Observing many students using difference-in-differences designs on the same data and hypothesis reveals a hidden universe of uncertainty

Alexander Wuttke Karolin Freitag Laura Kiemes
Linda Biester Paul Binder Bastian Buitkamp Larissa Dyk
Louisa Ehlich Mariia Lesiv Yannick Poliandri
Celina Schneider Adrian Brenner Andrey Samarskiy

2024-04-08

The recent advent of many-analysts studies highlighted significant variation in research outcomes when multiple teams independently explore identical datasets and hypotheses. This paper contributes to this emerging research paradigm by demonstrating how the many-analysts framework can be incorporated into social science methods classes and by pioneering its use in difference-in-differences (DiD) designs. By engaging students in independent DiD analyses on a specific research question, in this case related to local governance, the study uncovers a broad range of effect sizes and diverging conclusions reported by the analysts, highlighting the subjective nature of analytical choices and their potential impact on research outcomes. This variability underscores the importance of critical thinking and methodological rigor in social science education. The implications extend beyond pedagogy, suggesting that causal inference methods with significant researcher discretion such as DiD designs benefit from robustness analysis with multiple independent analysts.

1 Introduction

In the preceding decade, the paradigm of "many-analysts studies" has gained traction, offering a novel lens with which to examine the robustness of research conclusions (Camerer 2022; Landy et al. 2020). In many-analysts studies, multiple researchers independently investigate the same research question or hypothesis using the same data. Previous research employing the many-analysts approach has been conducted across various fields, including

psychology (Silberzahn et al. 2018), neuroscience (Botvinik-Nezer et al. 2020), gender studies (Schweinsberg et al. 2021), and sociology (Breznau et al. 2022), sometimes involving hundreds of analysts. Many-analysts studies offer insights inaccessible to the standard approach to academic knowledge production. Specifically, letting large numbers of researchers (rather than one team of researchers) explore the same research question reveals the variance of a particular finding across research teams. In this sense, many-analysts studies help understand the reliability of knowledge claims.

This type of uncertainty in scholarly findings has not been at the forefront of academic research or academic teaching. This article posits that many-analysts studies offer a valuable pedagogical tool, elucidating the intricate nature of scientific uncertainty. The simplicity of many-analysts studies, coupled with the potential to stimulate critical thought and discussion, makes them an exemplary educational resource. This paper's first contribution illustrates the integration of many-analysts studies into social science courses, enabling students to actively participate in these studies as analysts. Additionally, the article provides accessible teaching materials (template for slides and syllabus, CC-By licensed) for reuse (https://dx.doi.org/10.17605/OSF.IO/ZC9MH), advocating for a pedagogical approach that intertwines methodological education with insights into meta-science and the philosophy of science.

This study's second contribution lies in extending the many-analysts design to an unexplored research area. Given the limited scope of existing many-analysts studies, their generalizability remains unclear. Notwithstanding the observed variability in research outcomes, it is conceivable that the previously reported variability is not consistent across different disciplines and methods. Certain research fields might demonstrate a lower degree of variability, expressed by more uniform findings among independent researchers. This necessitates then a broader spectrum of many-analysts studies across diverse domains and methodologies to better understand the robustness of research outcomes.

This study conducts what is, to our knowledge, the first many-analysts study in causal inference using observational data. Specifically, 11 BA or MA students each independently conducted a difference-in-differences design on the same research question related to local politics in a class on causal inference methods at *Blinded University*. Although simplicity was one criterion for selecting the research question, the reported findings varied greatly. This suggests a need for more systematic inquiries into researcher variability in the field of causal inference using observational data.

2 Many-analysts studies in teaching

Incorporating many-analysts designs into courses on methodology or research design can offer a profound educational experience. Such integration allows for profound engagement with meta-scientific and philosophical questions about the essence and trustworthiness of science: What is science? When and why can we trust science? What distinguishes the scientific approach to knowledge generation from alternative paths to knowledge such as tradition, revelation, or industry research? What is the "scientific method", if it exists, and

(how) does it always lead to truth? These discussions, enriched by the provided teaching materials, help to foster dynamic and reflective classroom dialogues.

In addressing questions related to the nature and trustworthiness of science, the contemporary feminist philosophy of science offers a unique perspective. Contrary to traditional views that emphasize a 'scientific method,' thinkers like Naomi Oreskes assert that science's epistemic strength lies in its social institutions (Oreskes 2019; but see Wuttke 2020). This viewpoint considers scholarly consensus or dissent as crucial indicators of the credibility and acceptance of scientific knowledge. Oreskes's analysis of near-universal agreement among climate experts regarding anthropogenic climate change exemplifies this approach (Oreskes 2004), suggesting that the robustness and reliability of scientific findings are grounded in collective scholarly agreement. The many-analysts design operationalizes these philosophical considerations, providing a practical framework to explore such concepts.

Direct exposure to many-analysts studies provides students with a tangible understanding of the impact of scholarly consensus or dissent on the perception of reliable knowledge. A pedagogical approach might involve dividing the class for a learning activity: one half of the student review the findings of traditional single-team research, while the other examines the results of a many-analysts study on the same topic. Such an experiment could utilize subjects such as ego depletion in psychology (Dang et al. 2020; Vohs et al. 2021) or immigration in sociology (Breznau et al. 2022; Brady and Finnigan 2014). Students from both groups could then be paired to discuss the credibility of the phenomenon in question. Perhaps such a class assignment would show that many-analysts studies add a novel perspective by highlighting the influence of individual researcher choices and idiosyncrasies, alongside systematic biases (Nelson, Simmons, and Simonsohn 2018), in shaping scientific knowledge and its perceived reliability (also see Kahneman, Sibony, and Sunstein 2023).

3 Students as contributors to many-analysts designs: A difference-in-differences study

The next step in incorporating a many-analysts design into a quantitative methods course is to let students themselves serve as analysts who independently explore the same research question using the same data. Through this immersive experience, students gain a more profound understanding of researcher variability as they observe and compare their findings with those of their peers, each addressing the same question with identical data sets.

In this study, students at BA and MA levels contributed to a many-analysts study as part of their term papers. The project involved a difference-in-differences analysis of data related to bicycle traffic in a local area, specifically examining the accuracy of bicycle count machines that the local municipality uses to inform local transportation policies. The specific research questions (explained in greater detail in the supplementary material) concerned the question whether the observation station on the city's most frequently used bicycle lane systematically undercounted bicycles that took a convenient shortcut on the pedestrian lane, likely circumventing the counting machine. Convenient for a causal design, due to the exogenous

event of construction work cyclists were prohibited from using the shortcut for a limited period of time (treatment). Because all bicycles had to pass by the counting machine during that time period, a difference-in-differences design could be used to causally identify the amount of systematic undercounting as represented in the treatment's effect on the number of counted bicycles.

To ensure a unified approach for the many-analysts design, all students were assigned the same research question: "Does the municipality systematically undercount the number of bicyclists?" They were directed to use the municipality's open data portal, which offers public datasets from bike-counting machines, documenting the number of bicycles before, during, and after the specified treatment period.

Assigning the same task to all analysts ensures a degree of standardization, facilitating the comparison of results. However, the inherent analytical discretion in the Difference-in-Differences design (Huntington-Klein 2022) allows for various decision-making paths, each potentially introducing an element of variability in outcomes (Gelman and Loken 2014). These decisions include considerations such as data from which pre-treatment period is considered for the analysis, whether the parallel trends assumption is seen as met, which units of observation are used as the control group, and which covariates are used in the regression model.

4 Results

Altogether, 11 student analysts submitted 59 models. Slightly over half of the models (56 %) reported an effect indistinguishable from zero. At the same time, a significant portion (39 %) of the models reported a significant positive effect (Statistically significant negative effects: 2 %, no standard errors reported: 3 %)). This variation in retrieved estimates is reflected in the analysts' subjective conclusions. 46 % of analysts concluded in their verbal interpretation that no effect was present, while 54 % believed in a positive effect (negative effect: 0 %). These results show the lack of agreement whether the treatment led to more counted bicycles or not – but there is consensus that the treatment did not reduce the number of counted bicycles. This findings exemplifies that many-analysts designs can reveal agreement in some respects while also revealing disagreement and uncertainty in others respects, indicating what we can and cannot say with confidence.

Disagreements extend to key methodological decisions such as evaluating the parallel trends assumption, further underscoring the inherent variability in analytical approaches and interpretations within DiD designs. The parallel trends assumption is often informed by placebo tests (employed by 34 % of the analysts) or visualizations of trends in the outcome variable (utilized by 44 % of the analysts). For 85 % of models, analysts concluded that the parallel trends assumption was met, but 15 % arrived at the opposite conclusion.

Figure 1 presents a specification curve visualizing the variation in research outcomes and associated factors (Simonsohn, Simmons, and Nelson 2020). Because some analysts transformed the scale of the outcome variable, the figure's upper panel displays treatment effects

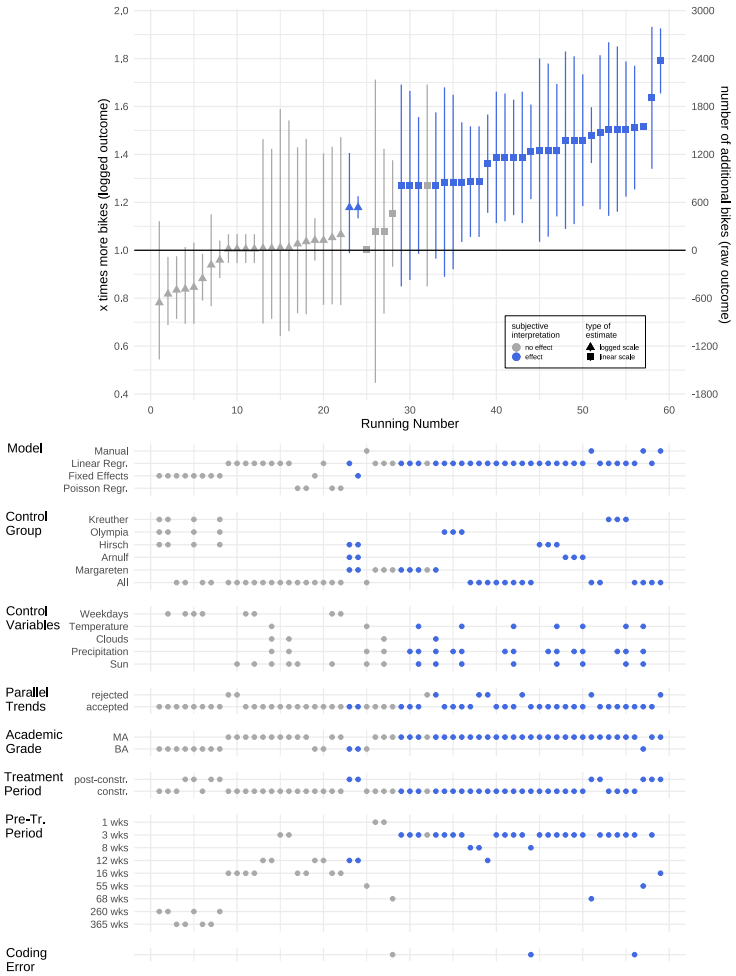

Figure 1: The figure shows all the analysts' effect estimates. The upper panel reports treatment effects on two y-axes for outcomes on a logged (triangle) or natural (square) scale, with corresponding 95% confidence intervals. Estimates are gray/blue when analysts concluded that no/an effect was present. The bottom panel shows the correlates of each provided model estimate. "Model" refers to the kind of statistical model used to estimate the effect. "Control Group" refers to the list of bicycle counting stations that were included in the non-treated experimental group. "Control variables" refers to the list of covariates that were included in a model. "Parallel Trends" denotes whether the analyst who contributed that model concluded that the parallel trends assumption was met. "Academic grade" shows whether the analyst was a BA or MA student. "Treatment period" denotes whether the model considers the period during or after the construction work as the treatment period. "Pre-Treatment period" shows the period included in the model as the pre-treatment period. "Coding errors" report the presence of obvious, consequential coding errors in the syntax.

on two y-axes for outcomes on a logged (triangle) or natural (square) scale. For each estimate, the bottom panel reports corresponding information, such as whether the estimate was provided by a BA or MA student or whether the model controlled for temperature. For instance, the model with the running variable 1 shows a non-significant treatment effect from a fixed effects model without control variables which analyzed the outcome variable on a logged scale and included Kreuther, Olympia and Hirsch counting stations in the control group.

The specification curve illustrates that outcomes from models using a logged scale often appear non-significant, compared to those using a natural scale. The plot also shows how model specifications differed. For instance, some analysts included no control variables, while others controlled for the day of the week or temperature. The relationship between the size or significance of the treatment estimates and other correlates is not clear, however, and should be interpreted with caution. Analysis choices might be correlated with other model decisions and observed factors, for example reflecting differing levels of competency among the analysts. Still, the specification curve effectively demonstrates the diversity in the models not only in the reported results but also in the methodological approaches adopted by the researchers.

5 Discussion

This study showcased the integration of many-analysts studies in social science education. Analyzing the same data for the same research question, the variability in findings among student researchers illuminated the intricacies and uncertainties of empirical research. The study also highlights the promise of many-analysts studies in enriching academic teaching and research. Letting students wander in the "garden of forking paths" (Gelman and Loken 2014) may foster critical thinking and a more in-depth understanding of the scientific process, promoting a more reflective approach to research and data analysis in future scholars.

This exercise underscores the need to take researcher variability more seriously in causal inference with observational data. There are reasons to believe that the field of causal inference with observational data might be prone to researcher variability. Due to the high level of statistical sophistication in this field, researchers face vast analytical discretion, which is rarely constrained by credible pre-registration plans due to the retrospective nature of the research. Although other many-analysts studies have not found that researcher variability decreases with higher method skills (Breznau et al. 2022), it is an open question whether the variation in findings presented here among analysts at the BA or MA level generalize to analysts with more advanced method skills. Further research should use the many-analysts approach to examine systematically analytical robustness across various causal inference methods on observational data.

References

Botvinik-Nezer, Rotem, Felix Holzmeister, Colin F. Camerer, Anna Dreber, Juergen Huber, Magnus Johannesson, Michael Kirchler, et al. 2020. "Variability in the Analysis of a Single Neuroimaging Dataset by Many Teams." *Nature* 582 (7810): 84–88. https://doi.org/10.1038/s41586-020-2314-9.

Brady, David, and Ryan Finnigan. 2014. "Does Immigration Undermine Public Support for Social Policy?" *American Sociological Review* 79 (1): 17–42. https://doi.org/10.1177/0003122413513022.

Breznau, Nate, Eike Mark Rinke, Alexander Wuttke, Hung H. V. Nguyen, Muna Adem, Jule Adriaans, Amalia Alvarez-Benjumea, et al. 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. https://doi.org/10.1073/pnas.2203150119.

Camerer, Colin F. 2022. "The Apparent Prevalence of Outcome Variation from Hidden "Dark Methods" Is a Challenge for Social Science." *Proceedings of the National Academy of Sciences* 119 (52): e2216020119. https://doi.org/10.1073/pnas.2216020119.

Dang, Junhua, Paul Barker, Anna Baumert, Margriet Bentvelzen, Elliot Berkman, Nita Buchholz, Jacek Buczny, et al. 2020. "A Multilab Replication of the Ego Depletion Effect." *Social Psychological and Personality Science* 12 (1): 14–24. https://doi.org/10.1177/1948550619887702.

Gelman, Andrew, and Eric Loken. 2014. "The Statistical Crisis in Science." *American Scientist* 102 (6): 460. https://doi.org/10.1511/2014.111.460.

Huntington-Klein, Nick. 2022. *The Effect: An Introduction to Research Design and Causality*. Boca Raton: CRC Press, Taylor & Francis Group.

Kahneman, Daniel, Olivier Sibony, and Cass R. Sunstein. 2023. *Noise: was unsere Entscheidungen verzerrt - und wie wir sie verbessern können*. 1. Auflage. München: Pantheon.

Landy, Justin F., Miaolei Liam Jia, Isabel L. Ding, Domenico Viganola, Warren Tierney, Anna Dreber, Magnus Johannesson, et al. 2020. "Crowdsourcing hypothesis tests: Making transparent how design choices shape research results." *Psychological Bulletin* 146 (5): 451–79. https://doi.org/10.1037/bul0000220.

Nelson, Leif D., Joseph Simmons, and Uri Simonsohn. 2018. "Psychology's Renaissance." *Annual Review of Psychology* 69 (1): 511–34. https://doi.org/10.1146/annurev-psych-122216-011836.

Oreskes, Naomi. 2004. "The Scientific Consensus on Climate Change." *Science* 306 (5702): 1686–86. https://doi.org/10.1126/science.1103618.

———. 2019. *Why Trust Science?* The University Center for Human Values Series. Princeton, NJ: Princeton University Press.

Schweinsberg, Martin, Michael Feldman, Nicola Staub, Olmo R. van den Akker, Robbie C. M. van Aert, Marcel A. L. M. van Assen, Yang Liu, 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 (July): 228–49. https://doi.org/10.1016/j.obhdp.2021.02.003.

Silberzahn, R., E. L. Uhlmann, D. P. Martin, P. Anselmi, F. Aust, E. Awtrey, Š. Bahník,

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): 337–56. https://doi.org/10.1177/2515245917747646.

Simonsohn, Uri, Joseph P. Simmons, and Leif D. Nelson. 2020. "Specification Curve Analysis." *Nature Human Behaviour* 4 (11): 1208–14. https://doi.org/10.1038/s41562-020-0912-z.

Vohs, Kathleen D., Brandon J. Schmeichel, Sophie Lohmann, Quentin F. Gronau, Anna J. Finley, Sarah E. Ainsworth, Jessica L. Alquist, et al. 2021. "A Multisite Preregistered Paradigmatic Test of the Ego-Depletion Effect." *Psychological Science* 32 (10): 1566–81. https://doi.org/10.1177/0956797621989733.

Wuttke, Alexander. 2020. "Naomi Oreskes, Why Trust Science? (Princeton, NJ: Princeton University Press, 2019). 376 Pages. ISBN: 9780691179001. Hardcover $24.95. - Garret Christensen, Jeremy Freese, and Edward Miguel, Transparent and Reproducible Social Science Research: How to Do Open Science (Berkeley: University of California Press, 2019). 272 Pages. ISBN: 9780520296954. Paperback $34.95." *Politics and the Life Sciences* 40 (1): 126–29. https://doi.org/10.1017/pls.2020.13.

Appendix

R Session Info

```
#> R version 4.3.2 (2023-10-31 ucrt)
#> Platform: x86_64-w64-mingw32/x64 (64-bit)
#> Running under: Windows 11 x64 (build 22631)
#>
#> Matrix products: default
#>
#>
#> attached base packages:
#> [1] stats     graphics  grDevices utils     datasets  methods   base
#>
#> other attached packages:
#>  [1] rmarkdown_2.25   quarto_1.3       scales_1.3.0     here_1.0.1
#>  [5] showtext_0.9-6   showtextdb_3.0   sysfonts_0.8.8   ragg_1.2.6
#>  [9] patchwork_1.1.3  specr_1.0.0      readxl_1.4.3     tidylog_1.0.2
#> [13] lubridate_1.9.3  forcats_1.0.0    stringr_1.5.1    dplyr_1.1.4
#> [17] purrr_1.0.2      readr_2.1.4      tidyr_1.3.0      tibble_3.2.1
#> [21] ggplot2_3.4.4    tidyverse_2.0.0  conflicted_1.2.0 groundhog_3.2.0
#>
#> loaded via a namespace (and not attached):
#>  [1] gtable_0.3.4      xfun_0.41         processx_3.8.2    lattice_0.22-5
#>  [5] tzdb_0.4.0        ps_1.7.5          vctrs_0.6.4       tools_4.3.2
#>  [9] generics_0.1.3    parallel_4.3.2    fansi_1.0.5       pkgconfig_2.0.3
#> [13] Matrix_1.6-4      lifecycle_1.0.4   farver_2.1.1      compiler_4.3.2
#> [17] textshaping_0.3.7 munsell_0.5.0     codetools_0.2-19  htmltools_0.5.7
#> [21] yaml_2.3.7        later_1.3.1       nloptr_2.0.3      pillar_1.9.0
#> [25] furrr_0.3.1       MASS_7.3-60       cachem_1.0.8      boot_1.3-28.1
#> [29] nlme_3.1-164      parallelly_1.36.0 tidyselect_1.2.0  digest_0.6.33
#> [33] stringi_1.8.2     future_1.33.0     listenv_0.9.0     splines_4.3.2
#> [37] rprojroot_2.0.4   cowplot_1.1.1     fastmap_1.1.1     grid_4.3.2
#> [41] colorspace_2.1-0  cli_3.6.1         magrittr_2.0.3    utf8_1.2.4
#> [45] withr_2.5.2       timechange_0.2.0  globals_0.16.2    igraph_1.5.1
#> [49] lme4_1.1-35.1     cellranger_1.1.0  clisymbols_1.2.0  hms_1.1.3
#> [53] memoise_2.0.1     evaluate_0.23     knitr_1.45        rlang_1.1.2
#> [57] Rcpp_1.0.11       glue_1.6.2        minqa_1.2.6       rstudioapi_0.15.0
#> [61] jsonlite_1.8.7    R6_2.5.1          systemfonts_1.0.5
```

Code

```
# library(quarto)
# quarto::quarto_render("index.qmd", output_format=c("pdf", "html", "docx"))
 knitr::opts_chunk$set(fig.format = ifelse(knitr::is_html_output(), 'svg', 'pdf'))
# before running the script, make sure that tinytex
# https://yihui.org/tinytex/
# is installed. Also, make sure that Quarto >= 1.4 is installed
# so that Quarto
# manuscripts work as intended https://quarto.org/docs/manuscripts/.
# Also make sure that groundhog is installed

# install.packages("groundhog")

library("groundhog")
# for reproducible package management, see https://groundhogr.com
#library("here")
# set.groundhog.folder(here::i_am("index.qmd"))

pkgs <- c("conflicted", "tidyverse", "tidylog",
          "readxl", "specr", "patchwork",
          "ragg", "showtext", "here", "scales",
          "quarto", "lubridate", "rmarkdown")
#list all packages here that are used in this project

groundhog.library(pkgs, '2023-12-01')


global_digit <- 0


# Use RAGG for PNGs, see https://ragg.r-lib.org/index.html
knitr::opts_chunk$set(dev = "ragg_png")

here::i_am("index.qmd")
data <- readxl::read_excel(
  here::here("raw_data", "models_handcoded.xlsx"),
  sheet = "treatment")
data <- data |> dplyr::mutate_at(
  dplyr::vars(estimate, std_error, p_value),
  as.numeric)
# number of weeks in pre_treatment_period
weeks_pre_treatment <- data |>
  dplyr::mutate(pre_treatment_period = ifelse(
```

```
    pre_treatment_period ==
      "01.01.2017-31.12.2019 & 01.01.2022-31.12.2023",
    "01.01.2017 -31.12.2023",
    pre_treatment_period)) |>
  separate(
    pre_treatment_period,
    into = c("start_pre_treatment_period", "end_pre_treatment_period"),
    sep = " -") |>
  dplyr::select(start_pre_treatment_period, end_pre_treatment_period)

weeks_pre_treatment <- weeks_pre_treatment |>
  cbind(data)

weeks_pre_treatment <- weeks_pre_treatment |>
  dplyr::mutate(
    start_pre_treatment_period = ifelse(
      pre_treatment_period ==
        "01.01.2017-31.12.2019 & 01.01.2022-31.12.2023",
      "01.01.2017",
      start_pre_treatment_period),

    end_pre_treatment_period = ifelse(
      pre_treatment_period ==
        "01.01.2017-31.12.2019 & 01.01.2022-31.12.2023",
      "31.12.2023",
      end_pre_treatment_period),

    wks = difftime(
        dmy(start_pre_treatment_period),
        dmy(end_pre_treatment_period),
        units = "weeks"),

    wks = ifelse(
      pre_treatment_period ==
        "01.01.2017-31.12.2019 & 01.01.2022-31.12.2023",
      wks - difftime(dmy("31.12.2019"),
                     dmy("01.01.2022"),
                     units = "weeks"),
      wks),

    wks = (round(wks, digits = 0)* -1),

    timespan = factor(wks,
                      levels = sort(unique(wks),
```

```
                                    decreasing = TRUE),
                        labels = paste0(sort(unique(wks),
                                             decreasing = TRUE),
                                        " wks"),
                        ordered = TRUE)
    )

data <- weeks_pre_treatment

overview_weeks <- weeks_pre_treatment |>
  dplyr::select(pre_treatment_period, wks) |>
  dplyr::group_by(pre_treatment_period) |>
  dplyr::distinct(wks)

# Data preparation ----

#Sort data by the size of the estimate
data <- data[order(data$estimate),]
data <- dplyr::mutate(data, position = 0)
for (i in 1:nrow(data)){
  data$position[i]= i
}

# Separate Logarithmic Models and not-logarithmic Models:
# Initialization of new variables
  data <- data |>
    dplyr::mutate(estimate_log = NA) |>
    dplyr::mutate(estimate_total = NA) |>
    dplyr::mutate(std_error_log = NA) |>
    dplyr::mutate(std_error_total = NA)

for (i in 1:nrow(data)) {
  # Set Estimate-Variables, if model is not logarithmic,
  # estimate_log stays NA, same with estimate_total
  # same for Standard Errors
  if (data$logarithm[i] == 1 | data$model[i] == "binomial"){
    #if estimation is logarithmic or model is binomial
    data$estimate_log[i] = data$estimate[i]
    data$std_error_log[i] = data$std_error[i]
  }
  else {
    data$estimate_total[i] = data$estimate[i]
    data$std_error_total[i] = data$std_error[i]
  }
```

```
}


# Mutate variables of specifications:
# mutate effect, parallel trends, coding error and
  # logarithm to factorial variables
data <- data |>
  dplyr::mutate(effect = as.factor(effect)) |>
  dplyr::mutate(parallel_trends = as.factor(
    ifelse(parallel_trends_assumed == 1, "accepted", "rejected")
    )) |>
  dplyr::mutate(coding_error = dplyr::case_when(coding_error == 1 ~ 1,
                                        coding_error == 0 ~ NA)) |>
  dplyr::mutate(logarithm = dplyr::case_when((logarithm == 1) ~ 1,
                                        logarithm == 0 ~ NA))|>
  dplyr::mutate(total_relative = as.factor(
    dplyr::case_when((logarithm == 1|model == "binomial") ~ "Relative",
                     .default = "Total")))

# mutate control variables, so they can fit in one plot:
data <- data |>
  dplyr::mutate(
    control_sun_n = as.factor(ifelse(control_sun == 1, 1, 0))
    )|>
  dplyr::mutate(
    control_precipitation_n = as.factor(
      ifelse(control_precipitation == 1, 2, 0))
    )|>
  dplyr::mutate(
    control_clouds_n = as.factor(
      ifelse(control_clouds == 1, 3, 0)
      )
    )|>
  dplyr::mutate(
    control_temperature_n = as.factor(
      ifelse(control_temperature == 1, 4, 0)
      )
    )|>
  dplyr::mutate(
    control_weekdays_n = as.factor(
      ifelse(control_weekdays == 1, 5, 0)
      )
    )
```

```r
# Mutate control group variable, so the manifestations fit in one plot:
data <- data |>
  dplyr::mutate(group_all = as.factor(
    ifelse(
      control_group == "All",
      1,
      0))) |>
  dplyr::mutate(group_margareten = as.factor(
    ifelse(
      control_group == "Margareten" |
        control_group == "Margareten, Arnulf, Hirsch",
      2,
      0))) |>
  dplyr::mutate(group_arnulf = as.factor(
    ifelse(control_group == "Arnulf" |
             control_group == "Margareten, Arnulf, Hirsch",
           3,
           0))) |>
  dplyr::mutate(group_hirsch = as.factor(
    ifelse(control_group == "Hirsch" |
             control_group == "Margareten, Arnulf, Hirsch"|
             control_group == "Kreuther, Hirsch, Olympia",
           4,
           0))) |>
  dplyr::mutate(group_olympia = as.factor(
    ifelse(control_group == "Olympia"|
             control_group == "Kreuther, Hirsch, Olympia",
           5,
           0))) |>
  dplyr::mutate(group_kreuther = as.factor(
    ifelse(control_group == "Kreuther"|
             control_group == "Kreuther, Hirsch, Olympia",
           6,
           0)))
# Create each plot individually
# Estimates log left, total right
plot_est <- ggplot2::ggplot(
  data,
  mapping = ggplot2::aes(
    x = position,
    color = effect,
    shape = total_relative)) +
  ggplot2::geom_point(ggplot2::aes(y = exp(estimate_log))) +
  ggplot2::geom_pointrange(ggplot2::aes(
```

```
    y = exp(estimate_log),
    ymin = exp(estimate_log - 1.96*std_error_log),
    ymax = exp(estimate_log + 1.96*std_error_log))) +
  ggplot2::geom_point(ggplot2::aes(y = estimate_total/3000 + 1)) +
  #Transforming estimates lineae and logged estimates into one scale
  ggplot2::geom_pointrange(ggplot2::aes(
    y = estimate_total/3000 + 1,
    ymin = (estimate_total - 1.96*std_error_total)/3000 + 1,
    ymax = (estimate_total + 1.96*std_error_total)/3000 + 1)) +
  ggplot2::geom_hline(yintercept = 1, color = "black") +
  ggplot2::scale_color_manual(
    name = "subjective\ninterpretation",
    values = c("darkgrey", "royalblue"),
    labels = c("no effect", "effect")) +
  ggplot2::scale_y_continuous(
    name = "x times more bikes (logged outcome)",
    sec.axis = ggplot2::sec_axis(
      ~.*3000-3000, name = "number of additional bikes (raw outcome)",
      breaks = seq(-1800,3000, 600)), breaks = seq(0.4, 2.0, 0.2)) +
  ggplot2::scale_shape_manual(
    name = "type of\nestimate",
    values = c(17, 15),
    labels = c("logged scale", "linear scale")) +
  ggplot2::labs(x = "Running Number") +
  ggplot2::theme_minimal() +
  ggplot2::theme(axis.title.y = ggplot2::element_text(
    vjust = 1,
    margin = ggplot2::margin(r = -120)),
                 legend.position = c(0.8, 0.15),
                 legend.title = element_text(size=7),
                 legend.text = element_text(size=6),
                 legend.box = "horizontal",
                 legend.box.background = element_rect(
                   color = "black",
                   size = 0.3),
                 legend.key.size = unit(8, "pt"))

# Specification: Model Calculation
plot_model <- ggplot2::ggplot(
  data,
  mapping = ggplot2::aes(x = position,
                         y = model,
                         color = effect)) +
  ggplot2::geom_point() +
```

```
  ggplot2::scale_color_manual(
    guide = "none",
    values = c("darkgrey", "royalblue")) +
  ggplot2::scale_y_discrete(
    labels = c("Poisson Regr.", "Fixed Effects", "Linear Regr.", "Manual")
    ) +
  ggplot2::labs(x = NULL, y = "Model") +
  ggplot2::theme_minimal() +
  ggplot2::theme(axis.title.x = ggplot2::element_blank(),
        axis.text.x = ggplot2::element_blank(),
        axis.title.y = ggplot2::element_text(
          hjust = 0,
          vjust = 1,
          angle = 0,
          margin = ggplot2::margin(l = -15)))

# Specification: Control Group, "All" as single line for clarity reasons
plot_contgroup <- ggplot2::ggplot(
  data,
  mapping = ggplot2::aes(x = position, color = effect)) +
  ggplot2::geom_point(ggplot2::aes(y = group_all)) +
  ggplot2::geom_point(ggplot2::aes(y = group_margareten)) +
  ggplot2::geom_point(ggplot2::aes(y = group_arnulf)) +
  ggplot2::geom_point(ggplot2::aes(y = group_hirsch)) +
  ggplot2::geom_point(ggplot2::aes(y = group_olympia)) +
  ggplot2::geom_point(ggplot2::aes(y = group_kreuther)) +
  ggplot2::scale_y_discrete(
    limits = c("1","2","3","4", "5", "6"),
    breaks = c("1","2","3","4", "5", "6"),
    labels = c("All", "Margareten", "Arnulf",
               "Hirsch", "Olympia", "Kreuther")) +
  ggplot2::scale_color_manual(
    guide = "none", values = c("darkgrey", "royalblue")
    ) +
  ggplot2::labs(x = NULL, y = "Control\nGroup") +
  ggplot2::theme_minimal() +
  ggplot2::theme(axis.title.x = ggplot2::element_blank(),
        axis.text.x = ggplot2::element_blank(),
        axis.title.y = ggplot2::element_text(
          hjust = 0,
          vjust = 1,
          angle = 0,
          margin = ggplot2::margin(l = -12))
        )
```

```
# Specification: Control Variables
plot_control <- ggplot2::ggplot(
  data,
  mapping = ggplot2::aes(x = position, color = effect)) +
  ggplot2::geom_point(
    ggplot2::aes(y = control_sun_n), na.rm = TRUE
    ) +
  ggplot2::geom_point(
    ggplot2::aes(y = control_clouds_n), na.rm = TRUE
    ) +
  ggplot2::geom_point(
    ggplot2::aes(y = control_precipitation_n), na.rm = TRUE
    ) +
  ggplot2::geom_point(
    ggplot2::aes(y = control_temperature_n), na.rm = TRUE
    ) +
  ggplot2::geom_point(
    ggplot2::aes(y = control_weekdays_n), na.rm = TRUE
    ) +
  ggplot2::scale_y_discrete(
    limits = c("1","2","3","4", "5"),
    breaks = c("1","2","3","4", "5"),
    labels = c("Sun", "Precipitation",
               "Clouds", "Temperature", "Weekdays")) +
  ggplot2::scale_color_manual(
    guide = "none", values = c("darkgrey", "royalblue")
    ) +
  ggplot2::labs(x = NULL, y = "Control\nVariables") +
  ggplot2::theme_minimal() +
  ggplot2::theme(axis.title.x = ggplot2::element_blank(),
        axis.text.x = ggplot2::element_blank(),
        axis.title.y = ggplot2::element_text(
          hjust = 0,
          vjust = 1,
          angle = 0,
          margin = ggplot2::margin(l = -1))
        )

# Specification: Parallel Trends accepted or rejected
plot_partrend <- ggplot2::ggplot(
  data,
  mapping = ggplot2::aes(x = position, y = parallel_trends, color = effect)
  ) +
  ggplot2::geom_point() +
```

```
  ggplot2::scale_color_manual(
    guide = "none", values = c("darkgrey", "royalblue")
    ) +
  ggplot2::labs(x = NULL, y = "Parallel\nTrends") +
  ggplot2::theme_minimal() +
  ggplot2::theme(axis.title.x = ggplot2::element_blank(),
        axis.text.x = ggplot2::element_blank(),
        axis.title.y = ggplot2::element_text(
          hjust = 0,
          vjust = 1,
          angle = 0,
          margin = ggplot2::margin(l = -11))
        )

# Specification: Academic Grade
plot_prof <- ggplot2::ggplot(
  data,
  mapping = ggplot2::aes(x = position, y = profession, color = effect)
  ) +
  ggplot2::geom_point() +
  ggplot2::scale_color_manual(
    guide = "none",
    values = c("darkgrey", "royalblue")
    ) +
  ggplot2::labs(x = NULL, y = "Academic\nGrade") +
  ggplot2::theme_minimal() +
  ggplot2::theme(axis.title.x = ggplot2::element_blank(),
        axis.text.x = ggplot2::element_blank(),
        axis.title.y = ggplot2::element_text(
          hjust = 0,
          vjust = 1,
          angle = 0,
          margin = ggplot2::margin(l = 1))
        )

# Specification: Which Treatment (construction or post-construction)
plot_treat <- ggplot2::ggplot(
  data,
  mapping = ggplot2::aes(x = position, y = treatment, color = effect)
  ) +
  ggplot2::geom_point() +
  ggplot2::scale_color_manual(
    guide = "none",
    values = c("darkgrey", "royalblue")
```

```
    ) +
  ggplot2::labs(x = NULL, y = "Treatment\nPeriod") +
  ggplot2::theme_minimal() +
  ggplot2::theme(legend.position = "top") +
  ggplot2::theme(axis.title.x = ggplot2::element_blank(),
        axis.text.x = ggplot2::element_blank(),
        axis.title.y = ggplot2::element_text(
          hjust = 0,
          vjust = 1,
          angle = 0,
          margin = ggplot2::margin(l = 0))
        )

# Specification: pre-treatment period included in Analysis
plot_time <- ggplot2::ggplot(
  data,
  mapping = ggplot2::aes(x = position, y = timespan, color = effect)
  ) +
  ggplot2::geom_point() +
  ggplot2::scale_color_manual(
    guide = "none", values = c("darkgrey", "royalblue")
    ) +
  ggplot2::labs(x = NULL, y = "Pre-Tr.\nPeriod") +
  ggplot2::theme_minimal() +
  ggplot2::theme(axis.title.x = ggplot2::element_blank(),
        axis.text.x = ggplot2::element_blank(),
        axis.title.y = ggplot2::element_text(
          hjust = 0,
          vjust = 1,
          angle = 0,
          margin = ggplot2::margin(l = -14))
        )

# Specification: Coding Errors
plot_error <- ggplot2::ggplot(
  data,
  mapping = ggplot2::aes(x = position, y = coding_error, color = effect)
  ) +
  ggplot2::geom_point() +
  ggplot2::scale_y_discrete(limits = c(1, 1)) +
  ggplot2::scale_color_manual(
    guide = "none",
    values = c("darkgrey", "royalblue")
    ) +
```

```
  ggplot2::labs(x = NULL, y = "Coding\nError") +
  ggplot2::theme_minimal() +
  ggplot2::theme(axis.title.x = ggplot2::element_blank(),
        axis.text.x = ggplot2::element_blank(),
        axis.title.y = ggplot2::element_text(
          hjust = 0,
          vjust = 1,
          angle = 0,
          margin = ggplot2::margin(l = -12)
          ),
        axis.text.y = ggplot2::element_blank())

# uniform x scales with breaks and minor breaks
plots <- list(plot_est, plot_model, plot_contgroup,
              plot_control, plot_partrend,
              plot_prof, plot_treat, plot_time, plot_error)

for (plot in 1:length(plots)) {
  plots[[plot]] <- plots[[plot]] + scale_x_continuous(
    breaks = seq(0, 60, 10),
    minor_breaks = seq(5, 55, 10))
}

# merge plots using patchwork

# patchwork::wrap_plots(
#   plot_est, plot_model, plot_contgroup, plot_control, plot_partrend,
#   plot_prof, plot_treat, plot_time, plot_error, ncol = 1) +
#   patchwork::plot_layout(heights = c(30, 4, 6, 5, 2, 2, 2, 9, 1))

patchwork::wrap_plots(
  plots, plot_error, ncol = 1) +
  patchwork::plot_layout(heights = c(30, 4, 6, 5, 2, 2, 2, 9, 1))

print(sessionInfo(), local = FALSE)
```

Codebook

1. **analyst**: Identifier for the analyst who conducted the analysis (1 to 11).
2. **model_ID**: Identifier for the model used in the analysis (1 to 59).
3. **profession**: Level of education of the analyst (BA, MA).
4. **control_sun**: Binary variable indicating whether sunlight was controlled for (0 = not controlled, 1 = controlled).
5. **control_precipitation**: Binary variable indicating whether precipitation was controlled for (0 = not controlled, 1 = controlled).
6. **control_clouds**: Binary variable indicating whether cloud cover was controlled for (0 = not controlled, 1 = controlled).
7. **control_temperature**: Binary variable indicating whether temperature was controlled for (0 = not controlled, 1 = controlled).
8. **control_weekdays**: Binary variable indicating whether weekdays were controlled for (0 = not controlled, 1 = controlled).
9. **control_group**: List of bicycle stations that were considered as the untreated control group (out of: Arnulf, Hirsch, Kreuther, Margareten, Olympia).
10. **logarithm**: Binary variable indicating whether logarithmic transformation was applied (0 = no, 1 = yes).
11. **coding_error**: Binary variable indicating whether an apparent coding error was present (0 = no, 1 = yes).
12. **treatment**: Whether the period during or after the construction work was considered as the treatment period (constr., post-constr.).
13. **model**: Statistical model used for analysis (binomial, feols, lm, manual).
14. **pre_treatment_period**: Period before treatment that was considered in the model.
15. **treatment_period**: Treatment period that was considered in the model.
16. **placebo_test_performed**: Binary variable indicating whether a placebo test was performed (0 = no, 1 = yes).
17. **parallel_trends_visual**: Binary variable indicating whether parallel trends were checked visually (0 = no, 1 = yes).
18. **parallel_trends_assumed**: Binary variable indicating whether the analysts concluded that the parallel trends assumption was met (0 = no, 1 = yes).
19. **effect**: Binary variable indicating whether the analyst concluded that a treatment effect was present (0 = no, 1 = yes).
20. **estimate**: Estimate of the effect.

21. **std_error**: Standard error of the estimate.
22. **p_value**: P-value associated with the estimate.

The case

This appendix describes the rationale for employing a Difference-in-Differences (DiD) design centered around an investigation into bicycle traffic counts at a municipal counting station in Munich, Germany.

Each student was given the same research question, originating from a distinctive situation at a specific municipal bicycle counting station in Munich, Germany. These counting devices, managed by the local municipality, serve to assess urban bicycle traffic. However, at the Erhardstraße station, evidence suggests that cyclists might have frequently opted to ride through the pedestrian lane instead of the designated bicycle path. This alternate route choice could potentially bypass the counting mechanism, leading to a consistent underestimation of the actual number of bicycles passing through the area.

Due to construction activities, the pedestrian lane at the Erhardstraße bicycle counting station was temporarily blocked. Consequently, all cyclists were compelled to use the designated bicycle lane, ensuring that each passing bicycle was accurately counted during this interval. This unique situation presented an opportunity to apply a Difference-in-Differences (DiD) analysis to explore the potential bias introduced by the ability to bypass the counting machine at this specific location. Analysts were tasked with addressing the question: "Does the municipality systematically undercount its cyclists?"

The Difference-in-Differences approach quantifies treatment effects by contrasting the trends over time on a specific variable in a treated group with the variation in an untreated group, both before and after the application of a treatment. In the scenario of our study, the bicycle counts at the target station (considered the treatment group) during pre-construction and construction phases were measured against the counts at other stations (the control group) over the same periods. Various underlying factors, such as the day of the week and weather conditions, can affect bicycle traffic. Assuming these factors impact all stations uniformly, the DiD method can isolate the treatment effect by estimating what the bicycle count at the focal station would have been in the absence of construction. This approach aims to neutralize the influence of unobserved variables if we can assume that the trends at the treated and untreated units would have developed similarly absent a treatment.

The dataset utilized for this research was sourced from the municipality's open data portal, encompassing not just the daily bicycle counts at each station but also meteorological data. This allowed analysts to selectively include control variables in their analyses. Variables such as the length of the pre-construction period under review, the method used to compute the Difference-in-Differences estimator, the choice of control group, and any variable transformations offered additional analytical flexibility. Analysts also had the option to incorporate data from the period following construction. Moreover, the introduction of a traffic regulation sign banning bicycle use on the pedestrian pathway during or after the construction phase was another factor that could be integrated into the analysis. Consequently, despite the uniformity in the foundational design and dataset, the models developed by analysts exhibited considerable divergence.