------
             |               Robust
 school      | Coef.      Std. Err.  t     P>|t| [95% Conf. Interval]
-------------+----------------------------------------------------------------
```

```
DACA         |  .0986071 .0060045   16.42 0.000  .0865468  .1106674
TreatedGroup | -.1588601 .0059554  -26.67 0.000 -.1708219 -.1468983
sex          |  .0844663 .0029183   28.94 0.000  .0786046  .090328
speakeng     | -.0050274 .0013929   -3.61 0.001 -.0078252 -.0022297
lfpr         | -.0008572 .0014806   -0.58 0.565 -.0038312  .0021167
unemp        |  .0002485 .0014923    0.17 0.868 -.0027489  .003246
```

D.2: Example Peer Feedback

The submitted analysis seeks to investigate the causal impact of eligibility for the Deferred Action for Childhood Arrivals (DACA) program on the probability to be employed full-time for ethnically Hispanic-Mexican Mexican-born people living in the US. To this end, the author uses survey data from the American Community Survey (ACS) from 2006 to 2016. The author restricts the analyses to individuals with Hispanic ethnicity born in Mexico, aged between 16 and 31, currently enrolled in school or with a high school degree, who are not citizens and did not receive immigration papers.

The submitted report in general gives a good overview of what was done and why. The analysis is quite clear on the translation of the DACA eligibility criteria into the variable definitions (and the limitations due to data availability). However, there are some points which remain open/unclear and warrant some attention:

(1) Definition of the control group

The analysis is less clear on the definition of the control group. Based on the description of the sample restrictions, I would have expected the following control group:

- All individuals with Hispanic ethnicity born in Mexico, aged between 16 and 31, currently enrolled in school or with a high school degree, who are not citizens and did not receive immigration papers AND
- who came to the US after reaching their 16th birthday OR
- who came to the US after June 15, 2007

However, the submitted survey response reads to me as if the control group was defined as follows:

- All individuals with Hispanic ethnicity born in Mexico, aged between 16 and 31 AND
- currently not enrolled in school or without a high school degree AND
- are citizens AND
- who came to the US after reaching their 16th birthday AND
- who came to the US after June 15, 2007 AND
- who came to the US less than five years ago

This would pose a very different control group and imply different hypotheses about the parallel trend. In the first case, it would be assumed that in the absence of DACA, individuals would have observed the same trend in employment irrespective whether they came to the US before or after their 16th birthday, and irrespective of whether they came to the US before or after 2007. While this would need to be argued in a full paper, I think this comparison is especially debatable in the second case, in which the control group is less educated and has a citizen status. I suggest to clarify the definition of the control group further (and potentially revise it in light of the implications for the difference-in-differences assumptions). This would also be important to interpret the estimated effect considering the chosen control variables: In the end, who is compared with whom?

(2) Empirical model

The author chose to include the year 2012 as a "before" year. As DACA was implemented in 2012, I would suggest to drop this year (probably, the current specification is "only" a downward-biased estimate, but there might be other, short-term anticipation effects). Also, the analysis is not very clear on the choice of control variables. This would need to be added. While no extensive robustness checks are asked for, a second regression without controls would be nice.

(3) Further considerations

I understood the task such that the impact on Hispanic-Mexican Mexican-born individuals was to be measured, such that I would not have included Hispanic-Non-Mexican individuals.

Supplemental Appendix 1: Hypotheses and Analysis Preregistration

This Appendix Section shows only 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.

1.1: 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 feedback and revision
- Round 3: Second replication task
- Round 4: Results after second peer feedback and revision
- Round 5: Third replication task
- Round 6: Results after third peer feedback and revision

Hypotheses. Note that, for consistency with the original registration, these refer to "peer review" instead of "peer feedback", as we have referred to the process in this paper:

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.

1.2: 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 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.

   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.

Supplemental Appendix 2: Additional Figures and Tables

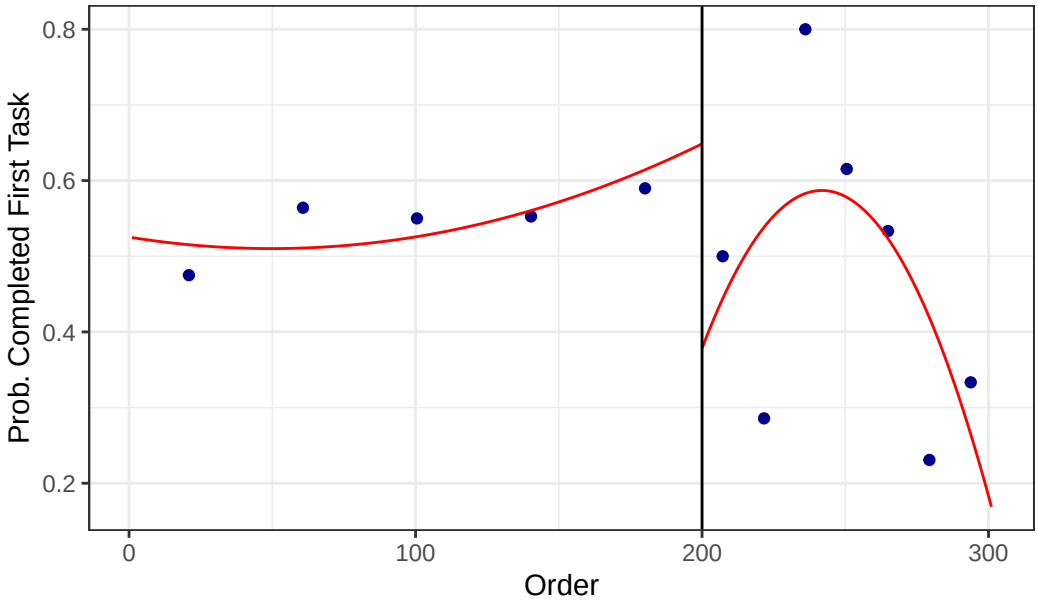


Figure 2.1: Impact of Guaranteed Payment on Probability of Task 1 Completion

Table 2.1: Linear and Quadratic Regression Discontinuity Estimates

	Linear	Quadratic
Intercept	0.543***	0.543***
	(0.036)	(0.035)
Above Cutoff	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 2.2: 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

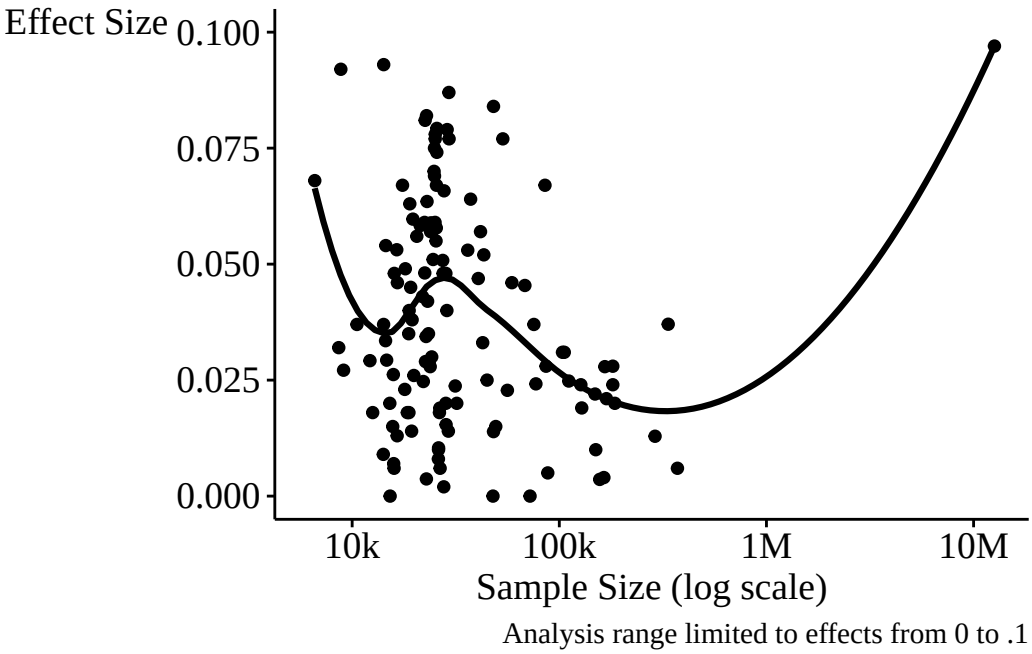


Figure 2.2: Task 2 Effect Size and Sample Size

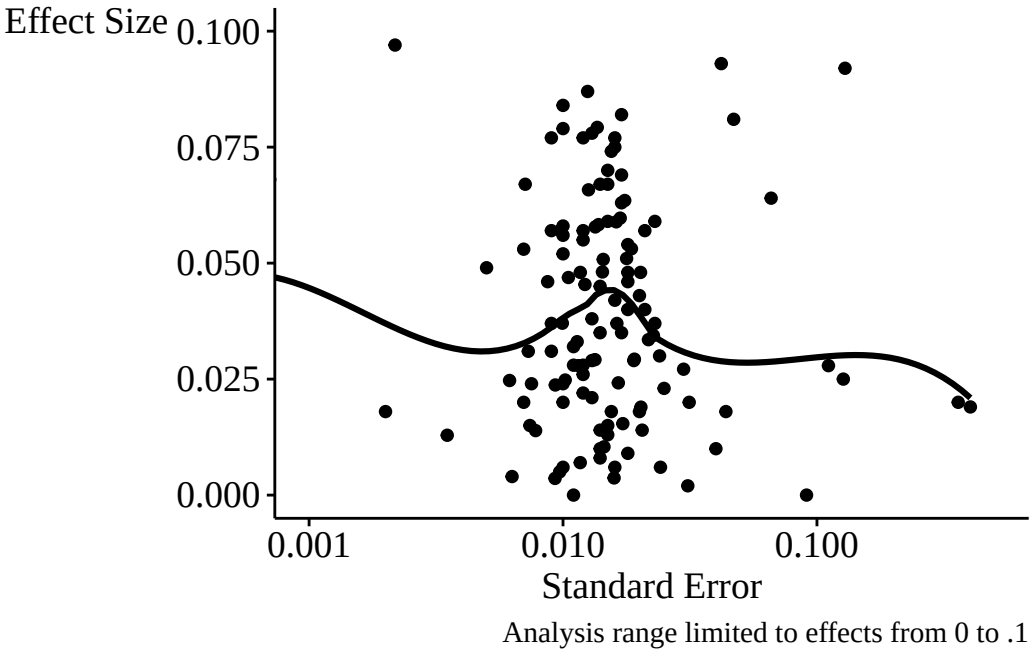


Figure 2.3: Task 2 Effect Size and Standard Error

Table 2.3: 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 2.4: 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 2.5: 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

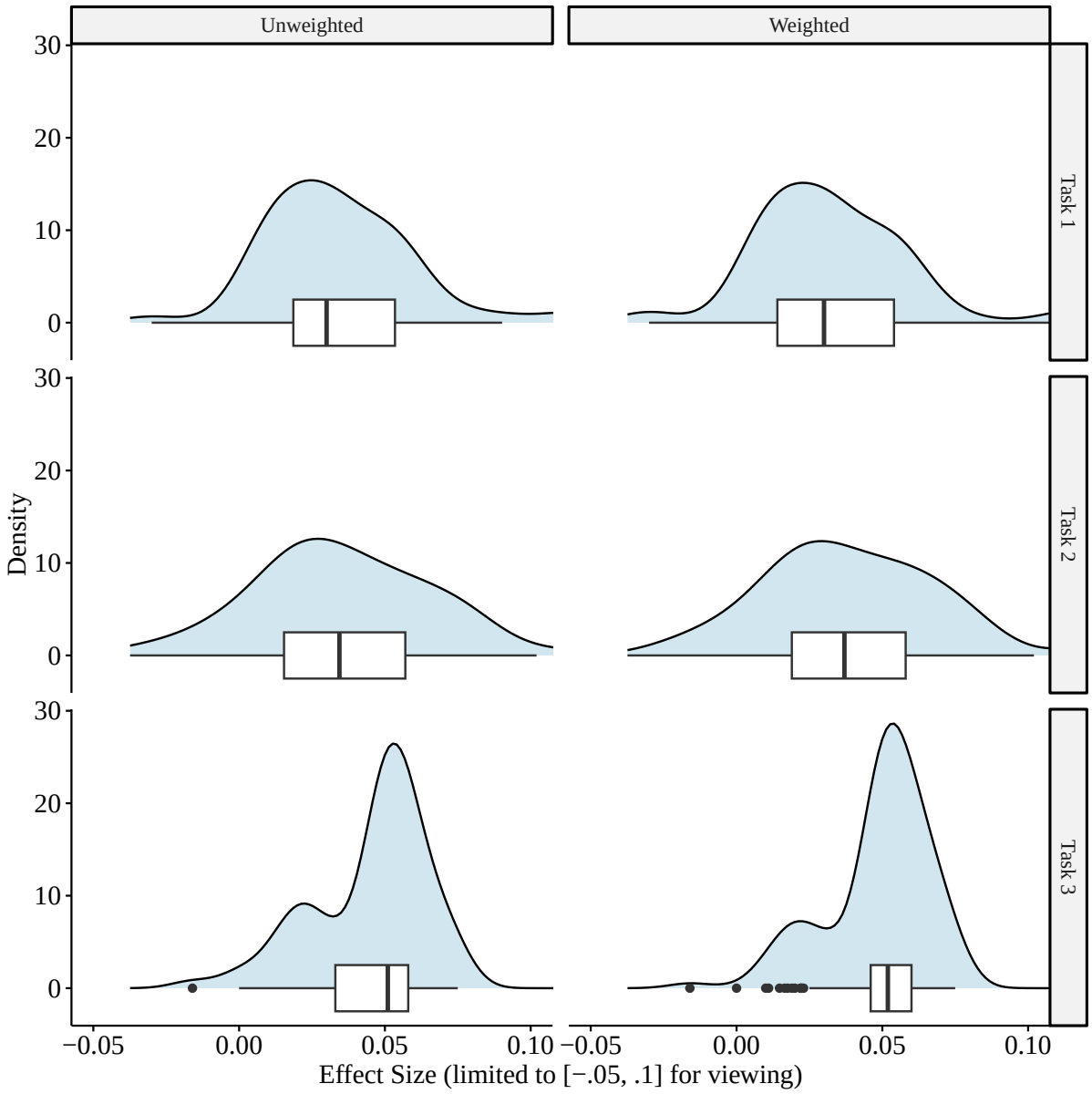


Figure 2.4: Distributions of Reported Effect Sizes with Peer Feedback Qualification

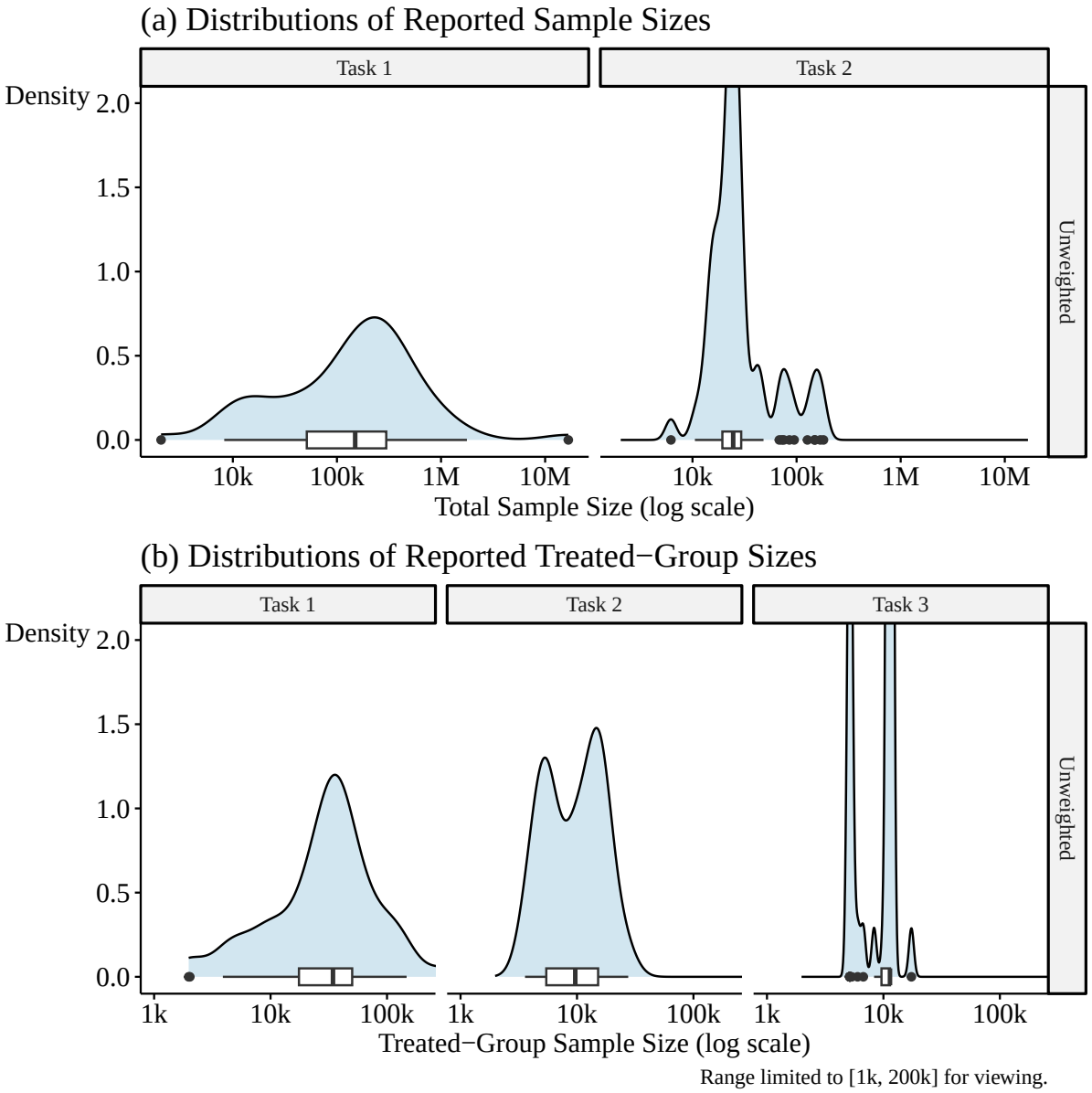


Figure 2.5: Distributions of Reported Sample Sizes and Treated-Group Sizes with Peer Feedback Qualification

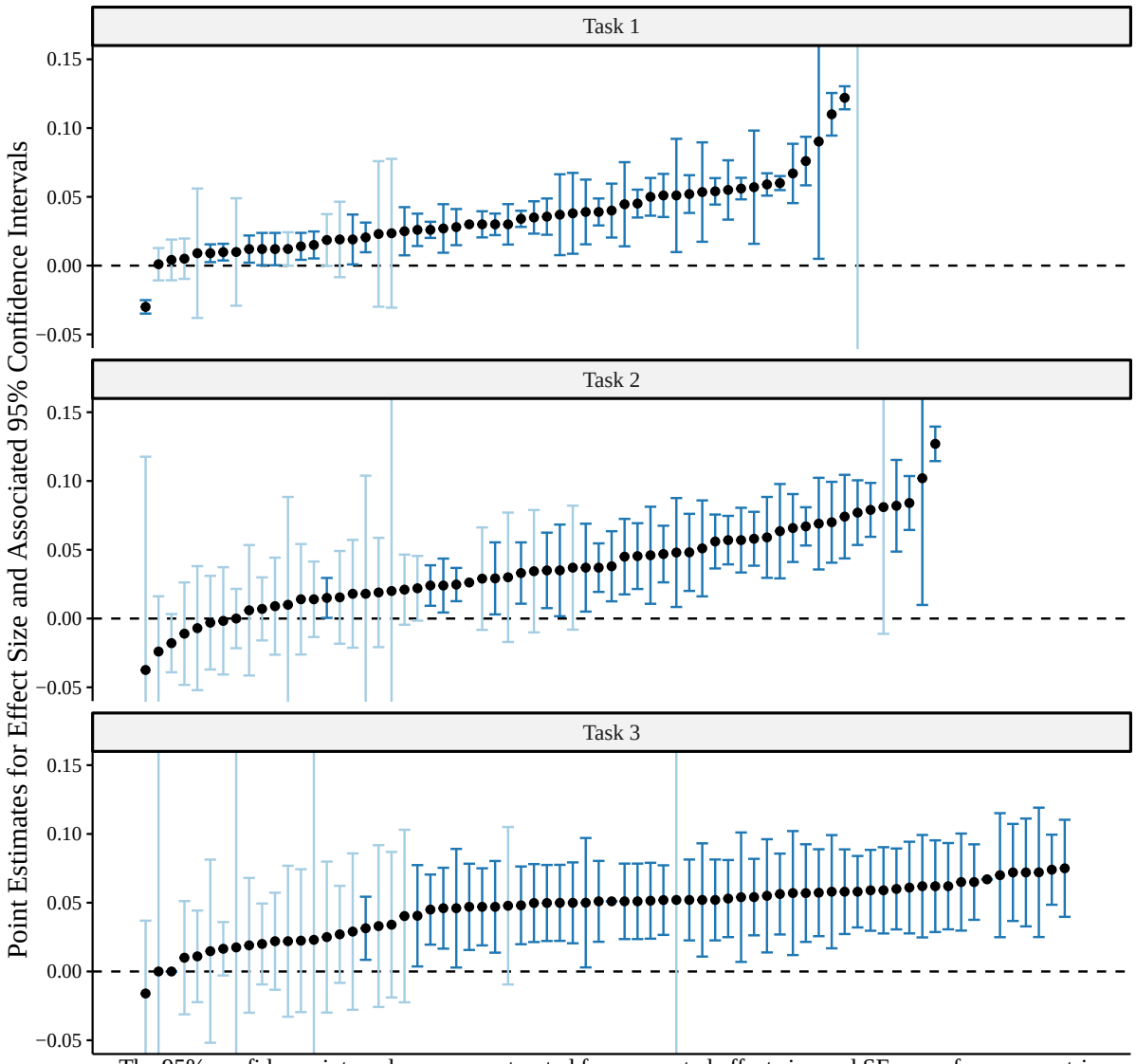


Figure 2.6: Specification Curve for All Reported Estimates with Peer Feedback Qualification

Supplemental Appendix 3: Multi-Analyst Evaluation of Researcher Characteristics

3.1: Analysis by Project Organizer A

Table 3.1 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 5.3, 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 3.2 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 3.1 and 3.2 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 0.05. In 3.2, the p-value for race as an explanatory variable in Task 1 had a p-value below 0.1, 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 3.1 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 0.1 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 3.1: 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	0.764	0.551	0.022	0.973	0.425	0.028	0.254	0.907	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 3.2: 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	1.623	0.172	0.045	0.096	0.984	0.003	0.567	0.687	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.

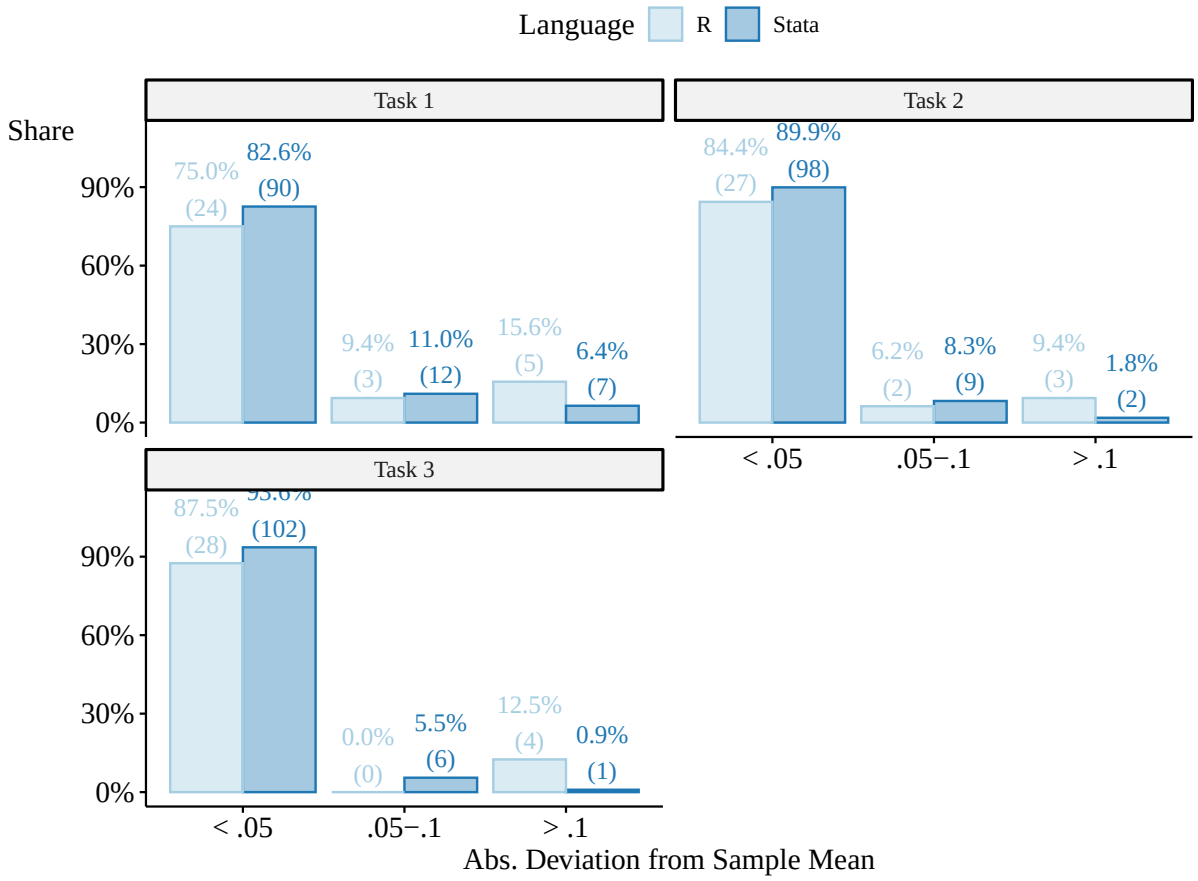


Figure 3.1: Deviation from Sample Mean of Reported Effect by Language

users is relatively low at 32,[35] 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 0.1 or more every round, while all other R researchers with deviations of 0.1 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.

3.2: Analysis By Project Organizer B

Table 3.3 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 $\alpha = 0.1$ 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

[35] This is one lower than the value reported in Section 5.3 because the researcher who was dropped from analysis, mentioned later in Section 4.2, was an R user.

Table 3.3: 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 $\alpha = 0.1$), and those with fewer papers also showed smaller absolute changes than those with 6+ papers (insignificant).

Supplemental Appendix 4: LLM Evaluation of Peer Feedback

In this section we describe how the written peer feedback provided by researchers was turned into scores that we used to limit the sample in Section 5.2.

We wrote an AI prompt (985 words) that explained in thorough detail the context of the peer feedbacks and how an LLM could evaluate them. The full written prompt includes examples, formatting instructions, and peer feedback features to look out for, and is available in the project repository at https://github.com/many-economists/analysis. The key elements are the scoring system used, and instructions given to the LLM on how to evaluate peer review.

Importantly, the specific scores given are sensitive to the prompt designed, and it would be straightforward to design a prompt with the intent of giving a higher or lower average score, if desired. So the LLM scores should not be used to judge the overall quality of work on average. Rather, the scores can be used to make internal comparisons of the quality of one completed research task against another. The desired behavior of the LLM responses, which we achieved, is that the LLMs produce a range of responses that allow us to distinguish one completed replication from another, and that they produce responses that are relatively consistent for the same task if elicited multiple times.

The scoring system given in the prompt was:

Score 1 (Recommendation: Reject & Abandon): This score is reserved for work that is truly unsalvageable or fundamentally misconceived. The reviewer identifies issues so deep that the entire research approach is invalid, the logic is broken, or the work is incoherent. This is a rare and extreme rating.

Score 2 (Recommendation: Major Revisions / Re-think): The reviewer sees merit in the research question, but the execution has serious, structural flaws. The paper requires a significant

re-analysis or a redesign of the core empirical strategy. The work is salvageable, but it needs a major overhaul, not just tweaks.

Score 3 (Recommendation: Standard Revisions): This is the default score for a solid, competent paper that needs improvement. The reviewer believes the core analysis is on the right track but requires specific, meaningful revisions. The review may sound negative and list several points, but the requested fixes do not require changing the fundamental research design.

Score 4 (Recommendation: Good, with Minor Polish): The reviewer is largely satisfied and finds the paper to be in good shape. The feedback consists of helpful suggestions, minor corrections, or requests for clarification that do not challenge the validity of the results. The reviewer may still sound demanding about these small points (nitpicking is common), but the requested changes are about polishing a strong product, not fixing a broken one.

Score 5 (Recommendation: Accept As-Is / Excellent): The reviewer is highly impressed and has no substantive requests for changes to the analysis. Any comments are trivial (e.g., typos) or framed as matters of personal preference. A review can be a 5 even if it contains a minor suggestion, as long as the overall tone is overwhelmingly positive and the suggestion is presented as non-critical.

Initial versions of the prompt tended to lead to very low scores, where the LLM would interpret most reviews as giving a 1 or a 2, justifying it by pointing to any criticism present in the peer review. This ignores that, in academic peer review, even positive reviews contain some criticism. The following guidance was added to the prompt:

A Lack of Praise is Not a Critique: It is normal for a review to contain only criticism. Do not penalize a paper for having no positive comments. Base your score solely on the severity of the critiques that are present. Further, a lot of small and minor critiques does not mean the

review is highly negative, even though there may be a lot of them. To be a 2 or worse a paper needs important flaws.

Volume vs. Severity: A long list of minor, cosmetic critiques is far less severe than a single, structural flaw. Do not be swayed by the number of points raised; focus on their substance and the work required to address them.

Distinguish Demands from Suggestions: Pay close attention to the reviewer's tone. "The author must…" or "The paper lacks…" points toward a 2 or 3 IF the point being raised is significant. "The author might consider…" or "I was curious about…" points toward a 4 or 5.

Ignore Writing/Grammar Comments: Your analysis must focus exclusively on critiques of the research substance. Ignore all comments about writing style, clarity, or grammar.

We trialed the full prompt with several example peer reviews with both Google Gemini 2.5-pro and OpenAI's GPT-5. In the examples evaluated by the project organizers, Gemini 2.5 Pro more closely matched the organizers' assessments of the severity of the substantive criticisms. It also used the full scoring scale, whereas GPT-5 concentrated heavily on the middle categories. We therefore used Gemini 2.5 Pro for the analysis.

We gave Gemini 2.5-pro each peer review ten times, producing a different response and score in each iteration. We averaged together the ten iterations to produce a final score for each peer review, and then rounded that score to the nearest integer.