From Bikers to Batters: Steering Statistics Through Real-World Problems

By

Amanda Glazer

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Statistics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Philip Stark, Chair
Professor Sam Pimentel
Professor Grace O'Connell
Professor Avi Feller

Spring 2024

Abstract

From Bikers to Batters: Steering Statistics Through Real-World Problems

by

Amanda Glazer

Doctor of Philosophy in Statistics

University of California, Berkeley

Professor Philip Stark, Chair

A problem-first approach to statistics develops statistical methods directly from real world questions and problems. This dissertation illustrates this approach through the development of statistics methods and tools in four disciplines: active transportation, higher education, election auditing and sports. Causal inference and nonparametric methods are emphasized as they avoid typically incorrect parametric assumptions.

The second chapter focuses on active transportation and problems with ensuring data quality. Sufficiently accurate bicycle and pedestrian counts are useful for improving safety analyses, planning infrastructure, and prioritizing funding. The accuracy of instrumental counts is affected by the instrument's sensing technology, details of siting and installation, calibration, random error, and malfunctions. Some of these errors cannot be detected without an independent, accurate count to compare to the instrumental count. But some failures can be detected (imperfectly) through their signal in the count data, which has led to a variety of algorithms to clean and interpolate instrumental count data. We present different methods for flagging questionable data and provide a detailed comparison of data cleaning approaches.

Higher education is the focus of the next chapter, and the central research question is "do female presenters receive more questions or comments than male presenters during academic job talks?" We collect a large dataset of academic job talks from eight UC Berkeley departments from 2013-2019 in order to answer this question. We find that differences in the number, nature, and total duration of audience questions and comments are neither material nor statistically significant. For instance, the median difference (by gender) in the duration of questioning ranges from zero to less than two minutes in the five departments. Moreover, in some departments, candidates who were interrupted more often were more likely to be offered a position, challenging the premise that interruptions are necessarily prejudicial. These results are specific to the departments and years covered by the data, but they are broadly consistent with previous research, which found differences of comparable in magnitude. However, those studies concluded that the (small) differences were statistically significant. We present evidence that the nominal statistical significance is an artifact of using inappropriate hypothesis tests. We show that it is possible to calibrate those tests to

obtain a proper P-value using randomization.

Motivated by the permutation test work in the previous chapter, the fourth chapter develops a method to construct fast exact/conservative Monte Carlo confidence intervals by inverting exact/conservative Monte Carlo tests about parameters. The method uses a single set of Monte Carlo samples, which both reduces the computational burden and ensures that the problem of finding where the P-value crosses α is well posed. For problems with real-valued parameters, if the P-value is quasiconcave in the parameter, a minor modification of the bisection algorithm quickly finds conservative confidence bounds to any desired degree of accuracy. Additional computational savings are possible for common test statistics in the one-sample and two-sample problem by exploiting the relationship between values of the test statistics for different values of the parameter. Examples across a wide range of disciplines are given to illustrate this new method.

The fifth, sixth, and seventh chapters focus on post-election audits. Post-election audits can provide convincing evidence that election outcomes are correct—that the reported winner(s) really won—by manually inspecting ballots selected at random from a trustworthy paper trail of votes. Risk-limiting audits (RLAs) control the probability that, if the reported outcome is wrong, it is not corrected before the outcome becomes official. RLAs keep this probability below the specified "risk limit." Chapter five compares RLAs to a proposed Bayesian alternative, Bayesian audits (BAs). BAs control a weighted average probability of correcting wrong outcomes over a hypothetical collection of elections; the weights come from the prior. RLAs and BAs make different assumptions, use different standards of evidence and offer different assurances. We illustrate these differences using simulations based on real contests. Historically, conducting RLAs of all contests in a jurisdiction has been infeasible, because efficiency is eroded when sampling cannot be targeted to ballot cards that contain the contest(s) under audit. States that conduct RLAs of contests on multi-card ballots or of small contests can dramatically reduce sample sizes by using information about which ballot cards contain which contests—by keeping track of card-style data (CSD). We present a method for using CSD to drastically decrease RLA sample sizes in chapter six. Chapter seven describes an open-source Python implementation of RLAs using CSD for the Hart InterCivic Verity voting system and the Dominion Democracy Suite® voting system. The software is demonstrated using all 181 contests in the 2020 general election and all 214 contests in the 2022 general election in Orange County, CA, USA, the fifth-largest election jurisdiction in the U.S., with over 1.8 million active voters.

In the final chapter, we develop a novel method to quantify the impact of injuries on player performance in baseball. To quantify this impact we can look at the difference between performance the player would have achieved in the absence of injury and after a given injury. This quantity can be estimated by matching injured players to similar non-injured players. However, matching in observational studies faces complications when units enroll in treatment on a rolling basis (e.g., players are injured at different times). To address this issue, we introduce a new matched design, GroupMatch with instance replacement, allowing maximum flexibility in control selection. Second, we propose a block bootstrap approach for inference in matched designs with rolling enrollment and demonstrate that it accounts properly for complex correlations across matched sets in our new design and several other contexts. Third, we develop a falsification test to detect violations of the timepoint agnosticism assumption, which is needed to permit flexible matching across time.

To my 11-year-old self and Fermat's Last Theorem for setting me off on this journey.

Contents

Acknowledgements

First, I want to extend a massive thank you to my advisor, Philip Stark. I am so grateful for his guidance and mentorship. Because of him, my PhD has not only been an incredible learning experience but also genuinely enjoyable. Philip has been extremely encouraging of my wide range of interests and has always supported me in pursuing what I find most meaningful. I am so appreciative of the endless hours he spent with me editing papers, thinking through logic, and going over proofs and arguments.

Thank you to my committee members: Sam Pimentel, Grace O'Connell, and Avi Feller. I have learned so much from each of you and am so grateful for your continual mentorship and support. I appreciate everything I have learned from Sam about causal inference and the research process, as well as clear communication and presentation of statistical ideas. I admire Sam's ability to present complicated statistical concepts in a way that is accessible and engaging to a diverse audience. I am appreciative of the many hours Grace spent with me digging into statistical questions and editing papers. I am impressed with her commitment to finding the right statistical tools to answer impactful, real-world questions. Thanks to Avi for always being willing to share great advice and for his insights on applying causal inference to real-world problems.

Many thanks to all my collaborators over the last few years. Thank you to Jeremy Rue, Maria Smith, Dave Harding, Alex Skabardonis, and the Orange County Registrar of Voters and its staff, including Justin Berardino, Roxana Castro, and Imelda Carrillo. Look Who's Talking was made possible due to a large group of impressive and dedicated students: Hubert Luo, Shivin Devgon, Catherine Wang, Xintong Yao, Steven Siwei Ye, Frances McQuarrie, Zelin Li, Adalie Palma, Qinqin Wan, Warren Gu, Avi Sen, and Zihui Wang.

I was lucky enough to meet in my cohort not only a great collaborator but also a close friend in Jake Spertus. I am beyond grateful that we were able to go through this journey together.

Thank you to the Computational Research for Equity in the Legal System Training Program, Hearts to Humanity and David P. Gardner fellowships for providing me with funding and support during my PhD. Solidarity forever and gratitude to our union, UAW 2865, for providing support, community and an avenue to fight for change.

Hank Aaron once said, "my motto was always to keep swinging. Whether I was in a slump or feeling badly or having trouble off the field, the only thing to do was keep swinging." I am forever indebted to baseball for always reminding me to keep swinging. Thank you to the San Francisco Giants for giving me the best job in the world throughout my PhD. I am so grateful to Paul Bien, Brian Huey and the rest of the Giants baseball operations team for the opportunity to work with them and for inspiring me with their unwavering enthusiasm

for and dedication to the game.

I am so fortunate to have had Yasmin Kouchesfahani and Rami Rashmawi in my life since elementary school. They have inspired and supported me throughout this journey. Thank you to Katie Werner, a weaving goddess, for teaching me to weave, for all our knit-alongs throughout this PhD, and for always being there for me.

Big woof to Donna Grazer for the countless hours she spent watching me write this dissertation and for never letting me work too long.

Thank you to my partner, Nate Grates, for listening to and supporting every crazy idea, long rant and excited update, and for always being willing to pick up brown butter cookies and watch another season of Love Island.

So much love to my family – Tara Weston, Dr. Jeff Weston, Ashley Weston, Jordan Weston, Stephanie Weiss, and Grandma Lena – for their endless love, support and humor.

Thank you to my sister, BB Glazer, whose resilience, passion and ability to get up to excercise at 5 a.m. I admire, for always believing in me and hyping me up.

Finally, I am so appreciative of my parents, without whom I certainly would not be where I am today. I am so grateful for the countless hours they spent sitting with me, teaching me math, and supporting all of my dreams. A tremendous amount of love and thanks to them for their unwavering support and encouragement, for always inspiring me to think bigger, for helping me find the humor in everything, and for always picking up my call, even when it's the tenth time that day.

Chapter 1

Introduction

Statistics is a powerful discipline because of its foundational role in developing and assessing empirical evidence in other fields. I strongly believe that the development of statistics methods should stem directly from real world questions and problems: *a problem-first approach.*

In line with this belief, my overarching research goal is to collaborate closely with experts in other fields to understand methodological problems and develop rigorous, practical solutions to real scientific problems. This process involves learning the context, science, language, sources of evidence, and ways of thinking special to that discipline. This dissertation will center around illustrations of this process across a variety of disciplines.

Causal inference and nonparametric approaches are emphasized in this dissertation. Nonparametric methods are important because they enable valid inference by avoiding typically incorrect parametric assumptions. However, there is a lack of tools that allow researchers in other disciplines to take full advantage of these methods. To this end, a key focus of this dissertation is to develop a large suite of open-source code and tools that allow researchers to easily conduct nonparametric analyses.

In line with my problem-first approach, I will use studies I conducted in a variety of disciplines to illustrate my statistical and computational advances. Each section will focus on research in a different discipline: active transportation, higher education, election auditing, and sports. I adhere to the general data science workflow illustrated in Figure 1.1.

1.1 Active Transportation

The second chapter focuses on *active transportation* and the middle stage of the data science workflow: obtain data. "Garbage in, garbage out," is a phrase most researchers are familiar with: if we are using low quality data, the resulting analysis will also be low quality, regardless of our methods. To this point, data quality, a central challenge in active transportation, is the main focus of this chapter.

Research in active transportation often relies on data from continuous count stations that can be unreliable and inaccurate. The accuracy of the counting device depends on the technology (e.g., inductive loops, pneumatic tubes, infrared detectors), location and details of the installation.

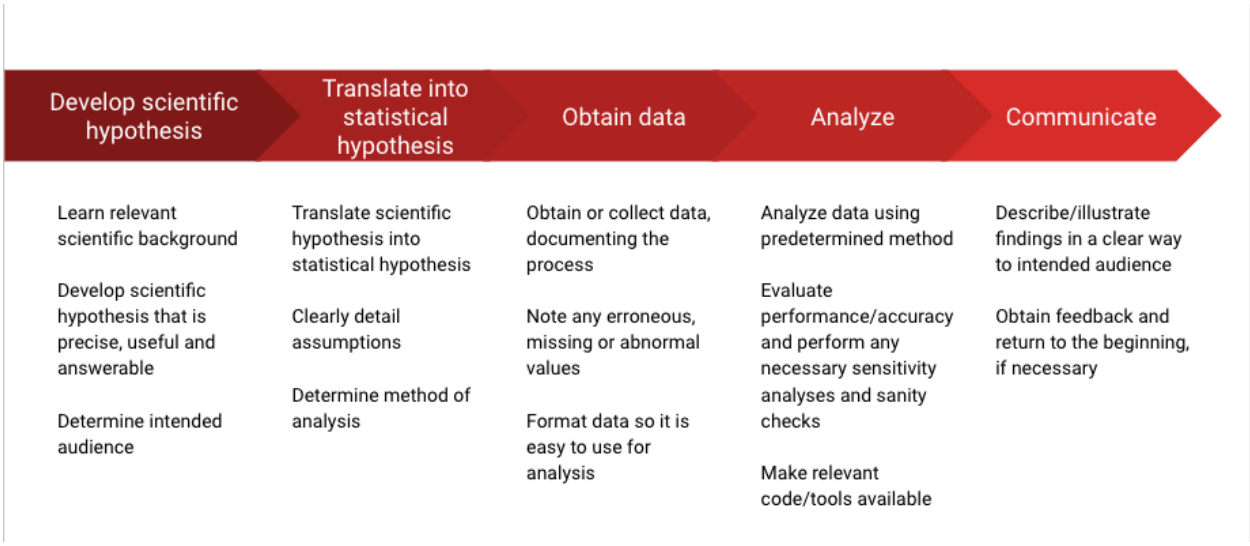

Figure 1.1: Five stages of the data science workflow illustrated: develop scientific hypothesis, translate into statistical hypothesis, obtain data, analyze and communicate.

Without appropriate data checks and cleaning, researchers risk analyzing data that are highly inaccurate. We present a deep dive into data quality procedures exemplifying common failures modes and illustrating the impact of various criteria on the number of datapoints flagged. We present a potential workflow for evaluating data quality and cleaning data for research in active transportation.

1.2 Higher Education and Permutation Tests

The third chapter focuses on *higher education* and issues that can arise in translating a scientific hypothesis into a statistical hypothesis. The central research question in this chapter is whether female presenters receive more questions and comments than male presenters during academic job talks.

To answer this question, we obtained academic job talks from eight UC Berkeley departments from 2013-2019. As in the previous chapter, issues of data quality arise when annotating the videos to create a dataset of questions and comments that lends itself easily to analysis. We contrast our approach to data collection, which requires at least 3 annotators per video, to previous approaches.

In this chapter, we introduce a nonparametric method, permutation (or randomization) tests, to analyze our data. We compare this approach to previous approaches that use parametric methods. We show how the translation from scientific to statistical hypothesis differs between nonparametric and parametric approaches. To partly remedy this gap, we propose a nonparametric callibration of parametric test statistics. To make it easy for others to implement these methods, we have updated the Python packages `permute` and `cryptorandom`.

Motivated by this work on permutation tests, the fourth chapter develops a method for constructing confidence intervals by inverting permutation tests, as well as other Monte Carlo

tests. When the tests are exact or conservative—as some families of such tests are—so are the confidence sets. Because the validity of confidence sets depends only on the significance level of the test of the *true* null, every null can be tested using the same Monte Carlo sample, substantially reducing the computational burden ($O(n)$, where n is the number of data) of constructing confidence sets. The Monte Carlo sample can be arbitrarily small, although the highest attainable confidence level generally increases as the number of Monte Carlo replicates increases. When the parameter is real-valued and the P-value is quasiconcave in that parameter, it is straightforward to find the endpoints of the confidence interval using bisection in a conservative way. For some test statistics, values for different simulations and parameter values have a simple relationship that make more savings possible. Numerical examples are given across a wide range of disciplines showing the usefulness of this new method. A Python implementation of this method is available at `https://github.com/akglazer/monte-carlo-ci`.

1.3 Election Auditing

The fifth, sixth and seventh chapters focus on *election auditing*. *Evidence-based elections* should provide convincing evidence that the reported winners really won. In line with this principle, *risk-limiting audits* (RLAs) have a known minimum chance of correcting a reported outcome if it is wrong (but never change correct outcomes).

Bayesian audits (BAs) have been proposed as an alternative to RLAs. BAs stop without a full hand count only if the posterior probability that the reported winner(s) actually lost, for a particular prior, given the audit sample, is below a pre-specified threshold. While RLAs and BAs purport to address the same problem, they make different assumptions, use different standards of evidence and offer different assurances. The fifth chapter, similar to the third chapter, illustrates challenges that can arise in translating the scientific hypothesis into a statistical hypothesis, through comparison of risk-limiting audits and Bayesian audits.

In the sixth and seventh chapters, the importance of the last stage of the data science workflow comes into focus. Historically, conducting RLAs of all contests in a jurisdiction has been infeasible as it would often lead to a full hand count because ballots were not optimally targeted. In response to this problem we show how keeping track of which ballot cards contain which contests can reduce audit sample sizes by orders of magnitude in typical elections. Code to implement these methods is available in the Python package `SHANGRLA` and described in detail in chapter six. In chapter seven, we demonstrate the software on all 181 contests in the 2020 general election and all 214 contests in the 2022 general election in Orange County, CA, USA, the fifth-largest election jurisdiction in the U.S., with over 1.8 million active voters.

1.4 Sports

The final chapter focuses on *sports*. In this chapter, we develop a method to quantify the impact of injuries on player performance in baseball. One way to quantify impact is as the difference between the value of a performance metric the player would have achieved

in the absence of injury and the value of the same metric achieved after a given injury. This quantity can be estimated by matching injured players to similar non-injured players, a technique called "matching." However, players do not get injured at the same time and non-injured players do not have an injury (or "treatment") time. We propose a method that allows us to match flexibly across time: GroupMatch with instance replacement. Additionally, we propose a block-boostrap approach to inference and a falsification test that can be used to check a key assumption underlying our method's validity. Code to run these methods is available in the R package `GroupMatch`.

Chapter 2

Data checks for bicycle and pedestrian data[1]

2.1 Introduction

Bicycle and pedestrian counts from continuous count stations are foundational for studying active transportation. They are used to track walking and bicycling over time, to study safety, and to prioritize infrastructure projects [Ryus et al., 2014]. Technologies for counting bicycles and pedestrians vary, for instance, inductive loops and pneumatic tubes are widely used to count bicycles, and infrared detectors are used to count pedestrians and bicycles [Ryus et al., 2014]. The accuracy of a counting device depends not only on the technology the counter uses, but also on the details of the installation and location, and there are no set standards or requirements for calibrating counters after they are installed.

Counter data are subject to error due to interruptions in cellular service, weather, power supply variations, insects, lack of maintenance, and other issues [Jackson et al., 2017, Turner et al., 2019]. A critical principle of quality assurance is that it should begin before data are collected through inspection, testing, maintenance and calibration of equipment: otherwise we are not addressing the cause of bad data quality [Turner and Lasley, 2013]. While continuing to push for quality assurance measures, one way to mitigate data quality issues is by applying algorithmic data checks to flag suspicious data. While there are some situations where it is obvious that the data are incorrect (e.g., hourly counts equal to 17 for a year, or exceeding a million in a single day), generally one cannot tell with certainty whether data are bad from the data alone. (Data flagged as bad by algorithmic checks generally should be reviewed and investigated by a human to determine whether the flag was a false alarm.) In practice, data checks must compromise between the rate of flagging good data as bad (false alarms), and the rate of failing to flag bad data (failures to detect problems).

In this chapter, we investigate this tradeoff. We review the literature on data checks. We try to connect the checks to the failure modes of the underlying counting technology they are likely to detect, and to situations in which they are likely to lead to false alarms. Most checks involve one or more "tuning parameters," for instance, the number of consecutive zeros

[1]This chapter comprises a paper submitted for publication co-authored by Philip B. Stark, Krista Nordback, Julia Griswold, Md Mintu Miah, and Alexander Skabardonis.

considered suspicious or the ratio of 3 a.m. to 3 p.m. traffic considered suspicious. We apply common checks to data from 311 continuous bicycle and 144 pedestrian counters in California from 2019 to 2022, using a variety of tuning parameters for each check. We tabulate the percentage of data each check flags as suspicious. We analyze bicycle and pedestrian checks separately and highlight differences. Reasonable choices of tuning parameters can differ substantially for bicycle and pedestrian data, for sites with different volumes, for sites with different land use, etc.: there is no one-size-fits-all optimal choice of checks or of tuning parameters for a given check. We make recommendations based on the California data, which may be a reasonable starting point for other researchers, but our goal is to provide researchers with more information about the differences in checks, as well as additional checks and tools to customize to their counters.

This chapter is organized as follows. Section 2.2 gives background on common data quality issues. We introduce our California dataset in Section 2.3. In Sections 2.4 and 2.5 we go through the data checks for bicycle and pedestrian data respectively, discuss recommendations in the literature, and examine the behavior of the checks for the California data, for a variety of tuning parameters. Section 2.6 gives conclusions and recommendations, and discusses next steps, including optimal imputation of missing data for the purpose of estimating means and totals.

2.2 Failure modes of continuous count stations

Continuous count stations are susceptible to many problems [Jackson et al., 2017, Turner et al., 2019] for instance:

- Installation issues, including
  - poor location
  - misalignment
- Electrical or mechanical failures, including
  - battery depletion
  - damage to inductive loops or burial of loops during road construction or repaving
  - electrical interference
- Detector or data logger hardware or software problems
- Vandalism
- Obstruction of telemetry signal or loss of cellular connection
- Weather
- Insect activity
- Interference from vegetation

- Infrared sensor blockage

These issues manifest in the data in many different ways; we now present some examples.

Insect infestations can cause consecutive zeros and otherwise distort counts. Figure 2.1 shows data before, during and after an infestation of earwigs blocked an infrared counter. After the counter was cleaned, pedestrian volumes returned to normal as shown below. Since this was a combined passive infrared and inductive loop counter, the bicycle count, which used the inductive loop, was not impacted; but insects caused the pedestrian count to be zero for 17 days.

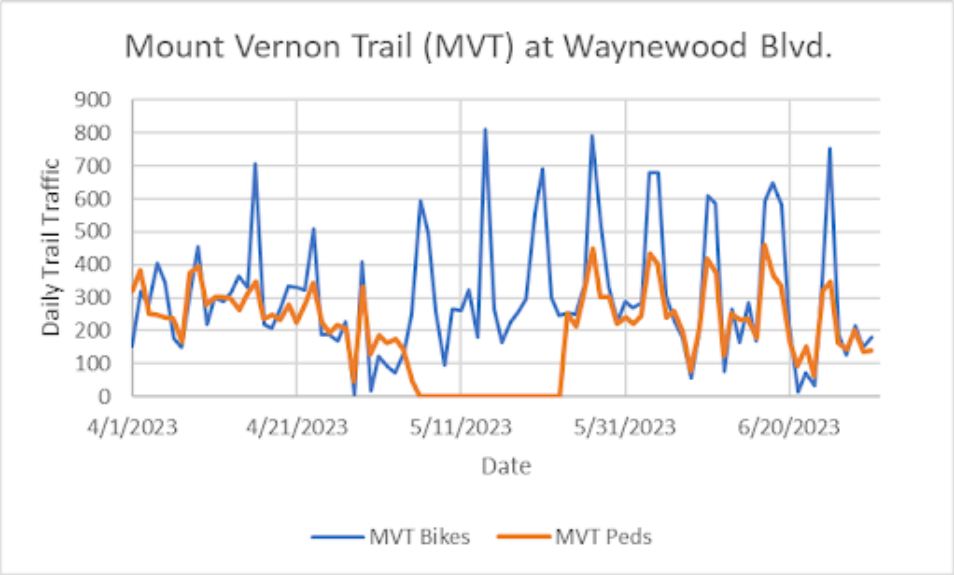

Figure 2.1: Bicycle and pedestrian daily counts from "Mount Vernon Trail at Waynewood Blvd.," in Washington, D.C. Metropolitan Area in 2023.

Location / alignment. Figure 2.2 is an example of erroneous high volume caused by installing an ECO-MULTI with the infrared counter pointed towards a road, so that some cars were counted as pedestrians. (Bicycle counts collected using an associated inductive loop were not affected).

Experience with data quality issues from continuous count stations, as well as previous literature, helped us develop a list of data checks, discussed in Sections 2.4 and 2.5.

2.3 Data

Our study comprises several hundred counters in California, mapped in Figure 2.3 below. The data come from 311 unique bicycle counters and 144 unique pedestrian counters from 2019 to 2022, all manufactured by Eco-Counter: inductive loop detectors for bicycles and/or passive infrared counters for pedestrians. Infrared sensors only detect warm objects. A detection by infrared and the inductive loop is inferred to be a bicyclist on a bicycle; a detection by infrared alone is inferred to be a pedestrian. The raw bicycle data comprised 11,991,546 hours across the 311 sites and the raw pedestrian data comprised 5,715,269 hours across the 144 sites.

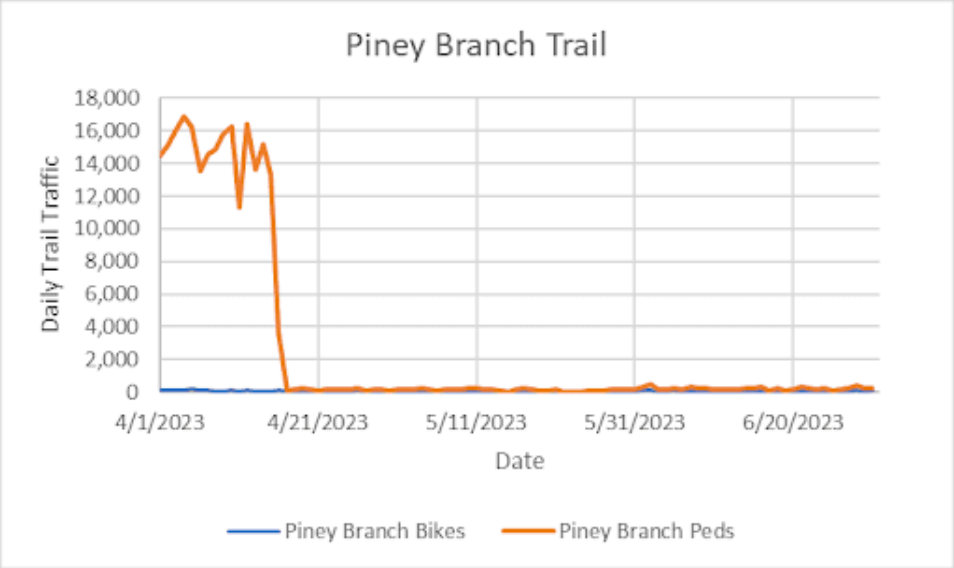

Figure 2.2: Bicycle and pedestrian daily counts from ECO-MULTI on the Piney Branch Trail in Washington, DC in 2023.

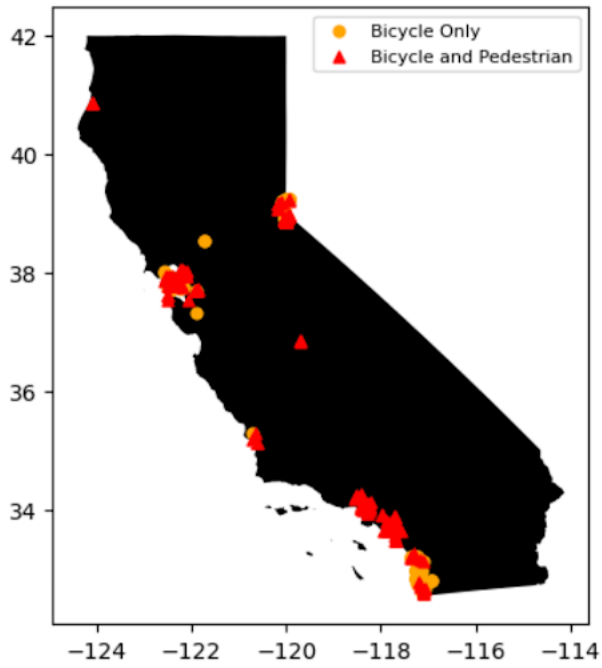

Figure 2.3: Map of the 167 bicycle-only counters (orange dots) and 144 combined bicycle/pedestrian counters (red triangles) in California.

2.4 Data checks for bicycle data

We now turn to the challenge of flagging some data as "bad" based on the data themselves. The method for identifying data as potentially "bad" depends on the type of problem we hope

to detect. As mentioned above, any method that relies solely on the count data will have false positives and false negatives, and human examination of data flagged as "bad," including research to determine possible causes is essential (e.g., to determine whether surprisingly high bicycle counts were due to a bicycle race or whether a set of consecutive zero counts was due to a path closure).

2.4.1 Values repeated consecutively

Consecutive zeros

Many studies recommend flagging consecutive zeros [Jackson et al., 2017, Turner et al., 2019, CDOT, 2016, McNeil and Tufte, 2019, Roll, 2021, Lindsey et al., 2024]. Counters may record zero values if the battery fails or something blocks an infrared sensor, such as an insect infestation. However, consecutive zeros may be correct counts, for instance, at sites on roads or paths that are closed, or at night at sites with low daytime traffic.

Some studies recommend flagging as few as seven [Federal Highway Administration, 2016] or 15 hours [Turner et al., 2019] of consecutive zeros as suspicious; others recommend setting the threshold at 72 hours [Jackson et al., 2017, Roll, 2021, Lindsey et al., 2024] or seven days [MDOT, 2018, Kothuri et al., 2022]. McNeil and Tufte [2019] assert that 12.5 or more hours of consecutive zeros is possibly suspicious and 25 or more hours is suspicious.

The lower the threshold for flagging consecutive zeros, the higher the rate of false positives—flagging good data as bad. For example, some sites have no bicycle or pedestrian traffic at night, in which case the threshold of seven hours is too low. The higher the threshold, the higher the rate of mistaking bad data as good. However, in our experience, resolving a mechanical or electrical counter issue takes on the order of two weeks, so a threshold of even 48 hours of consecutive zeros may be unnecessarily low.

We processed the California bicycle data using various thresholds in the literature (7, 15, 48, 72 and 168 hours). The number of site-days flagged for each consecutive zero threshold is summarized in Table 2.1.

Threshold (in hours)	Number site-days flagged
7	91,931
15	39,290
48	31,453
72	30,820
168	29,820

Table 2.1: Number of site-days that are flagged in our California bicycle dataset using a threshold of 7, 15, 48, 72, and 168 hours of consecutive zeroes.

The number of hours flagged stabilizes for a threshold of about 48 hours. We now look at a few examples to better understand what is being flagged and missed for each choice of threshold.

Seven or more consecutive zeros

A threshold of 7 consecutive zeros often flags low-volume sites that have consecutive zeros at night, which are not suspicious. For this reason, we think a threshold of 7 is too low, at least for low-volume sites. For example, on the night of February 18/19, 2019 there were consecutive zero counts from 11 p.m. to 10 a.m. at 'LIWH: Serrano Trail (Gate 5)' in Orange County (Figure 2.4), but because the volume at the site is generally low, it is not surprising for there to be no traffic at night or in the early morning.

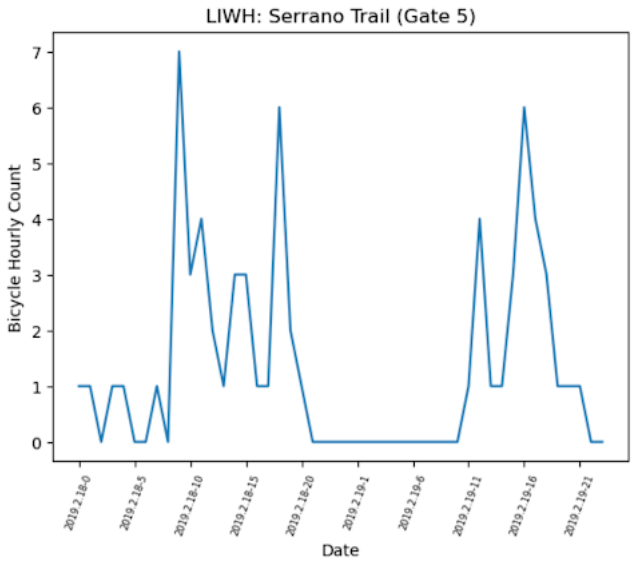

Figure 2.4: Bicycle hourly count data for 'LIWH: Serrano Trail (Gate 5)' in Orange County, CA, from 2/18/2019 at 00:00 to 2/19/2019 at 23:00.

Fifteen or more consecutive zeros

Thresholds of 15 and 7 consecutive zeros have similar issues. Both often flag low-volume sites at night. For example, consider 'BART AT SPIRE' in San Francisco, the night of 1/5/2019–1/6/2019, plotted in Figure 2.5. A threshold of 15 consecutive zeros flags this day as suspicious, but a threshold of 48 does not. This is a low volume site where zeros at night are not uncommon. We conclude that the threshold of 15 consecutive zeros is too low, at least for low-volume sites.

Fourty-eight or more consecutive zeros

The number of runs of more than a threshold number of consecutive zeros decreases slowly once the threshold exceeds 48 hours. We believe that for thresholds of at least 48 hours, most runs of zeros really are bad data, with the exception of sites on closed paths, where the zeros may be real; conversely, when zeros result from counter malfunctions (e.g., because of dead battery or insect infestation), we generally expect bad data to continue for more than 48 hours. We conclude that a threshold of 48 hours of contiguous zeros in the

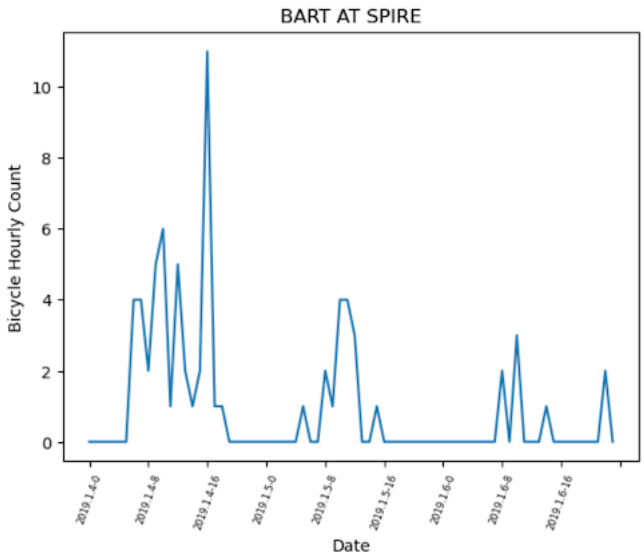

Figure 2.5: Bicycle hourly count data for 'BART AT SPIRE' in San Francisco, CA, on 1/4/2019 through 1/6/2019.

data is reasonable for catching bad data due to instrument malfunctions and unlikely to raise false alarms at low-volume sites, but it may erroneously flag good data on paths or roads that are closed for more than 48 hours.

There are reasons that a threshold of 48 consecutive zeros might correctly identify bad data that higher thresholds would not identify, for instance, battery or transmission problems. Figure 2.6 shows data from 'San Diego: Harbor Dr Multi-Use Path EB & WB' in San Diego, CA, in 2019. In May, 2019, a data transmission issue caused more than 48 hours—but fewer than 72 hours—of consecutive zeros. Attempts to fix the issue never entirely succeeded. The initial gap on May 14–16, 2019, is more than 48 hours but less than 72 hours.

Strings of consecutive zeros also result from street or path closures that eliminate bicycle traffic. Depending on the goal, data during closures might be considered "correct" and included (because the counts really are zero), or "exceptional" and excluded (because they are not typical for the location, unless the closures recur daily, seasonally, or on some other schedule). Similarly, in rural and low-volume areas, the true counts might be zero for 48 or more hours. We recommend manual reviewing flagged data and investigating the cause of extended runs of zeros.

Consecutive identical nonzero values

Some studies filter out consecutive non-zero values repeated more than three times [Turner et al., 2019] or six times [Kothuri et al., 2022]. In our experience, it is rare that identical consecutive non-zero values are bad data; instead, they usually result from coincidences (for small counts) and intentional data manipulation (for large counts, e.g., using 9, 99, or -99 to denote missing or bad data). McNeil and Tufte [2019] found "at least one near-certain glitch of this type" in their dataset. They recommend rejecting counts with more than nine consecutive identical nonzero values, regardless of site volume, and give detailed

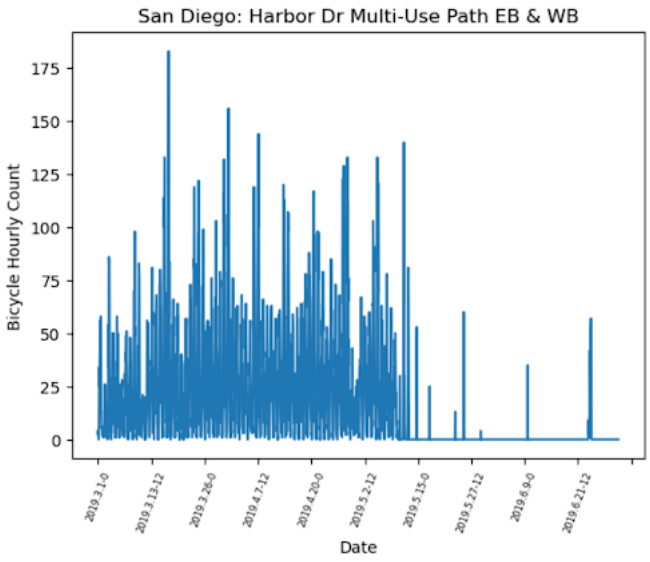

Figure 2.6: Bicycle hourly count data for 'San Diego: Harbor Dr Multi-Use Path EB & WB' in San Diego, CA, on 3/1/2019 through 6/30/2019.

recommendations based on typical count volume (e.g., flag runs of eight or more if the count exceeds two, and runs of six or more if the count exceeds ten).

In practice, this test often erroneously flags good data as bad. Especially at low-volume sites, runs of more than six repeated low counts (e.g., of one, three or even ten) are not unusual. In our dataset, because we have several low volume sites, we find that flagging three consecutive values results in thousands of flags, which we conclude are not suspicious after manual review. Even flagging runs of six or more consecutive values may generate a large number of false alarms, unless the consecutive values are also required to be large. The number of site-days flagged for repeated nonzero values lasting longer than 3 and 6 hours, for values greater than 1, 2, and 10, is summarized in Table 2.2.

Count threshold	Site-days flagged, 3+ consecutive values	Site-days flagged, 6+ consecutive values
1	31,825	483
2	15,642	190
10	2,031	22

Table 2.2: Number of site-days flagged for runs of three or more and six or more consecutive nonzero values, where those values are greater than 1, 2, or 10.

We do not recommend using this check without manually investigating the data it flags. If a particular value might have been used to encode null data (e.g., 9, 99 or -99) we recommend searching for that value as those flags will correctly identify bad data. We also recommend checking for negative or noninteger values and filtering those out. After manual review, none of the consecutive values in our bicycle dataset seemed suspicious.

2.4.2 Extreme Values

Night/Day Ratio

Some studies [Turner et al., 2019] compare nighttime counts to daytime counts. A high ratio may indicate a number of issues; for instance, the timestamps might have been shifted by 12 hours or the instrument might be counting nocturnal insect activity as traffic.

This test requires specifying the daytime and nighttime hours to compare, for instance, comparing 3 a.m. to 3 p.m. We find that flagging days where the 3 a.m. to 3 p.m. ratio exceeds one and the 3 p.m. count is at least three is a reasonable check. Requiring the 3 p.m. count to be at least three eliminates situations where the 3 p.m. count is zero, and reduces flags that we determined were false alarms after manually inspecting the data. We considered ratios based on longer intervals, such as 8 p.m. to 7 a.m. count versus 8 a.m. to 7 p.m. but those tests generated a large number of false alarms, especially at lower count sites. For instance, that range flagged all the data from Placer and El Dorado counties.

Checking whether the 3 a.m. count exceeds the 3 p.m. count and that the 3 p.m. count is at least 3, flags 155 dates with 3 a.m. counts ranging from 4 to 4,838 and ratios ranging from slightly above 1 to 39.5. If instead we require 3 p.m. counts to be greater than 10, 20 or 50, the number of flags reduces to 83, 61, and 44 respectively. If instead we require the ratio to be greater than 1.5, 2, or 3, the number of flags becomes 86, 66, and 53, respectively.

Hard Limits

Some studies recommend flagging hourly and daily counts that exceed some limit. However, there is a wide range of recommended limits [Turner et al., 2019, McNeil and Tufte, 2019, Roll, 2021, Kothuri et al., 2022]. We would expect that an appropriate limit would depend on the site, so it is necessary to study each site before choosing a limit for that site.

Spikes or unusually high counts may be caused by battery failure or insects being counted. On the other hand, data spikes can be legitimate, for instance, because there was an unusual event such as a race or parade. Researchers must decide whether to remove data from special events because they are unusual, or keep the data because the data reflect the actual traffic. This decision depends on how the results will be used. Figure 2.7 shows the number of hours flagged for a variety of hourly limits. The data include 2,854,755 nonzero hourly counts, so the total number of hours flagged for many thresholds is a small percentage of the data, e.g., 0.05% of hours with nonzero count have a count greater than 1500.

The maximum hourly count in our dataset is 10,419. There are 1,493 hours (0.05%) with counts greater than 1,000. We conclude that a threshold of 1,000 is reasonable for our data. We recommend that researchers plot their data in the form of Figure 2.7 to inform their choice of an upper limit for such flags.

Figure 2.8 shows hourly counts for '(14) Duboce Bike Path behind Safeway' in San Francisco, CA, for 2022. Beginning October 8, 2022, the counts vary wildly for the rest of the year, ranging from 597 to 10,419. Many of these hours are flagged if we use a threshold of 1,000. We find this period of counts unusually high; an hourly count of 10,419 means just under 3 bicyclists per second were passing the counter. The high counts at all hours of the day also indicate that this data is suspicious.

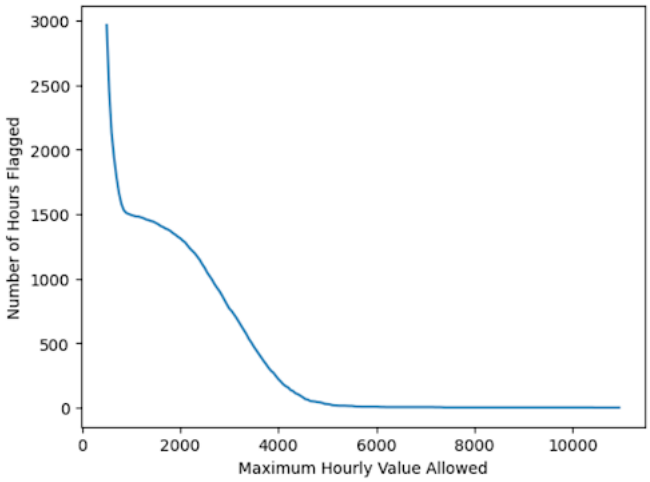

Figure 2.7: Number of hours flagged in our dataset versus the maximum hourly count value allowed.

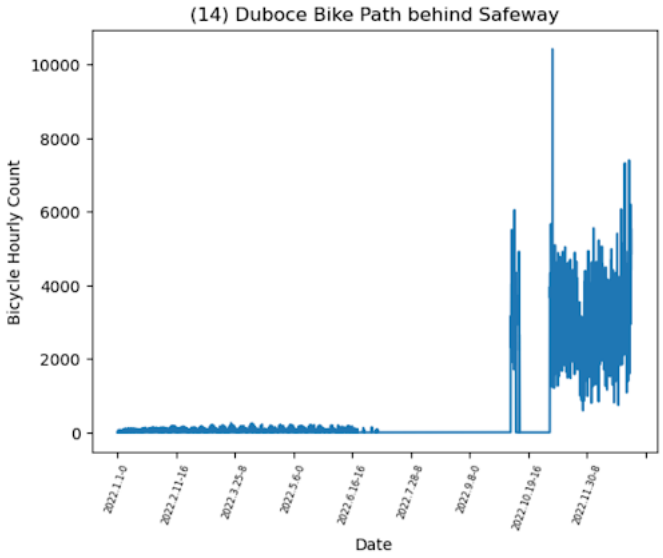

Figure 2.8: Bicycle hourly count data for '(14) Duboce Bike Path behind Safeway' in San Francisco, CA, in 2022.

An example of a spike in legitimate hourly data is from 'Bayshore Bikeway (Chula Vista)' in San Diego, CA, on August 25, 2019. See Figure 2.9. There was an hourly count of 1,165. However, this was due to the annual "Bike the Bay" event. With a threshold of 1,000 this hour would be flagged and it would be up to the reviewer to determine if hourly data on this special event should be included in analysis or not. On November 16, 2019, at 10 a.m. there was an hourly count of 569. However, a google search did not reveal any special events.

We might also consider hard limits for the *daily* counts. Figure 2.10 shows the number of days flagged versus the maximum daily count allowed for the full California dataset. The number of days flagged decreases substantially for daily limits around 1,800–3,000. There

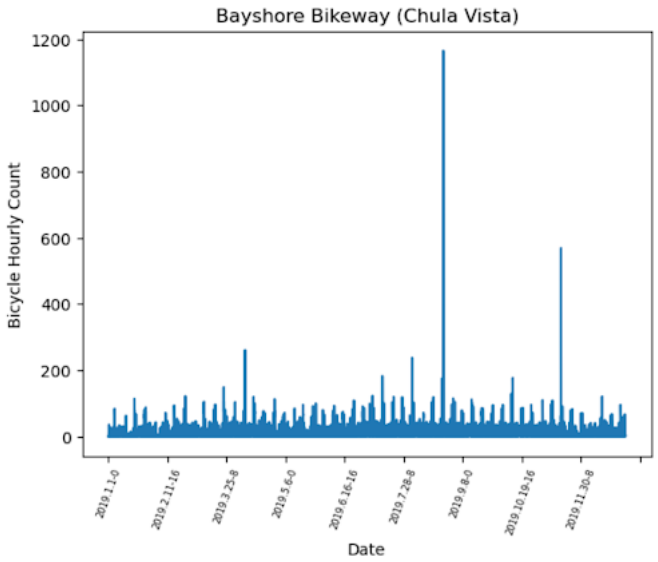

Figure 2.9: Bicycle hourly count data for 2019 at 'Bayshore Bikeway (Chula Vista)' in San Diego, CA.

are 913 days with counts greater than 3000 (0.2% of the 499,662 data days). There are 255 data days (0.05%) with counts greater than 4,000. We find 4,000 to be a reasonable daily limit for our data.

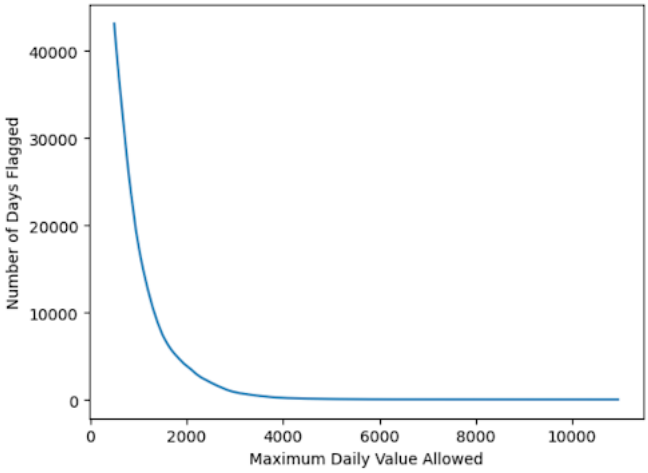

Figure 2.10: Number of site-days with daily counts that exceed a given limit, as a function of that limit, for the California dataset.

Adaptive Upper Bounds

Some researchers advocate flagging data that exceeds a time-varying upper bound [Jackson et al., 2017, Turner and Lasley, 2013, Turner et al., 2019, CDOT, 2016, Roll, 2021]. For example, one might flag data that exceed (Q3 + C * IQR), where Q3 is the upper quartile,

IQR is the interquartile range, and C is a constant or a multiple of the IQR—all computed in a moving window, e.g., two months. Such checks flag counts that are unusual for a particular site, and can adapt to seasonal variation. However, they are not sensitive to problems that persist for a substantial portion of the window: the bad data become the bulk of the data against which new data are tested. This is illustrated in Figure 2.8, where the bad data persists for months.

Flexible upper bounds of this type tend to have high false alarm rates at low-count sites, where the IQR and standard deviation are also small. Consider the count data at 'BART AT SPIRE' in San Francisco, CA, plotted in Figure 2.11. The count of nine on March 3, 2019, and the count of 11 on March 7, 2019 are flagged by many checks, because the two-month rolling lower and upper quartile are 0 and 2, respectively, and the mean and standard deviation are 1.5 and 2.6 respectively. We advise caution when using such checks at low-volume sites.

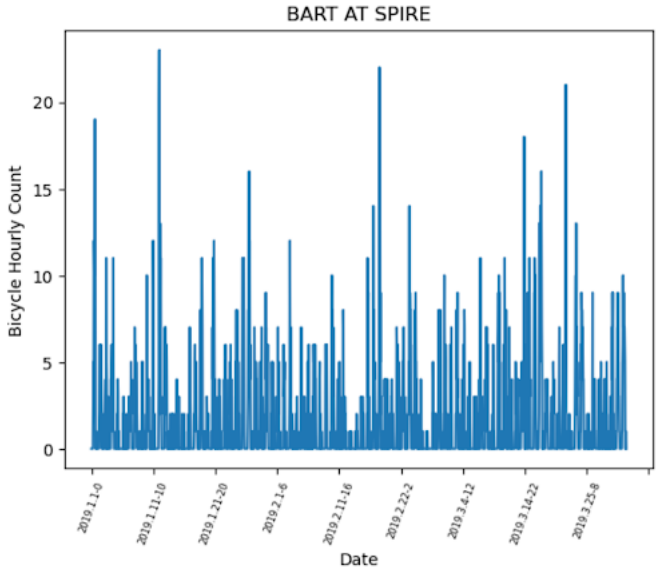

Figure 2.11: Bicycle hourly count data from 'BART AT SPIRE' in San Francisco, CA, from January 1, 2019, through March 31, 2019.

Table 2.3 shows the number of hours flagged for various upper bounds.We also tabulate the number of hours flagged for each upper bound if we only flag hours that are greater than the upper bound if the upper bound is greater than 10, 20 and 50. For those tests, if the upper bound is less than or equal to 10, 20 or 50 we do not flag any data. Consecutive zeros lasting longer than 48 hours were removed before applying this check to avoid diluting the good data with spurious zeros. Each upper bound is calculated using a two-month moving window. The majority of hours are flagged by an upper bound that is less than 50, for all checks: data at lower volume sites tend to be flagged more often. Thresholds involving a rolling mean and standard deviation tend to be less sensitive than those involving percentiles. Some studies filter more aggressively; for example, Turner and Lasley [2013] remove every zero count before calculating the mean, Q3, and IQR.

Upper Bound	$Q3 + 3 * IQR$	$Q3 + 5 * IQR$	$Mean + 2 * SD$	$Mean + 3 * SD$	$Mean + 5 * SD$
Hours Flagged	79,240	21,895	196,092	73,893	9,927
Hours Flagged, Upper Bound > 10	45,957	17,130	160,128	63,445	8,430
Hours Flagged, Upper Bound > 20	35,337	13,368	115,371	51,372	7,109
Hours Flagged, Upper Bound > 50	20,120	8,055	69,011	31,584	4,194

Table 2.3: Number of hours for which the count exceeds a threshold computed from a sliding 2-month window of data. Q3: upper quartile. IQR: interquartile range. Mean: arithmetic mean. SD: sample standard deviation. Rows considering upper bound > 10, 20 and 50 do not flag data when the upper bound is less than or equal to 10, 20 and 50.

Next hour multiple

Another type of adaptive upper bound we can consider is flagging hourly counts that increase by more than some factor of the previous hour [Turner et al., 2019]. Similar to the hard limit and adaptive upper bond checks, this check is intended to flag volume spikes due to insects, battery issues, and other maintenance issues. This check could be a good alternative to the adaptive upper bounds discussed in the previous section as this check still flags data that is unusual relative to what we have seen, but is less sensitive. The number of hours flagged in the California data for which one of the counts exceed a factor 5, 10, 15, or 20 times the count in the previous hour, where the previous hours' count is at least 10, is summarized in Table 2.4.

Factor	Hours flagged
5	1,437
10	76
15	27
20	12

Table 2.4: The number of hours flagged in the California bicycle data for which one of the counts exceed a factor times the count in the previous hour, where the previous hours' count is at least 10, and the factor is 5, 10, 15, or 20.

A factor of 5 generally results in many false positives. A factor of 10 or greater generally seems to work well, and there are far fewer flags for factors larger than 10.

2.5 Data checks for pedestrian counts

Few studies address data checks for pedestrian data specifically, and those that do generally use the same data checks for bicycle and pedestrian data [Jackson et al., 2017, Roll, 2021, Lindsey et al., 2024]. Nordback et al. [2016] stress the need for data checks for pedestrian data. This section applies the same kinds of checks used above for bicycle data to pedestrian data and discusses differences. The results show that pedestrian-specific checks are needed.

2.5.1 Values repeated consecutively

Consecutive zeros

There are more runs of consecutive zeros for pedestrian data than for bicycle data. However, the number of runs drops more slowly starting at a lower threshold for the pedestrian data. The number of site-days flagged for a threshold of 7, 15, 48, 72 and 168 hours is summarized in Table 2.5.

Threshold (in hours)	Number site-days flagged
7	170,533
15	136,768
48	132,246
72	131,936
168	131,237

Table 2.5: Number of site-days that are flagged in our California pedestrian dataset using a threshold of 7, 15, 48, 72, and 168 hours of consecutive zeroes.

We find that a lower threshold is likely appropriate for the pedestrian data. While a threshold of 7 hours of consecutive zeros still flags nighttime counts at some low-count sites, a threshold of 15 hours seems to have relatively few false alarms. For example, Figure 2.12 shows pedestrian count data from 'Albany Bay Trail' in Albany, CA, which contains a string of zeros from October 23, 2021, at 8 p.m. through October 25, 2021, at 6 a.m.. We believe these zeros are erroneous, but they would not be flagged by a threshold of 48 hours of consecutive zeros.

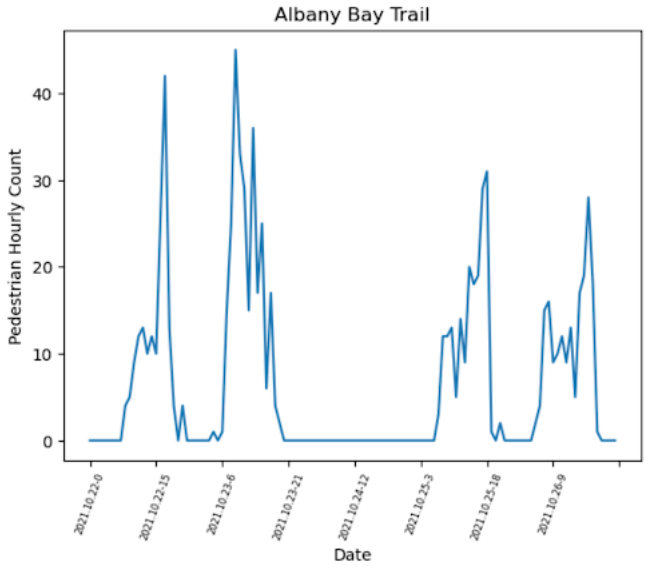

Figure 2.12: Pedestrian hourly count data in October, 2021 from 'Albany Bay Trail' in Albany, CA.

Repeated nonzero values

Flagging short runs of repeated nonzero values or longer runs of small nonzero values identifies many hours of pedestrian data as "bad," albeit fewer hours than the same checks applied to bicycle data. The number of site-days flagged for repeated nonzero values lasting longer than three and six hours, for values greater than 1, 2, and 10, is summarized in Table 2.6.

Count threshold	Site-days flagged, 3+ consecutive values	Site-days flagged, 6+ consecutive values
1	9,945	224
2	5,189	135
10	880	119

Table 2.6: Number of pedestrian site-days flagged for runs of three or more and six or more consecutive nonzero values, where those values are greater than 1, 2, or 10.

Moreover, data correctly identified as "bad" by this check would generally be caught by other checks. We conclude that this check has little added utility.

For example, all of the flags for six consecutive identical values greater than or equal to ten come from the same site, 'PCT,' in San Francisco, CA (see Figure 2.13). On March 10, 2020, a count of 8,191 is repeated 8 times; these are likely erroneous values, but they would have also been flagged as "extreme" values by tests described in the next section. We strongly recommend manually reviewing data flagged by checks for repeated nonzero values, especially for low count sites.

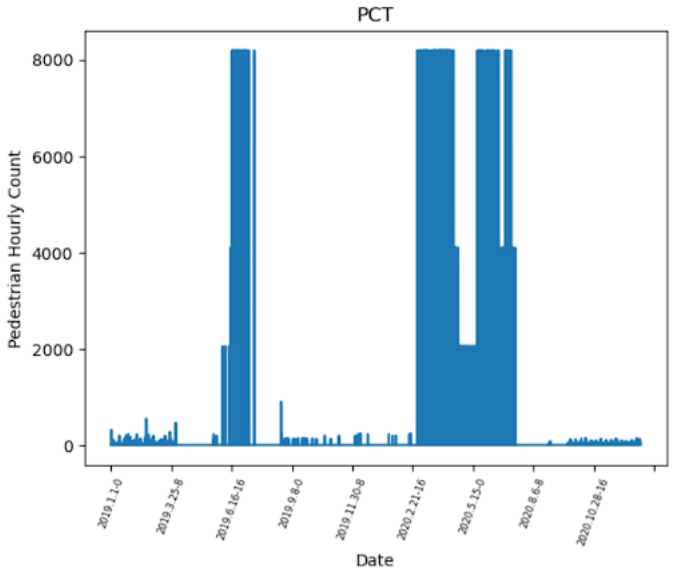

Figure 2.13: Pedestrian hourly count data from 2020 for site 'PCT' in San Francisco, CA.

2.5.2 Extreme Values

Pedestrian counts are generally higher than bicycle counts, leading to different choices for the parameters in the extreme value checks.

Night/Day Ratio

There are more days of pedestrian data for which the 3 a.m. count exceeds the 3 p.m. count and the 3 p.m. count is at least 3: 425 days across 37 sites versus 155 days across 30 sites. For the pedestrian data, the 3 a.m. counts ranged from 4 to 8,191, and the maximum ratio of 3 a.m. to 3 p.m. counts was 341.3. The number of days where the 3 a.m. count is greater than the 3 p.m. count and the 3 p.m. count is at least 10, 20, or 50, is 335, 256, and 137, respectively. The number of days with 3 a.m. to 3 p.m. ratios greater than 1.5, 2, or 3, is 330, 226, and 165, respectively.

We find that flagging site-days when 3 a.m. counts exceeds 3 p.m. counts generally does a good job of identifying bad data. For example, a site in Mill Valley, CA, had several days in November, 2021, with 3 a.m. counts above 2,500 and 3 p.m. counts below 300. That same site had a 3 a.m. count of 274 and 3 p.m. count of 48 on December 2, 2021. Figure 2.14 shows the November and December counts for this site. While some days at this site would also be flagged by other checks (e.g., hard limit), not all days would have.

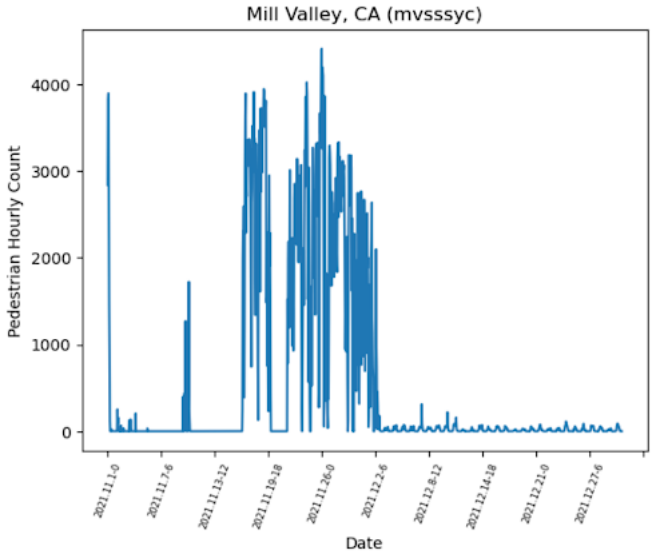

Figure 2.14: Pedestrian hourly counts from November 1, 2021 to December 30, 2021 at site 'mvsssyc' in Mill Valley, CA.

Hard Limits

Figure 2.15 shows the number of hours of pedestrian data with counts that exceed a threshold, as a function of the threshold. The number decreases substantially by around 2,000 pedestrians per hour. Of the 1,293,521 non-zero counts, 0.3% are greater than 2,000

and 0.1% are greater than 4,000. Flagging counts greater than 1,000 (appropriate for bicycle counts) flags 0.5% of hours.

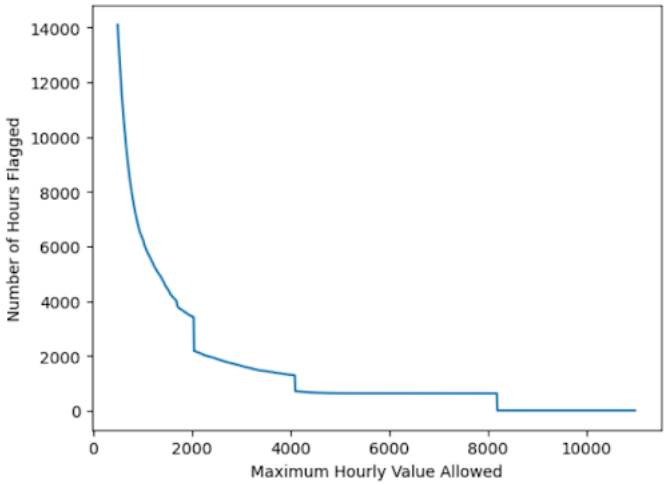

Figure 2.15: Number of hours of California pedestrian data with hourly counts greater than a given threshold, as a function of that threshold.

For example, 'CF2 West Crissy' in San Francisco, CA, has unusually high counts at the end of April and beginning of May 2020, shown in Figure 2.16. These data spikes, which range from more than 2,000 to less than 4,000, seem suspicious compared to the data at the beginning of April and end of May. These spikes indicate that during some hours a pedestrian was passing the counter more than every two seconds, which seems high during the COVID-19 pandemic when everyone was trying to keep six feet apart, even outside. If the data was legitimate we would also expect the eventual reduction to be more gradual.

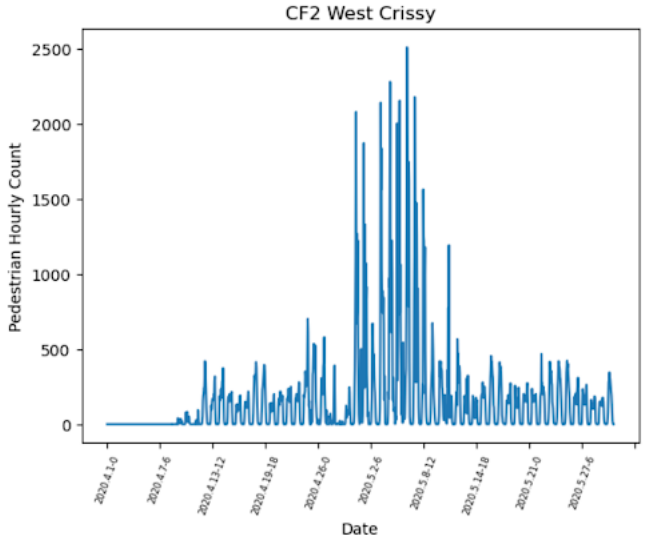

Figure 2.16: Pedestrian hourly counts for 'CF2 West Crissy' in San Francisco, CA, for April and May, 2020.

Flagging pedestrian counts greater than 2,000 makes sense in most places, especially because the incremental number of hours flagged compared to a threshold of 4,000 is small, and subsequent human review of flagged hours can reduce false positives.

We also consider thresholds for daily maxima. Figure 2.17 shows the number of days with total counts exceeding a threshold, as a function of the threshold.

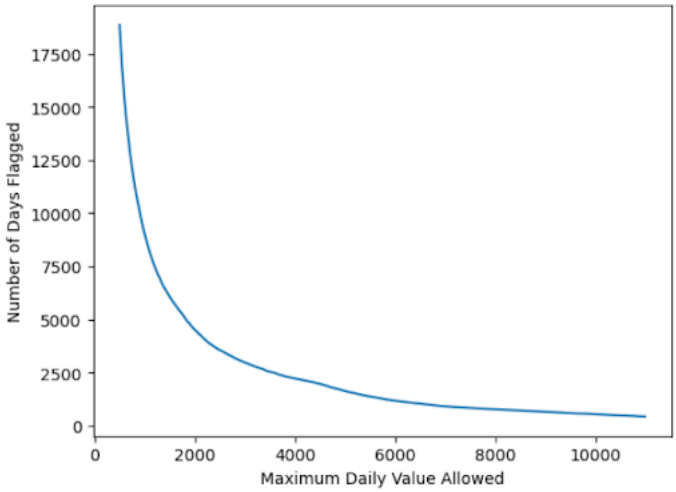

Figure 2.17: Number of site-days of California data with total pedestrian counts that exceed a threshold, as a function of that threshold.

The number changes more slowly once the threshold is about 2,000, and continues to drop. Of the 80,490 site-days with nonzero daily counts, 3% of the daily totals are greater than 4,000 and 1% are greater than 6,000. A threshold of about 6,000 seems appropriate for the California pedestrian data.

Adaptive upper bounds

We consider various adaptive upper bounds. As before, we use a 2 month rolling window applied to data after removing runs of 48 hours or more of consecutive zeros to calculate the mean, standard deviation, and percentiles. Results are summarized in Table 2.7.

Similar to the bicycle data, the adaptive upper bounds for the pedestrian data are very sensitive (although slightly less so than the upper bounds for the bicycle data) and tend to flag good data as bad. Upper bounds based on a rolling mean and standard deviation tend to flag fewer hours than those based on percentiles.

Next hour multiple

The number of hours flagged in the California pedestrian data for which one of the counts exceed a factor 5, 10, 15, 20, or 50 times the count in the previous hour, where the previous hours' count is at least 10 or 20, is summarized in Table 2.8.

As was the case for bicycle data, a factor of 5 results in a large number of false positives. While more data is flagged with this check for the pedestrian data than the bicycle data,

Upper Bound	$Q3 + 3 * IQR$	$Q3 + 5 * IQR$	$Mean + 2 * SD$	$Mean + 3 * SD$	$Mean + 5 * SD$
Hours Flagged	42,170	23,089	77,900	32,074	8,863
Hours Flagged, Upper Bound > 10	32,464	18,867	69,846	29,578	8,428
Hours Flagged, Upper Bound > 20	25,583	15,997	53,947	24,879	7,670
Hours Flagged, Upper Bound > 50	15,075	10,830	31,237	15,924	5,581

Table 2.7: Number of hours in which counts exceed various upper bound thresholds. Q3 is the upper quartile; Mean is the arithmetic mean; SD is the standard deviation; and IQR is the interquartile range. All four of these statistics are calculated using a rolling 2 month window after removing strings of 48 or more zero counts. Rows considering upper bound > 10, 20 and 50 do not flag data when the upper bound is less than or equal to 10, 20 and 50.

Factor	Hours flagged, previous count >10	Hours flagged, previous count >20
5	2,799	1,447
10	561	300
15	249	133
20	146	72
50	27	12

Table 2.8: The number of hours flagged in the California pedestrian data for which one of the counts exceed a factor times the count in the previous hour, where the previous hours' count is at least 10 or, and the factor is 5, 10, 15, 20, or 50.

we still find that this check with a multiplier of 10, conditioning on the previous hour being greater than 10, does a good job of flagging bad data.

2.6 Discussion

We examined data checks common in the literature and the instrumental failures they can detect. We tabulated the number of hours and days these checks flag as 'bad' in California bicycle and pedestrian data from 2019 to 2022 and investigated the data flagged as "bad" to determine whether the label was accurate or a false alarm. For the California data, we find it appropriate to flag the following occurrences:

- Hourly counts that are null, negative, or not integers
- Zeros that occur in runs of 48 hours or more in bicycle data, or in runs of 15 hours or more consecutive zeros in pedestrian data
- Site-days with runs of six or more identical hourly values greater than ten. These data should be reviewed manually to ensure the flag is not a false alarm
- Site-days with 3 a.m. counts that are greater than 3 p.m. counts, provided the 3 p.m. count is greater than two

- For bicycle and pedestrian data, site-days with at least one hourly count that is at least ten times larger than the previous hour's count, provided the previous hour's count is greater than 10
- Hourly counts greater than 1,000 bicycles or 2,000 pedestrians.
- Site-days with total counts greater than 4,000 bicycles or 6,000 pedestrians

Researchers should customize these checks for their sites by performing an analysis similar to that described above, and should manually review the data these checks identify

In general we find that it is better to have false positives than false negatives, because manual review can correct the false positives but there is no mechanism to catch and correct false negatives. Manual review of flagged data should include examining counter maintenance logs and searching for information about possible events (e.g., races, parades) or notices of road or park closures that might affect the counts. Keeping up-to-date maintenance logs is helpful for identifying bad data.

Once data are flagged as 'bad,' and manual review confirms the flags, researchers must decide what to do. When estimating aggregate quantities over different time periods (e.g., daily, monthly and yearly totals and means), provided "not too many" values are missing or flagged 'bad,' researchers might ignore or impute missing data.

If they plan to ignore missing data, then they will consider a day, week, or month "usable" if it has at least some threshold number (which can be zero) of "good" data. Higher thresholds will result in more data loss but potentially higher quality data. In our bicycle data, after filtering out data according to our checks, if we require at least 22, 23, or 24 valid hours in a day to count the daily data for a site, then we lose 0.2%, 0.2%, or 0.6% of our data (out of a total of 166,865 days). If we remove days with fewer than 22 hours of "good" data and require a month to have "good" data every day, the filters remove 851 (15%) of 5,816 months of data. If instead we require just 1 of each weekday for a given site and month, we filter out 177 months of data (3%).

An alternative strategy is to impute this missing data, provided not too many are missing (the larger the gaps, the less accurate imputed data are likely to be, and the more the estimates of aggregate quantities will depend on assumptions about the missing data). In future work, we will investigate various imputation methods and compare the accuracy of those methods for estimating aggregate quantities over different time spans, different types of sites (e.g., low/high volume, rural/urban), and with differing amounts of missing data.

Chapter 3

Look Who's Talking: Gender Differences in Academic Job Talks[1]

3.1 Introduction

Women are underrepresented among U.S. university faculty in Science, Technology, Engineering, and Mathematics (STEM). Why?

Gender bias pervades academia, including academic hiring processes [Moss-Racusin et al., 2012, Reuben et al., 2014], student evaluations of teaching [Boring et al., 2016], citation counts [Caplar et al., 2017], grant applications [Kaatz et al., 2014, Witteman et al., 2018], letters of recommendation [Schmader et al., 2007, Madera et al., 2009], credit for joint work [Sarsons, 2015], and the journal refereeing process [Budden et al., 2008]. Because of the prevalence of gender bias in so many areas of academia, it is important to understand where the bias is largest and most impactful in order to target gender equity efforts most effectively.

Some recent studies have concluded that audiences treat academic seminar (e.g., job talks, conference talks, departmental seminars) speakers differently depending on the speaker's gender [Blair-Loy et al., 2017, Davenport et al., 2014, Dupas et al., 2021].

Here, we examine whether female job applicants received more questions or spent more time responding to questions than male job applicants in five STEM departments between 2013–2019: Civil and Environmental Engineering (CEE), Electrical Engineering and Computer Science (EECS), Industrial Engineering and Operations Research (IEOR), Mechanical Engineering (ME), and Physics. Table 3.1 shows the proportion of female faculty and female interviewees in these departments. Presenters' self-identified genders were not available. We inferred gender from pronouns on the presenter's website (if available), name, and appearance. We did not infer that any presenter's gender was non-binary but our analysis is easily extended to include more gender categories.

Our study and analysis differ substantially from previous work. Our data are for a different institution, cover more STEM disciplines, and include more categories of questions and other interruptions. To address inter-rater reliability, at least three raters examined every

[1]This chapter comprises a publication [Glazer et al., 2023a] co-authored by Hubert Luo, Shivin Devgon, Catherine Wang, Xintong Yao, Steven Siwei Ye, Frances McQuarrie, Zelin Li, Adalie Palma, Qinqin Wan, Warren Gu, Avi Sen, Zihui Wang, Grace D. O'Connell, and Philip B. Stark.

talk, while other studies generally used only a single rater. One recent study [Dupas et al., 2021] found very large differences among raters, but concluded—based on an inappropriate use of the correlation coefficient—that those differences could be ignored.

Our data include whether each speaker was ultimately offered a faculty position, allowing us to examine the relationship between interruptions and successful applications. We also investigated department culture around asking questions during job seminars, which revealed differences across departments.

Previous work used parametric tests and were based on differences in means. We use nonparametric randomization tests based on differences in medians. The tests frame the scientific null hypothesis that "speaker gender does not matter" as the statistical null hypothesis that speaker gender is an arbitrary label that might as well have been randomly assigned (within each department). Medians represent what is "typical" for speakers of each gender, whereas means are sensitive to extreme values.

We generally find small gender differences in the medians, on the order of 0–4 questions, not all in the same direction. The differences are not statistically significant.

Whether the differences are statistically significant or not, it is implausible that differences so small have a material impact on whether a candidate is hired. Moreover, the data do not support the hypothesis that interruptions are always detrimental to the presenter: in some departments, candidates who were interrupted more often were more likely to be offered a position.

Our study was inspired by that of Blair-Loy et al. [2017], who examine a slightly smaller data set (119 talks in Engineering departments versus 156 talks in STEM departments in our study) and find gender differences comparable in magnitude to those we find—but conclude that those small differences are statistically significant.

Section 3.2 discusses our data and statistical methods. Section 3.3 presents the randomization test results. Section 3.4 examines differences between our study and previous work, presenting evidence (from simulations and experiments with negative controls) that the apparent statistical significance of the small effects found by Blair-Loy et al. [2017] results from using an inappropriate hypothesis test. It also explains how to calibrate parametric tests using randomization, to obtain genuine P-values in some situations where the parametric assumptions do not hold. Section 3.5 discusses the findings and limitations. Section 3.6 presents our conclusions.

3.2 Data and Methods

3.2.1 Data

Many UC Berkeley departments record academic job talks for tenured and tenure-track positions. We received Berkeley IRB approval to use such videos in this research.

We obtained videos from 2013–2019 for eight departments: civil and environmental engineering (CEE), electrical engineering and computer science (EECS), industrial engineering and operations research (IEOR), materials science and engineering (MSE), mechanical engineering (ME), nuclear engineering (NE), physics, and statistics. Not all the videos were adequate for our purpose (e.g., prior to 2018, the statistics department did not use an audi-

ence microphone: the audience voices were often unintelligible), and some departments had too few male or female applicants for any test to have much power: we omitted departments for which (# presenters) choose (# female presenters) is less than 20, because that makes it impossible to have a P-value less than 5%, no matter what the data are. That left CEE, EECS, IEOR, ME, and Physics. In all, 156 videos from the five departments were annotated.

Department	CEE	EECS	IEOR	ME	Physics
Female Faculty	25%	18%	30%	18%	12%
Female Pre-tenure Faculty	50%	31%	33%	25%	22%
Female applicant pool, 2015-2019	28%	22%	22%	22%	20%
Videos	31	65	8	35	17
Female interviewees	48%	34%	38%	40%	29%
Median events, female	9	11	23	16	7
Median events, male	9	10	24	20	8

Table 3.1: Percentage of faculty who are women and number of job talk videos for the five STEM departments in the study (the counts include lecturers and adjunct faculty but not emeriti). Median events refers to the median number of audience utterances (e.g., questions, comments). Pre-tenure faculty includes tenure-track assistant professors but neither lecturers nor adjunct faculty. Faculty counts and applicant pool data were obtained from the UC Berkeley Office for Faculty Equity and Welfare. Faculty full-time equivalent (FTE) data as of 4/30/2020.

3.2.2 Annotation Methodology

We developed a set of tags for audience interactions using an iterative process that involved eight raters tagging the same videos, then assessing inter-rater reliability. The category definitions were adjusted until all annotators agreed on the annotations across several videos. We tried to capture "tone" to the extent that it could be labeled consistently by different raters. We ended up with 13 categories, listed in Table 3.2. An annotation refers to each time a member of the audience spoke. A typical video might have 8–20 annotations (the number of annotations ranged from 2 to 57); variation across departments was substantial.

Each video was reviewed by three undergraduate researchers. Two students independently annotated each video; a third student resolved any discrepancies. Data quality is discussed in Appendix A.1.

3.2.3 Randomization (Permutation) Tests

We consider the null hypothesis that the gender of the presenter is not related to the number, duration, or nature of questions the audience asks, as if gender were an arbitrary label assigned at random to presenters. This hypothesis naturally leads to randomization tests.

We condition on the number of female and male presenters in each department and consider the distribution of test statistics under the null hypothesis. Conceptually, we imagine

Category	Definition
Begin	Time a distinct person starts speaking.
End	Time that person stops speaking.
Speaker	Whether the speaker was an audience member or host/other
Acknowledged	Presenter (or host) paused and either verbally or nonverbally recognized the speaker before the speaker spoke, e.g., "I see you have a question" or "yes?" If the speaker cut off the presenter or host, the speaker is unacknowledged.
Attempted Interruption	An audience member interrupted the presenter or host but the presenter or host continued without giving the audience member a chance to question or comment, or ignored the question or comment.
Follow-up	The question/comment came from the same person as the previous question/comment. If a new person asks a related question it is not a follow-up.
Scientific Comment	The audience member commented about the science, beyond providing context for a question.
Non-scientific Comment	The audience member made a comment that is not related to a scientific concept.
Positive Comment	The audience member made a positive comment (e.g., "very interesting work!").
Clarifying Question	A question about what the presenter did, how they did it or what it means (e.g. "what does that variable mean?", "How does this model work?"). Questions about the presenter's background, previous research, or approach to various problems are clarifying questions (e.g., "Can you describe the research you are working on with Professor X?" or "How would you teach this concept to others?").
Furthering Question	A question that bring in new concepts or information (e.g. "you mentioned X, have you considered Y?", "Do you have thoughts on the effect of Z on X?").
Critical	A question/comment that expresses skepticism, doubt, or concern about the validity of the work (e.g., "Are you sure that method works in this context?").
Ad hominem	A question/comment impugning the presenter's identity rather than addressing the presenter's work (e.g., "how could a woman be expected to understand this?", "only somebody who studied at Stanford would use that method").
Self-referential	An audience member makes a statement about themselves (e.g. "in my experience/work . . . ," "My work on X shows . . . ").

Table 3.2: Characteristics of utterances noted by raters, and their definitions.

randomly re-labeling presenters in such a way that each department keeps its observed numbers of female and male presenters, but the gender labels are "shuffled" across presenters. That induces a (null) probability distribution for any test statistic we might choose to examine (including the test statistic used by Blair-Loy et al. [2017], as discussed below). The probability that the test statistic is greater than or equal to the value observed for the original data, computed on the assumption that the null hypothesis is true, is a P-value for the null hypothesis.

In principle, the randomization distribution can be found exactly by enumerating all assignments of genders to presenters that keeps the total number of female and male presenters fixed. When there are many presenters and more than a few of each gender, it is impractical to enumerate all assignments. Instead, P-values can be constructed by assigning gender pseudorandomly B times, then basing the P-value on the distribution of the test statistic in that simulation. This can be viewed as a simulation approximation to the "true" P-value that would be obtained by examining all assignments or it can be viewed as an exact P-value for a randomized test [Dwass, 1957, Ramdas et al., 2023], if the P-value is computed as

$$\frac{(\#\text{ assignments for which the test statistic is as large or larger than observed}) + 1}{B + 1}. \quad (3.1)$$

In the latter approach, the smallest attainable P-value is $1/(B + 1)$.

It is important to select the test statistic before examining the data, to prevent "P-hacking." We chose to use the difference in the median number of questions asked of female and male presenters as the test statistic, primarily for two reasons. First, we are interested in "typical" behavior, which the median measures but the mean does not. Second, in our experience, the total number of questions varies considerably; we did not want the results to be driven by a small number of talks that generated unusually many questions. Note that there is more than one definition of the median. We use the "smallest" median: the smallest number that is greater than or equal to at least 50% of the observations.

As described in Section 3.4 and Appendix A.2, we also used the test statistic adopted by Blair-Loy et al. [2017], namely, the gender coefficient in a ZINB regression of the number of questions on covariates that included the presenter's gender and the percentage of faculty in the department who are female.

Blair-Loy et al. [2017] examined the pre-Q&A portion of job talks but not the Q&A portion: they hypothesized that presenters are injured by questions (interruptions) during the pre-Q&A period because it takes time away from their exposition. We analyzed pre-Q&A questions and other interruptions to compare with their results. However, we also analyzed entire talks, including the Q&A period, to examine whether male and female presenters are treated differently overall. Because pre-Q&A questions are relatively rare, restricting attention to the pre-Q&A period would have limited our ability to detect differences.

Our primary analysis kept departments separate because departments have different customs and etiquette for asking questions and interrupting presenters. We did not stratify by year. Stratifying by year might reduce the possibility of Simpson's Paradox affecting the results, for instance, if the percentage of presenters who are female varies substantially from year to year and department practices also change. However, stratifying by year might

also decrease power because there are relatively few female applicants and relatively few applicants in all annually.

The randomization tests work as follows:

1. For each category, calculate the test statistic for the original data.

2. Randomly reassign the presenter gender labels $B = 10,000$ times, holding constant the number of female and male labels. For each assignment, recalculate the test statistic for each category.

3. Calculate the P-value for each category as in (3.1).

In addition, we used nonparametric combination of tests (NPC) [Pesarin and Salmaso, 2010] to combine categories into a single multivariate randomization test (see below).

We performed the following randomization tests:

- One-sided randomization test using the difference in the median number of *acknowledged* questions between male and female presenters

- One-sided randomization test using the difference in the median number of *unacknowledged* questions (i.e., interruptions) between male and female presenters

- One-sided randomization test using the difference in the median number of *attempted interruptions* between male and female presenters

- One-sided randomization test using the difference in the median *time* spent on audience questions/comments between male and female presenters

- Nonparametric combination of tests combining all 13 categories, where the individual tests were one-sided for the four variables mentioned above and two-sided for the other 9 categories.

We use 1-sided tests for the four primary categories because previous research suggests that women receive more questions, are interrupted more often, and spend more time answering questions than men [Blair-Loy et al., 2017, Davenport et al., 2014, Dupas et al., 2021]. We also combine all 13 categories into a single, omnibus test using the nonparametric combination of tests method (NPC) [Pesarin and Salmaso, 2010], a general method for creating multivariate tests by combining univariate permutation tests. The test statistic for the multivariate test is calculated by applying a *combining function* to the P-values from the univariate tests.

The randomization distribution of that combination of P-values under the null hypothesis is used to calibrate the omnibus test, as follows:

1. Create B randomized versions of the dataset by randomly reassigning the presenter gender labels B times, yielding a total of $B + 1$ datasets, including the original.

2. For each of the $B + 1$ datasets, calculate the test statistic for each of the 13 categories (the difference in medians for that category between male and female presenters).

3. Then, for each dataset, replace the value of the test statistic for the jth category with the fraction of values (across the $B+1$ versions of the dataset) for which the jth test statistic is greater than or equal to the value of the test statistic for that dataset (for two-sided tests, the test statistic is the absolute value of the "raw" difference). This replaces each observed value of the test statistic by its corresponding P-value. That gives $B+1$ 13-vectors; the components of each 13-vector are numbers between 0 and 1.

4. Apply Fisher's combining function ($-2\sum_j \ln P_j$) to each of the $B+1$ 13-vectors to get the NPC test statistic for each dataset.

5. The overall P-value is the fraction of the NPC test statistics (among the $B+1$ values) that are greater than or equal to the NPC statistic for the original dataset.

In total, five tests (the four one-sided randomization tests and the NPC test) were performed on 4 subsets of each department's data—all presenters or pre-tenure presenters, and the entire talk or pre-Q&A portion of the talk—a total of 20 tests. We adjusted for multiplicity using the Holm-Bonferroni correction, but nothing was statistically significant even before the adjustment.

3.3 Randomization Test Results

Descriptive statistics do not illuminate differences in the number or nature of questions asked to female versus male presenters. Figure 3.1 shows the distribution of acknowledged, unacknowledged, and follow up questions asked broken down by department and gender.

Here we present our results based on randomization tests. These findings were compared to an analysis using the parametric ZINB method of Blair-Loy et al. [2017], which was calibrated parametrically and nonparametrically using randomization (Section 3.4 and Appendix A.2).

We considered the entire talk (pre- and post-Q&A), pre-Q&A by itself, all presenters, and only pre-tenure presenters: four analyses in all. Table 3.3 show entire talk results for each department and all presenters. Results for pre-Q&A and only pre-tenure presenters were qualitatively the same and are presented in the Appendix.

Some differences were positive (women received more questions of a given type than men) and some negative; most were zero. The smallest P-value was 0.14, for unacknowledged questions in ME. Almost half of the 65 P-values for individual categories were equal to 1; only 3 are below 0.2 (in EECS and ME). The non-parametric combination of tests yields a combined P-value of 1 for all departments. In summary, the statistical evidence that audience members interact with female and male presenters differently is weak.

3.4 Comparison with Previous Studies

We find relatively small differences in the median number of questions between female and male presenters: the difference was 0 for most types of questions and comments, but depending on the department and the type of question or comment, the difference ranged

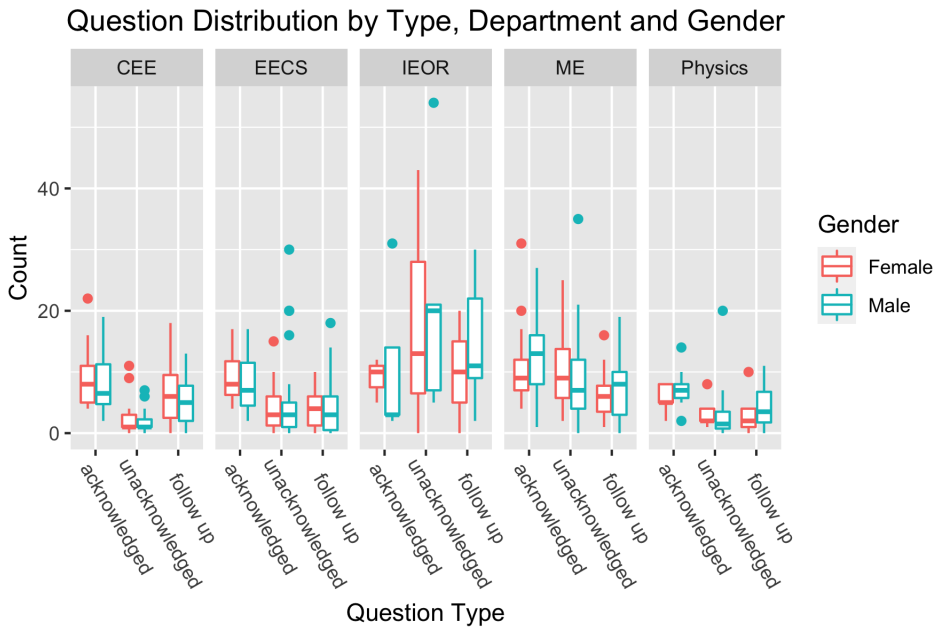

Figure 3.1: Box plots showing the distribution of acknowledged, unacknowledged and follow up questions asked broken down by department and gender. The box extends from the 25th to 75th percentile with a line for the median in between. The whiskers extend from the minimum to maximum values, with points plotted above or below if they are outside of 1.5 times the interquartile range.

from -7 (women received fewer questions) to 7 (women received more questions). Median time for questions was 0–2 minutes more for women than for men. The randomization tests do not find any of these differences to be statistically significant.

We are aware of three studies related to ours: Blair-Loy et al. [2017], Davenport et al. [2014], and Dupas et al. [2021]. Blair-Loy et al. [2017] is the closest to ours: the other studies primarily look at non-job talks (or do not look at job talks at all), annotated talks "live" while the talk was underway, and examine talks in other disciplines (Economics and Astronomy versus Engineering and Physics).

Davenport et al. [2014] is an "informal report" based on 225 talks at the 223rd Meeting of the American Astronomical Society (AAS), held in January 2014. Attendees of AAS were asked to recall and report the number of questions asked in talks they attended through an online form. The online form was advertised to the attendees via email, social media, and blogs. Attendees were not given any training on recording information about the talks, and the report does not analyze the reliability of annotations. The mean number of questions for female presenters was 3.28 (SE 0.20) and for male presenters was 2.64 (SE 0.12). It is not clear how to assess whether the difference, 0.64, is meaningful or statistically significant. Our best understanding is that the reporters were self-selected; not every talk was included and observers had no training. We are not aware of any study of the accuracy of recall data (by untrained observers) in this context: the analysis consolidated multiple observations of a single talk on the assumption that the highest number was correct.

Dupas et al. [2021] analyzed 462 Economics seminars at 32 institutions and from 84

	CEE	EECS	IEOR	ME	Physics
Time on Questions (in seconds)	43	37	117	31	0
	0.36	*0.15*	*0.50*	*0.33*	*0.64*
Acknowledged Question	2	1	7	-5	-2
	0.29	*0.30*	*0.37*	*0.97*	*1*
Unacknowledged Question	0	0	-7	2	1
	0.83	*0.59*	*0.65*	*0.14*	*0.35*
Attempted Interruption	0	0	-2	0	0
	1	*1*	*0.65*	*0.85*	*0.67*
Follow-up Question	1	1	-1	-2	-1
	0.49	*0.35*	*0.65*	*0.85*	*0.77*
Scientific Comment	0	0	-3	2	0
	1	*1*	*0.57*	*0.17*	*1*
Non Scientific Comment	0	0	-2	1	0
	1	*1*	*0.59*	*0.55*	*1*
Positive Comment	0	0	0	0	0
	1	*1*	*1*	*1*	*1*
Clarifying Question	3	3	-5	-5	-3
	0.36	*0.18*	*0.72*	*0.29*	*0.32*
Furthering Question	1	0	4	0	2
	1	*1*	*0.38*	*1*	*0.63*
Critical Element	0	0	-1	0	-1
	1	*1*	*1*	*1*	*0.65*
Ad Hominem	0	0	0	0	0
	1	*1*	*1*	*1*	*1*
Self Referential	0	0	0	0	0
	1	*1*	*1*	*1*	*1*

Table 3.3: For each department, difference in medians (female−male) for each category of audience utterance, for entire talks (pre- and post-Q&A), for all applicants (non-tenured and tenured). P-values for permutation tests are italicized in the second row for each question category.

seminar series. Of those, 176 talks (38%) were job talks. The talks were annotated by 77 graduate students from many institutions. It is not clear whether the annotators received any training. Most talks were annotated by a single student; a small percentage were annotated by two students. Annotators recorded the start and end time of each interaction, information about who asked the question (e.g., male or female, professor or student), and whether the question was answered, deferred, ignored, or interrupted. Qualitative data were also collected about the type and tone of question. Coding the tone was optional and most annotators chose not to report tone. The data was analyzed using a large number of linear regressions, regressing the outcome (e.g., number of questions) on a subset of presenter gender, a vector of talk level controls (dummy variables for official seminar duration in minutes and whether the seminar is internal (presenter is from institution hosting the seminar)), seminar series

fixed effects, coder fixed effects, home institution group fixed effects, and paper JEL fixed effects. The regression was weighted by the inverse of number of coders recording a given talk. The paper does not mention multiplicity adjustments, despite the fact that at least eight models were fit using four different treatments of clustered standard errors, along with dozens of other tests and regression models. (We estimate that the analysis includes hundreds of combinations of models and assumptions about errors.) Dupas et al. [2021] conclude that women are asked 3.5 more questions than men on average.

Blair-Loy et al. [2017] examined 119 videos from two years of job talks in Computer Science and Electrical Engineering departments at two highly ranked R1 universities and the Mechanical Engineering department at one of those universities. They do not mention the years the talks were recorded. Not every video was annotated; they annotated videos of all female presenters ($N = 41$) and a sample of male presenters ($N = 78$) (matched to the female presenters by years from PhD).

They found small differences comparable to those we found: women had an average of 1.18 (SE 1.04) more unacknowledged interruptions than men, 0.097 (SE 0.89) fewer acknowledged questions, 1.83 (SE 1.09) more follow up questions, 2.91 (SE 2.40) more total questions, and 0.012 (SE 0.0065) proportion of the time more on audience questions. The t-statistics for the individual Blair-Loy et al. [2017] estimates are 1.8 or below: formally, the differences are not statistically significant, even before adjusting for multiplicity. Thus, our data and theirs agree in broad brush.

However, we disagree over the statistical and practical significance of the (generally) small observed differences. Blair-Loy et al. [2017] find the gender differences to be statistically significant—but not on the basis of the t-statistics. Instead, they introduce an ungrounded parametric model for audience questions: zero-inflated negative binomial (ZINB), which they fit to the data by regression. They find the gender coefficient in a ZINB model to differ significantly from zero at significance level 0.05 for follow-up questions and at level 0.1 for total questions. Appendix A.2 discusses differences in more detail, including differences in the data collection and the statistical analysis. It applies their parametric analysis to our data and shows that randomization P-values for the same test statistic are substantially larger, and that the parametric test may produce the spurious appearance of statistical significance.

3.5 Discussion

3.5.1 Are interruptions bad?

So far we have considered whether there are gender differences in how audiences treat speakers. Generally, the observed differences are small and neither material nor statistically significant. However, we might also wonder whether asking women more questions than men disadvantages women at all. Blair-Loy et al. [2017] suggest that women are disadvantaged by frequent audience questions.

Our study is observational, not a controlled experiment; it is hard to draw reliable causal inferences from observational data. However, our data suggest that (at least in some departments) questions reflect genuine interest in the talk: departments that spent more time

asking female presenters questions also hired women more frequently during that time period. Table 3.1 shows that the proportion of female pre-tenure faculty in CEE, EECS, and IEOR is higher than the proportion of women in their applicant pools. These departments also spent more time questioning women than men. On the other hand, women and men spent equal time on questions in Physics, which hires women roughly in proportion to their representation in the applicant pool.

Furthermore, in CEE, faculty presenters who received offers generally were asked more questions during their talk than presenters who did not receive offers, which is consistent with the chair's description of departmental culture (see Section 3.5.2). The median number of questions asked of presenters who received offers was larger than the median for presenters who did not receive offers by 2 acknowledged questions and 1 unacknowledged question. While more study is needed, these descriptive statistics suggest that , at least in CEE, candidates who receive more questions may be treated more favorably—not less favorably—in hiring decisions.

In summary, questions and interruptions could signal many different things, including:

- audience interest, curiosity, engagement, or excitement
- audience confusion, related to the audience's familiarity with the material
- audience confusion, related to the quality or clarity of the exposition
- disrespect, hostility, or harassment

Our data suggest that all four of these things happen, depending on departmental culture.

3.5.2 Departmental Culture

Descriptive statistics and our randomization test analysis indicated substantial differences in the way departments tend to act towards speakers—regardless of the speaker's gender. For example, the median number of audience utterances was 9 for both male and female presenters in CEE whereas it was 23 and 24, respectively, for IEOR. In IEOR, talks by 3 of the 8 speakers (all men) had no formal Q&A period, but had many pre-Q&A questions.

For the Engineering departments, women generally spent more time on questions and were asked more questions (except in ME), however these differences are small and not statistically significant. In Physics, male presenters generally spent more time on questions and comments and were asked more questions than female presenters, although the differences are not statistically significant.

We asked the department chairs to describe the general department question etiquette.

In CEE, the audience is encouraged to hold questions until the end of the talk. Audiences are generally courteous, but questions at the end of the talk are encouraged and better talks typically stimulate more questions.

In EECS, etiquette is evolving. For the years included in our analysis departmental culture embraced interrupting speakers during their talk.

In IEOR, questions are frequently asked during the talks. The culture condones asking questions and interrupting, especially if the question is clarifying.

In ME, questions are generally asked during talks. If there are too many questions in a row, the moderator might ask the audience to hold their questions.

In physics, questions are encouraged and it is common to interrupt the speaker with clarifying questions during the talk. However, many audience members hold other kinds of questions until the end.

3.5.3 Leaky Pipeline

Blair-Loy et al. [2017] suggest that differences in audience interactions during academic job talks exemplify the "leaky pipeline." But as Dupas et al. [2021] points out, there is a difference between disparate *treatment*, i.e., whether the audience interacts differently with female versus male presenters, and disparate *impact*, i.e., whether job outcomes are different for equally qualified female and male applicants. We do not find gender based differences in our academic job talks (disparate treatment). We also note that all of the departments we analyzed interviewed a greater proportion of women than the proportion of women in their applicant pool. However, for some departments, the proportion of interviewees who are women is much larger than the proportion of pre-tenure faculty who are women. This indicates potential bias in making job offers (disparate impact), yield, or retention of junior female faculty; we do not examine the issue here.

3.5.4 Limitations

Presenter gender self-identification was not available to us, so we had to infer gender based on name, appearance and pronouns on presenter website (if available). We did not infer any of the presenters to be non-binary, so we were not able to analyze differences in audience interactions with non-binary presenters.

Some departments had video quality that was very poor, e.g., Statistics, that we were ultimately unable to use in our analysis.

Ideally we would have liked to have additionally stratified our analysis by year, but we were unable to do this due to small sample size. Therefore, we were unable to account for whether a department implemented bias training or specifically tried to diversify the faculty hiring process.

It is unclear to us what magnitude of difference is important. For example, is a difference of one question a material difference? We do not believe the median differences observed in this study are material differences.

3.6 Conclusion

Neither our main analysis (randomization tests with difference in the median number of questions asked of female and male presenters as the test statistic) nor our nonparametric calibration of a parametric test finds material or statistically significant differences in audience interaction with female versus male presenters (P-value≥ 0.1).

Of course, women are discriminated against in other ways. Previous studies have shown that women and faculty from under-represented minority groups face conscious and uncon-

scious biases in STEM and academia [Boring et al., 2016, Brunsma et al., 2017, Caplar et al., 2017, Ozgumus et al., 2020, Schmader et al., 2007].

It is clear that commitment and leadership can bring large changes in gender equity in hiring in a relatively short period of time: three years after instituting systematic changes to recognize and value contributions to community engagement, fully 50% of the faculty hired by the College of Engineering were women.

Moreover, hiring is not the end of the story. For example, relying on student evaluations of teaching for employment decisions disadvantages women and other groups protected by employment law [Boring et al., 2016]. Universities must also pay attention to mentoring, assessment, and promotion to ensure that everyone is supported and evaluated fairly.

Chapter 4

Fast Exact/Conservative Monte Carlo Confidence Intervals[1]

4.1 Introduction

While it is widely thought that tests based on Monte Carlo methods are approximate, it has been known for almost 90 years that Monte Carlo methods can be used to construct exact or conservative tests.[2] In particular, P-values can be defined in a conservative way for:[3]

- *simulation tests*, which sample from the (known) null distribution or from a 'proposal distribution' with a known relationship to the null distribution [Barnard, 1963, Birnbaum, 1974, Bølviken and Skovlund, 1996, Davison and Hinkley, 1997, Foutz, 1980, Harrison, 2012]
- *random permutation tests*, which sample from the orbit of the observed data under the action of a group of transformations, a group under which the null distribution is invariant [Davison and Hinkley, 1997, Dwass, 1957, Hemerik and Goeman, 2018, 2021, Pitman, 1937, Ramdas et al., 2023]
- *randomization tests*, which sample re-randomizations of the observed data. [Hemerik and Goeman, 2021, Kempthorne and Doerfler, 1969, Ramdas et al., 2023, Zhang and Zhao, 2023]

Simulation tests require the null distribution to be known. Random permutation tests require that the null satisfy a group invariance; they sample from the null conditional on the orbit of the observed data under that group. Randomization tests also condition on the set of observed values and require that the data arise from (or be modeled as arising from)

[1] This chapter comprises a paper co-authored by Philip B. Stark. A pre-print is available at `https://arxiv.org/abs/2405.05238`.

[2] A nominal significance level α hypothesis test is *exact* if the chance it rejects the null when the null is true is α. It is *conservative* if the chance is at most α. A P-value P for a null hypothesis is *exact* if, when the null hypothesis is true, $\mathbb{P}\{P \le p\} = p$ for all $p \in [0,1]$; it is *conservative* if, when the null hypothesis is true, $\mathbb{P}\{P \le p\} \le p$ for all $p \in [0,1]$.

[3] This terminology/taxonomy is not universal and usage has changed over time; see, e.g., Hemerik [2024].

randomizing units into 'treatments' following a known design. Some tests can be thought of as either random permutation tests or randomization tests.

It has also been noted—but is not widely recognized—that to construct confidence sets by inverting tests,[4] the same Monte Carlo simulations can be used to to test every null [Harrison, 2012]. Putting these two ideas together gives a computationally efficient, easily understood procedure for constructing exact or conservative confidence sets.[5] When the simulation P-value is quasiconcave in the hypothesized value of a real-valued parameter, exact or conservative confidence intervals for the parameter can be found efficiently using modified bisection searches.

Surprisingly (at least to us), many texts that focus on permutation tests do not mention that there are exact random permutation tests. Instead they treat random permutation P-values as an approximation to the non-randomized "full group" P-value corresponding to the entire orbit of the observed data under the group [Good, 2006, Higgins, 2004, Salmaso and Pesarin, 2010]. That point of view has led to methods for inverting permutation tests that are both computationally inefficient and approximate [Bardelli, 2016, Garthwaite, 1996, Garthwaite and Jones, 2009, Pagano and Tritchler, 1983, Tritchler, 1984] even though there are more efficient, exact or conservative methods, as described below.

Consider the two-sample shift problem [Lehmann et al., 2005], in which a fixed set of units are randomly allocated to control or treatment by simple random sampling. Under the null hypothesis that $\eta = \theta \in \Re$, if a unit is assigned to treatment, its response differs by η from the response it would have if assigned to control.[6] Suppose we use the mean response of the treatment group minus the mean response of the control group as the test statistic for hypotheses about the shift θ. Many published numerical methods for finding confidence sets for θ from random permutation tests incorrectly assume that the P-value is a continuous function of the shift and crosses α at exactly one value of the shift (for one-sided intervals) or two values (for two-sided intervals) [Bardelli, 2016, Garthwaite, 1996, Garthwaite and Jones, 2009, Pagano and Tritchler, 1983, Tritchler, 1984]. In fact, when the test is the difference in sample means, the P-value is a step function of the hypothesized shift, as illustrated in Figure 4.1.

The quasiconcavity makes it straightforward to find the endpoints of a conservative confidence interval for the shift to any desired precision, for instance, by using bisection in a conservative way; see Section 4.5.

The values of some test statistics for different permutations of the data and for different hypothesized values of the parameter have a simple relationship that makes even more computational savings possible, obviating the need to compute the test statistic from scratch for each Monte Carlo iteration for each hypothesized value of the parameter.

This chapter is organized as follows. Section 4.2 gives an overview of randomized tests, P-values, and the duality between tests and confidence sets, highlighting the fact that *a*

[4] See Theorem 1.

[5] A nominal confidence level $1 - \alpha$ confidence procedure is *exact* if the chance it produces a set that contains the true value of the parameter is $1 - \alpha$; it is *conservative* if the chance is at least $1 - \alpha$.

[6] Caughey et al. [2017] show that tests for a shift that is assumed to be the same for all units have an interpretation that does not require that assumption: they are tests for the maximum or minimum shift for all units. Moreover, calculations involving constant shifts can be used to make confidence bounds for percentiles of a shift that may differ across units.

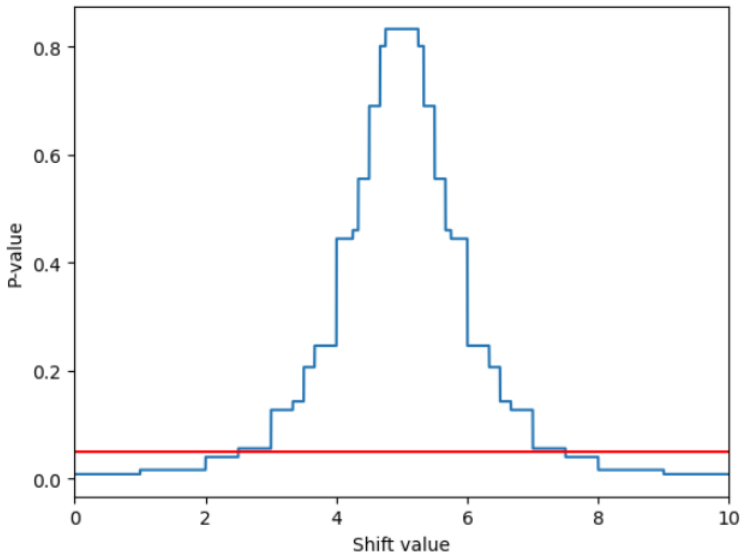

Figure 4.1: Exemplar (full-group) permutation P-value for the two-sided, two-sample shift problem as a function of the hypothesized shift, using the difference in sample means as the test statistic. The two samples are $\{5, 6, 7, 8, 9\}$ and $\{0, 1, 2, 3, 4\}$. The P-value is piecewise constant, discontinuous, and quasiconcave. Many algorithms for inverting permutation tests to form confidence sets incorrectly assume that the P-value is continuous in the parameter η and equal to α at exactly two values of η. In the general case, the two-sided P-value for the two-sample shift problem using the difference in sample means (i) is equal to α on two intervals, (ii) is equal to α on one interval and jumps through α at a discontinuity, or (iii) jumps through α at two discontinuities. The red horizontal line intersects the plot of the P-value as a function of the shift at values where the P-value equals 0.05. In this example, the P-value jumps through 0.05 at two discontinuities.

confidence set derived by inverting hypothesis tests is conservative or exact if and only if the test of the true null is conservative or exact. Section 4.3 reviews some Monte Carlo tests that are exact or conservative despite relying on simulation. The general approach to computing confidence sets developed below can be used with any of them. Section 4.4 describes how to construct confidence intervals for real-valued parameters when the P-value is quasiconcave in the hypothesized value of the parameter, using the bisection method with a slight modification. Section 4.5 improves that result for two problems with additional structure: finding a confidence interval for the shift in one-sample and two-sample problems with real-valued data. Section 4.6 compares the run time and confidence bounds for the new approach with those of some other methods. Section 4.7 discusses extensions and provides additional context.

4.2 Randomized tests

Suppose we have a family of probability distributions $\{\mathbb{P}_\eta : \eta \in \Theta\}$ on the measurable space $\mathcal{X}$, indexed by the abstract parameter $\eta \in \Theta$. We will observe data $X \sim \mathbb{P}_\theta$ for some

$\theta \in \Theta$. We want to make tests and confidence sets for θ using X, possibly relying in addition on an independent, auxiliary random variable U.

We treat the auxiliary randomness U abstractly: it is not necessarily a uniformly distributed real-valued random variable, as is common in the Neyman-Pearson framework. In the examples below, U comprises randomness arising from Monte Carlo sampling.

For each $\eta \in \Theta$, let $\phi_\eta(x,u) : (\mathcal{X},\mathcal{U}) \to \{0,1\}$ be the rejection function for a level α test of the hypothesis $\theta = \eta$: reject the hypothesis $\theta = \eta$ for data $X = x$ and auxiliary randomness $U = u$ iff $\phi_\eta(x,u) = 1$.[7] The test defined by ϕ_η has significance level α for the hypothesis $\theta = \eta$ iff

$$\mathbb{E}_\eta \phi_\eta(X,U) \le \alpha, \tag{4.1}$$

where the expectation is with respect to the joint distribution of X and U computed on the assumption that $\theta = \eta$. The test is exact if equality holds in inequality 4.1; otherwise, it is conservative.

The duality between tests and confidence sets establishes that the set of all $\eta \in \Theta$ for which the hypothesis $X \sim \mathbb{P}_\eta$ would not be rejected at level α is a $1-\alpha$ confidence set for θ:

Theorem 1 (see, e.g., Lehmann et al. [2005], § 3.5)**.** *For each $\phi \in \Theta$, let $\phi_\eta(x,u) : (\mathcal{X},\mathcal{U}) \to \{0,1\}$ be the rejection function for a test of the hypothesis $\theta = \eta$ at significance level α. Define $S(X,U) := \{\eta \in \Theta : \phi_\eta(X,U) = 0\}$. Then if the observed data $X = x$ and the auxiliary randomness $U = u$, $S(x,u)$ is a (possibly randomized) $1-\alpha$ confidence set for θ.*

Proof.

$$\mathbb{P}_\theta\{S(X,U) \ni \theta\} = \mathbb{P}_\theta\{\phi_\theta(X,U) < 1\} \ge 1 - \alpha, \tag{4.2}$$

where the probability is with respect to the joint distribution of $X \sim \mathbb{P}_\theta$ and U. □

Remark 1 The proof of Theorem 1 shows that the coverage probability of the confidence set rides entirely on the fact that the (single) test of the true null $\eta = \theta$ has level α. The tests involving other values of $\eta \in \Theta$ play no role whatsoever. In particular, the tests of different nulls do not need to be valid simultaneously; dependence among them does not matter; and a single U can be used for every test. Hence, if the Monte Carlo test of the *true* null $\theta = \eta$ is exact or conservative, a single set of simulations (i.e., a single realization of U) can be used to test every $\eta \in \Theta$.

Section 4.3.2 gives examples of exact and conservative Monte Carlo tests. Many of the tests we consider are derived from P-values, random variables with distributions that are stochastically dominated by the uniform distribution on $[0,1]$ when the null hypothesis is true: $P_\eta = P_\eta(X,U)$ is a P-value for the hypothesis $\theta = \eta$ iff

$$\mathbb{P}_\eta\{P_\eta(X,U) \le p\} \le p, \quad \forall p \in [0,1], \tag{4.3}$$

where the probability is with respect to the joint distribution X and U. If equality holds in inequality 4.3 for every p, the P-value is exact; otherwise, it is conservative. If P_η is a P-value, then (in the previous notation)

$$\phi_\eta(x,u) := \mathbf{1}_{P_\eta(x,u) \le \alpha}$$

[7] We assume that $\{\mathbb{P}_\eta\}_{\eta\in\Theta}$ have a common dominating measure and that ϕ_η is jointly measurable with respect to $\mathcal{X}$ and $\mathcal{U}$.

defines a test with significance level α.

In turn, many of the P-values we consider arise from a *test statistic* $T : \mathcal{X} \to \Re$. Monte Carlo simulation can be used to calibrate T to produce conservative or exact P-values, as discussed in the next section.

4.3 Exact and Conservative Monte Carlo Tests

We list exact or conservative versions of the three types of Monte Carlo tests listed above: simulation tests, random permutation tests, and randomization tests. All will depend on a test statistic $T : \mathcal{X} \to \Re$, and we consider larger values of T to be stronger evidence against the null, i.e., the P-value decreases monotonically with T. Most of the development does not require additional assumptions about T, but in section 4.5 we show that when T has special structure, additional savings are possible.

All three types of test use Monte Carlo to simulate new data sets and use the test statistic applied to each of those sets to find a P-value in various ways. In simulation tests, those sets are simulated directly from the null (possibly using importance sampling). In permutation tests, those sets are generated by applying randomly selected elements of a group to the original data. In randomization tests, those sets are generated by randomly re-allocating the data using the randomization design that was used to collect the original data.

We use Y_j to denote the jth simulated data set and n to denote the number of simulated data sets. The tests thus involve X, T, and $\{Y_j\}_{j=1}^n$.

4.3.1 Simulation tests

Consider the null hypothesis that $X \sim \mathbb{P}$, where $\mathbb{P}$ is a known distribution. Let $\{Y_j\}_{j=1}^n$ be IID $\mathbb{P}$. Then the following is a valid P-value [Barnard, 1963, Birnbaum, 1974, Bølviken and Skovlund, 1996, Foutz, 1980]:

$$P := \frac{1 + \sum_j 1\{T(Y_j) \geq T(X)\}}{1+n}. \tag{4.4}$$

The proof relies only on the fact that, if the null hypothesis is true, all rank orders of $\{T(X), T(Y_1), \ldots, T(Y_n)\}$ are equally likely. This result can be extended to sampling $\{Y_j\}$ from a known "proposal distribution" $\mathbb{Q}$ instead of sampling from the null distribution $\mathbb{P}$. Suppose that $\mathbb{P}$ and $\mathbb{Q}$ are absolutely continuous with respect to a dominating measure $\mathbb{F}$ and that $d\mathbb{P}/d\mathbb{F}(x) = 0$ whenever $d\mathbb{Q}/d\mathbb{F}(x) = 0$ (i.e., any set with strictly positive probability under $\mathbb{P}$ has strictly positive probability under $\mathbb{Q}$). Let $\{Y_j\}_{j=1}^n$ be an IID sample from $\mathbb{Q}$. Define the *importance weight*

$$w(x) := \frac{d\mathbb{P}/d\mathbb{F}}{d\mathbb{Q}/d\mathbb{F}}(x),$$

with the convention $0/0 := 0$. Then the following are valid P-values [Harrison, 2012]:

$$P := \frac{w(X) + \sum_{j=1}^n w(Y_j) 1\{T(Y_j) \geq T(X)\}}{1+n} \tag{4.5}$$

$$P := \frac{w(X) + \sum_{j=1}^{n} w(Y_j) 1\{T(Y_j) \geq T(X)\}}{w(X) + \sum_{j=1}^{n} w(Y_j)}. \tag{4.6}$$

In the special case $\mathbb{Q} = \mathbb{P}$, $w \equiv 1$, and both of these reduce to equation 4.4.

4.3.2 Random permutation tests

Consider the null hypothesis that the probability distribution $\mathbb{P}$ of the data $X \in \mathcal{X}$ is invariant under some group $\mathcal{G}$ of transformations on $\mathcal{X}$, so that $X \sim g(X)$ for all $g \in \mathcal{G}$.

Random permutation tests involve selecting elements of $\mathcal{G}$ at random. The sample can be selected in a number of ways: with or without replacement, with or without weights, from all of $\mathcal{G}$ or from a subgroup of $\mathcal{G}$; or a randomly selected element of $\mathcal{G}$ can be applied to a fixed subset of $\mathcal{G}$. Below are some valid P-values for those sampling approaches.

1. $\{\hat{g}_i\}_{j=1}^{n}$ is a uniform random sample (with or without replacement) of size n from $\mathcal{G}$ or a subgroup of $\mathcal{G}$. Let $Y_j := \hat{g}_j(X)$. Then

$$P := \frac{1 + \sum_{j=1}^{n} 1\{T(Y_j) \geq T(X)\}}{1 + n} \tag{4.7}$$

   is a valid P-value [Dwass, 1957, Davison and Hinkley, 1997, Ramdas et al., 2023].

2. $\{g_j\}_{j=1}^{n}$ is a fixed subset of n elements of $\mathcal{G}$, not necessarily a subgroup; $\hat{g}$ is drawn uniformly at random from $\mathcal{G}$; and $Y_j := g_j \hat{g}^{-1}(X)$.

$$P := \frac{\sum_{j=1}^{n} 1\{T(Y_j)) \geq T(X)\}}{n} \tag{4.8}$$

   is a valid P-value [Ramdas et al., 2023].

3. $\mathcal{G}$ is a finite group; $\hat{g}$ is selected at random from $\mathcal{G}$ with probability $p(g)$ of selecting $g \in \mathcal{G}$; and $Y_j := g_j \hat{g}^{-1}(X)$.

$$P := \sum_{j=1}^{|\mathcal{G}|} p(g_j) \cdot 1\{T(Y_j) \geq T(X)\} \tag{4.9}$$

   is a valid P-value [Ramdas et al., 2023].

Note that 4.7 has the same form as equation 4.4, but 4.7 in general involves sampling from the conditional distribution given the orbit of the observed data, while equations 4.4, 4.5, and 4.6 involve sampling from the unconditional distribution. The P-values in 4.8 and 4.9 involve drawing only one random permutation $\hat{g}$, then applying it to other elements of $\mathcal{G}$ to create $\{Y_j\}_{j=1}^{n}$.

4.3.3 Randomization tests

Unlike (random) permutation tests, which rely on the invariance of the distribution of X under $\mathcal{G}$ if the null is true, randomization tests rely on the fact that generating the data involved randomizing units to treatments (in reality or hypothetically).

Suppose there are N units, each of which is assigned to one of K possible treatments, $t = 1, \ldots K$. A *treatment assignment* r assigns a treatment to each unit: it is a mapping from $\{1, \ldots, N\}$ to $\{1, \ldots, K\}^N$. The assignment might use simple random sampling, Bernoulli sampling, blocking, or stratification; the assignment probabilities might depend on covariates. Let $\mathcal{R}$ be the set of treatment assignments the original randomization design might have produced, and let $p(r)$ be the probability that the original randomization made the assignment r, for $r \in \mathcal{R}$. Let r_0 denote the actual treatment assignment.

The test statistic T depends on the data X' and the (observed) treatment assignment r_0; we combine the two into a single random variable $X = (X', r_0)$ and write $T(X)$. Suppose we draw a weighted random sample of n treatment assignments $R_j \in \mathcal{R}$, $j = 1, \ldots, n$, with or without replacement, with probability $p(r)$ of selecting $r \in \mathcal{R}$ in the first draw (adjusting the selection probabilities appropriately in subsequent draws if the sample is without replacement). Let $Y_j := (X', R_j)$, $j = 0, \ldots, n$. Then

$$P := \frac{\sum_{j=0}^{n} p(R_j) \cdot 1\{T(Y_j) \geq T(X)\}}{\sum_{j=0}^{n} p(R_j)} \tag{4.10}$$

is a valid P-value. Note that 4.10 has the same form as 4.6, and when all $r \in \mathcal{R}$ are equally likely, 4.10 has the same form as 4.4 and 4.7.

Another valid P-value has the same form as 4.9 and involves a weighted sum over all possible assignments, just as 4.9 involves a weighted sum over all group elements. Let $Y_j := (X', r_j)$, $r_j \in \mathcal{R}$. Then

$$P := \sum_{j=1}^{|\mathcal{R}|} p(r_j) \cdot 1\{T(Y_j) \geq T(X)\} \tag{4.11}$$

is a valid P-value [Hemerik and Goeman, 2021, Kempthorne and Doerfler, 1969, Zhang and Zhao, 2023, Ramdas et al., 2023].

4.3.4 Tests about parameters

The tests above do not explicitly involve parameters. They can be extended to a family of tests, one for each possible parameter value $\eta \in \Theta$, in various ways.

For example, suppose there is a bijection $f_\eta : \mathcal{X} \to \mathcal{X}$ such that for any $\mathbb{P}_0$-measurable set A, $f_\eta(A)$ is $\mathbb{P}_\eta$-measurable and $\mathbb{P}_\eta(f_\eta(A)) = \mathbb{P}_0(A)$, and for any $\mathbb{P}_\eta$-measurable set A, $f_\eta^{-1}(A)$ is $\mathbb{P}_0$-measurable and $\mathbb{P}_\eta(A) = \mathbb{P}_0(f_\eta^{-1}(A))$. Then a P-value for the hypothesis $\theta = \eta$ can be calculated by using $T(f_\eta^{-1}(X))$ and $\{T(f_\eta^{-1}(Y_j))\}$ in place of of $T(X)$ and $\{T(Y_j)\}$ in 4.4, 4.7, or 4.8.

One common example is a location family with location parameter $\theta \in \Theta \subset \Re^m = \mathcal{X}$. For any $\mathbb{P}_\mathbf{0}$-measurable set $A \in \mathcal{X}$, $\mathbb{P}_\eta(A) := \mathbb{P}_\mathbf{0}(A - \eta)$. In other words, $f_\eta(x) = x + \eta$ and

$f_\eta^{-1}(x) = x - \eta$. A test of the hypothesis $\theta = \mathbf{0}$ can be used to test the hypothesis $\theta = \eta$ by applying the test to the transformed data $X - \eta$ and the transformed values $\{Y_j - \eta\}$.

For simulation test P-values using importance sampling (definitions 4.5 and 4.6) we can test the hypothesis $\theta = \eta$ by using the weights $w_\eta(x) := d\mathbb{P}_\eta / d\mathbb{Q}$; nothing else needs to be changed. In particular, the same sample can be used.

4.4 Confidence sets for scalar parameters

Remark 1, above, notes that it suffices to use a single set of Monte Carlo simulations to test the hypothesis $\theta = \eta$ for all $\eta \in \Theta$. As mentioned in section 4.3.4, how the simulations can be re-used depends on how the test depends on the parameter η. We use $P(\eta)$ to denote the P-value of the hypothesis $\theta = \eta$, $\eta \in \Theta$ for the family of tests under consideration. We now explore a simple case in more detail: confidence intervals for scalar parameters.

4.4.1 Confidence intervals when the P-value is quasiconcave

Suppose that the parameter θ is a scalar and that T is defined in such a way that $P(\eta)$ is quasiconcave in η, i.e., if η_1, $\eta_2 \in \Theta$ with $\eta_1 \le \eta_2$, then for all $\eta \in [\eta_1, \eta_2]$,

$$P(\eta) \ge \min\{P(\eta_1), P(\eta_2)\}. \tag{4.12}$$

(The two-sample test using the difference in sample means as the test statistic has this property.) Then the confidence set—all $\eta \in \Theta$ for which the hypothesis $\theta = \eta$ is not rejected at level α—is a connected interval $[\ell, u]$. The lower endpoint ℓ is the largest η such that $P(\zeta) \le \alpha$ for all $\eta \in (-\infty, \eta]$; the upper endpoint u is the smallest η such that $P(\zeta) \le \alpha$ for all $\eta \in [\eta, \infty)$.

Quasiconcavity makes it straightforward to find ℓ and u (or to approximate them conservatively to any desired precision) using the bisection method with a slight modification to account for the fact that $P(\eta)$ is not continuous in general (see, e.g., 4.1). Algorithm 1 returns a value in $[u, u+e]$ for any specified tolerance $e > 0$. An analogous algorithm returns a value in $[\ell - e, \ell]$.

Algorithm 1 Bisection algorithm for the upper endpoint u of a confidence interval when the P-value is quasiconcave in the parameter

1: Set $\alpha \in (0,1)$ and $e > 0$, the acceptable amount by which the returned upper endpoint exceeds the true upper endpoint. Let $P(\eta)$ be a valid P-value for the hypothesis $\theta = \eta$.
2: Find an interval $[L, U]$ that contains u: $P(L) \ge \alpha \ge P(U)$.[8]
3: While $|U - L| > e$:
   a: Set $m \leftarrow (L + U)/2$.
   b: If $P(m) > \alpha$, set $L \leftarrow m$, otherwise set $U \leftarrow m$.
4: Return U.

[8] A value of L can generally by found using an estimate of θ from the data X. Given a value of L, a value of U can be found by adding a sufficiently large value to L, since the P-value eventually decreases monotonically as η increases.

4.5 Additional efficiency in the one-sample and two-sample shift problems

For the most common test statistics, the computational cost of finding each P-value can be reduced further in the one-sample and two-sample problems by saving the treatment assignments (or in some cases, just a one-number summary of each assignment: see equation 4.14) and the value of the test statistic for each treatment assignment [Nguyen, 2009, Pitman, 1937]. This section explains how.

The one-sample problem. The one-sample problem is to find a confidence interval for the center of symmetry of a distribution that is symmetric about an unknown point θ from an IID sample $X = \{X_j\}_{j=1}^n$ from that distribution.

For any two N-vectors $\mathbf{x}$, $\mathbf{y}$, define the componentwise product $\mathbf{x} \odot \mathbf{y} := (x_1 y_1, \ldots, x_N y_N)$. If $\theta = \eta$, the distribution of $X_j - \eta$ is symmetric around 0, i.e., conditional on $|X_j - \eta|$, $X_j - \eta$ is equally likely to be $\pm(X_j - \eta)$.

A common test statistic for the hypothesis $\theta = \eta$ is the sum of the signed differences from η:

$$T_\eta(\boldsymbol{x}) := \sum_{j=1}^{N} (x_j - \eta). \tag{4.13}$$

Random permutation tests for the one-sample problem involve the distribution of the test statistic when the signs of $\{X_j - \eta\}$ are randomized independently. Let $\boldsymbol{\sigma} \in \{-1, 1\}^N$ be IID uniform random signs. Then we can test the hypothesis $\theta = \eta$ by comparing the value of $T_\eta(\mathbf{X})$ to the values of $T_\eta(\boldsymbol{\sigma} \odot \mathbf{X})$ for n randomly generated vectors $\boldsymbol{\sigma}$. As noted above in section 4.4, we can re-use the values of $\boldsymbol{\sigma}$ when testing $\theta = \eta$ for different values of η. But we can save even more computation by relating the value of $T_\eta(\boldsymbol{\sigma} \odot \mathbf{X})$ to the value of $T_0(\boldsymbol{\sigma} \odot \mathbf{X})$:

$$\begin{aligned} T_\eta(\boldsymbol{\sigma} \odot \mathbf{x}) &= \sum_{j=1}^{N} \sigma_j (x_j - \eta) \\ &= \sum_{j=1}^{N} \sigma_j x_j - \sum_j \sigma_j \eta \\ &= T_0(\boldsymbol{\sigma} \odot \mathbf{X}) - \eta \sum_j \sigma_j. \end{aligned} \tag{4.14}$$

Thus, for each configuration of signs $\boldsymbol{\sigma}$, we need only keep track of $T_0(\boldsymbol{\sigma} \odot \mathbf{X})$ and $\sum_j \sigma_j$ to determine the value of the test statistic for any other hypothesized value of η.

The two-sample problem. The two-sample problem involves a set of n units and two "conditions," treatment and control. Each unit will be assigned either to treatment or to control. If unit j is assigned to control, its response will be $y_j(0)$; if it is assigned to treatment,

its response will be $y_j(1)$.[9] The two-sample problem assumes that there is some $\theta \in \Re$ such that for all j, $y_j(1) - y_j(0) = \theta$.[10] Units are assigned to treatment or control by randomly allocating m units into treatment and the remaining $n - m$ to control by simple random sampling, with n and m fixed in advance. We seek a confidence interval for θ from the resulting data. Let $\{X_1, \ldots, X_m\}$ denote the responses of the units assigned to treatment and let $\{X_{m+1}, \ldots, X_n\}$ denote the responses of the units assigned to control.

A common test statistic is the difference between the mean response of the treatment group and the mean response of the control group, $\frac{1}{m}\sum_{j=1}^{m} X_j - \frac{1}{m-n}\sum_{j=m+1}^{n} X_j$, or its Studentized version [Wu and Ding, 2021]. Randomization tests for the two-sample problem involve the distribution of the test statistic when units are randomly allocated into treatment and control using the same randomization design the original experiment used.

If an allocation assigns n_{tc} of the units originally assigned to treatment to the control group (and vice-versa), then the difference between the test statistic when the shift is zero and the test statistic when the shift is η is

$$T_\eta(\mathbf{x}) - T_0(\mathbf{x}) = \eta \cdot n_{tc} \left(\frac{1}{n} + \frac{1}{m} \right). \tag{4.15}$$

This result is implicit in Pitman [1937]; it is straightforward to show. Suppose that for a particular allocation, the responses of the treatment group are $\{x_i\}_{i=1}^{n}$ and the responses of the control group are $\{y_i\}_{i=1}^{m}$. Then, the difference in means can be written

$$\frac{m\sum x_i - n\sum y_i}{nm}.$$

For each of the n_{tc} units that switch from treatment to control (and vice-versa) the change in the test statistic is $\frac{m\eta - n(-\eta)}{nm} = \eta(\frac{1}{n} + \frac{1}{m})$. If n_{tc} units change their treatment assignment, the total change in the test statistic is given by equation 4.15.

Thus, as we saw for the one-sample problem, for each allocation we only need to keep track of the test statistic for $\eta = 0$ and the number of units that moved from treatment to control to find the value of the test statistic for that allocation for any other value of η. Section 4.6 gives numerical examples.

4.6 Comparison to previous methods

4.6.1 Examples of extant methods

Many proposed methods for finding confidence bounds using random permutation or randomization tests implicitly or explicitly work with the full-group (or all-possible-allocations)

[9] This is the *the Neyman model for causal inference* [Splawa-Neyman et al., 1990, Imbens and Rubin, 2015]. The Neyman model implicitly assumes that each unit's response depends only on whether that unit is assigned to treatment or control, not on the assignment of other units. It also assumes that there are no hidden treatments that change the potential outcomes: if unit j is assigned to treatment the outcome would be $y_j(1)$ and if it is assigned to control the outcome would be $y_j(0)$. Together, these two assumptions are called the stable unit treatment value assumption (SUTVA).

[10] Caughey et al. [2017] show that the model can be used to draw inferences about percentiles of the effect of treatment even when the effect varies across units.

P-value $P(\eta)$ and assume it is continuous and crosses α at two points. But in practice, $P(\eta)$—whether defined using the full group (or full set of allocations) or only a subset of them—is typically a step function (see Figure 4.1).

Instead of using random permutations, Tritchler [1984] calculates the full-group P-value by taking advantage of the fact that the probability distribution of a sum is the convolution of the probability distributions, which becomes a product in the Fourier domain. The proof that their algorithm works relies on Theorem 1 of Hartigan [1969] regarding "typical values," which in turn assumes that the data have a continuous distribution.[11] Despite the fact that the method computes the distribution of the sum efficiently, it evaluates that distribution at such a large number of points that in practice it is limited to small problems.

In contrast, Garthwaite [1996] uses random permutations, applying the Robbins-Monro stochastic optimization algorithm [Robbins and Monro, 1951] to find the endpoints of the confidence interval. It is not guaranteed to give conservative or exact confidence intervals. The Robbins-Monro algorithm is a stochastic iterative method for approximating a root of a function of the expectation of a random variable by using realizations of the random variable. It assumes that the function is strictly monotone at the root, which is assumed to be unique.

The algorithm is sensitive to the starting point. Garthwaite and Jones [2009] propose averaging the results of the last n^* iterations of the Garthwaite [1996] algorithm to increase efficiency. The step size in the Garthwaite [1996] algorithm is a function of the significance level α, a constant $c > 0$, and the step number i: either $-c\alpha/i$ or $c(1-\alpha)/i$. This choice of step size can cause the algorithm to get stuck far from any root. O'Gorman [2014] show that the Garthwaite [1996] and Garthwaite and Jones [2009] algorithms can produce very different confidence bounds, even with the same starting point; they recommend averaging eight different runs with the same starting point. This increases the computational burden but still provides no guarantee that the algorithm finds the correct confidence bounds.

Bardelli [2016] use random permutations. They propose using the bisection algorithm to find the endpoints, but each step involves a new set of Monte Carlo simulations. That both increases the computational burden and makes the bisection potentially unstable, because sampling variability can make the simulation P-value non-monotonic in the shift. Indeed, re-evaluating the P-value at the same value of η with a different Monte Carlo sample will generally give a different result. Moreover, as mentioned previously, the usual bisection method relies on continuity, but the P-value is not typically continuous in the shift.

4.6.2 Numerical comparisons

The new method has time complexity $O(n)$ for the one-sample and two-sample problems. Our implementation uses the `cryptorandom` Python library, which provides a cryptographic quality PRNG based on the SHA256 cryptographic hash function (see section 4.7). That PRNG requires substantially more computation than the Mersenne Twister (MT), the default in R, Python, MATLAB, and many other languages and packages. For instance, generating

[11] Furthermore, as Bardelli [2016] note, there is as far as we know no public implementation of this method nor a clear and complete written description of it. We were not able to implement this method as described in Tritchler [1984]. Tritchler [1984] note that their algorithm is polynomial time, as is ours. However, based on what we understand from Tritchler [1984]'s description, we believe that the algorithm proposed here is substantially faster.

10^7 pseudo-random integers between 1 and 10^7 using cryptorandom takes 16s on our machine, in contrast to 0.17s for the `numpy` implementation of MT: a factor of about 100. We nonetheless advocate using the higher-quality PRNG, especially for large problems, for the reasons given in [Stark and Ottoboni, 2018].

Comparison to Tritchler [1984] We first compare results for the one-sample problem in Tritchler [1984] using Darwin's data on the differences in heights of 15 matched pairs of cross-fertilized and self-fertilized plants. Table 4.1 reports full-group confidence intervals generated by enumeration, the confidence intervals reported by Tritchler [1984], and the confidence intervals generated by our method using $n = 10^4$ and $e = 10^{-8}$. It took 0.5s of CPU time to generate the confidence intervals on an Apple MacBook Pro running macOS 12.3, with an M1 Max chip and 64GB of unified memory. Open-source python code that generated the intervals in column 4 is available at `https://github.com/akglazer/monte-carlo-ci`. We

confidence level	full-group	Tritchler	new method
90%	[3.75, 38.14]	[3.75, 38.14]	[3.857, 38.250]
95%	[-.167, 41.0]	[-.167, 41.0]	[-0.167, 41.0]
99%	[-9.5, 47.0]	[-9.5, 47.0]	[-8.80, 47.20]

Table 4.1: One-sample confidence intervals at 90%, 95%, and 99% computed using the full group of 2^{15} reflections, the intervals reported by Tritchler [1984], and the intervals computed using our proposed method with $n = 10^4$ and $e = 10^{-8}$, for data on 15 matched pairs of plants:
$49, -67, 8, 6, 16, 23, 28, 41, 14, 29, 56, 24, 75, 60, -48, 49$.

do not have access to source code for the Tritchler [1984] method[12] nor to a CDC 6600, so we are unable to compare run times of our method to those of Tritchler [1984]. However, Tritchler [1984] notes that execution time increases from 3.2s to 11.2s if ten more observations are included, about a 250% increase. On the other hand, the execution time for the new method (with $n = 10^4$) increased from 0.5s to 0.67s, about a 31% increase.

Next we consider the two-sample problem example in Tritchler [1984], which uses data reported in Snedecor and Cochran [1967] on the effect of sleep on the basal metabolism (measured in calories per square meter per hour) of 26 college women.[13] The data are listed in table 4.2. Full-group confidence intervals generated by enumeration, the confidence intervals reported by Tritchler [1984], and the conservative confidence intervals generated

[12]We requested two implementations, one in Fortran and one in C++, but were told neither still exists. (D. Tritchler, personal communication, October 2021; N. Schmitz, personal communication, December 2021.) We also attempted to implement the algorithm ourselves, but missing details made that impractical or impossible. See also Bardelli [2016].

[13]Snedecor and Cochran [1967] say that the data are from a 1940 Ph.D. dissertation at Iowa State University. This seems to be an example of hypothetical randomization versus actual randomization: although students presumably were not randomly assigned to sleep different amounts of time, the statistical analysis assumes that labeling a student as having 0–6 hours versus 7+ hours of sleep would amount to a random label if sleep had no effect on metabolism. If this is indeed an observational study, confounding is likely: students who sleep less than 7 hours differ from those who sleep more than 7 hours in ways other than just how long they sleep.

7+ hours of sleep	35.3, 35.9, 37.2, 33.0, 31.9, 33.7, 36.0, 35.0, 33.3, 33.6, 37.9, 35.6, 29.0, 33.7, 35.7
0–6 hours of sleep	32.5, 34.0, 34.4, 31.8, 35.0, 34.6, 33.5, 33.6, 31.5, 33.8, 34.6

Table 4.2: Basal metabolism (measured in calories per square meter per hour) and hours of sleep of 26 college women (source: Tritchler [1984]).

by our method using $n = 10^4$ and $e = 10^{-8}$ are given in table 4.3. Tritchler [1984] note that their FORTRAN implementation took 2.6s of CPU time on a CDC 6600, increasing to 14.5s (a 458% increase) when the number of observations was doubled. The new method took 0.60s to execute on our machine, increasing to 0.82s (a 37% increase) if the number of observations is doubled.

confidence level	full-group	Tritchler	new method
90%	[-2.114, 0.386]	[-2.114, 0.386]	[-2.117, 0.380]
95%	[-2.34, 0.650]	[-2.34, 0.650]	[-2.333, 0.643]
99%	[-2.814, 1.180]	[-2.814, 1.180]	[-2.850, 1.167]

Table 4.3: Two-sample confidence intervals at 90%, 95%, and 99% computed using the full group of $\binom{26}{11}$ reflections, the intervals reported by Tritchler [1984], and the intervals computed using our proposed method with $n = 10^4$ and $e = 10^{-8}$, for the basal metabolism data in table 4.2. The fourth column is based on a single seed value.

For larger sample sizes, Tritchler [1984]'s method quickly becomes infeasible, but the new method still runs relatively quickly. For example, for groups of $1,000$ units each, our new method takes approximately 22s to execute using the `cryptorandom` PRNG and approximately 5s for the `numpy.random.random` MT implementation, both using 10^4 samples and $e = 10^{-8}$.

Comparison to Garthwaite [1996] Garthwaite [1996] gives an example involving the effect of malarial infection on the stamina of the lizard *Sceloporis occidentali*. The data are listed in table 4.4. Garthwaite [1996] reports $[-0.30, 10.69]$ as the nominal 95% confidence

infected lizards (X)	16.4, 29.4, 37.1, 23.0, 24.1, 24.5, 16.4, 29.1, 36.7, 28.7, 30.2, 21.8, 37.1, 20.3, 28.3
uninfected lizards (Y)	22.2, 34.8, 42.1, 32.9, 26.4, 30.6, 32.9, 37.5, 18.4, 27.5, 45.5, 34.0, 45.5, 24.5, 28.7

Table 4.4: Distance in meters run in two minutes by infected and uninfected lizards (source: Samuels et al. [2003], as reported by Garthwaite [1996]).

interval based on 6000 steps of his algorithm; as mentioned above, that interval is approximate, not exact or conservative. Our implementation of the Garthwaite algorithm produced the approximate confidence interval $[-0.29, 11.0]$ in 0.72s, using 6000 steps and the initial

values of 0 and 10 for the lower and upper endpoints. For $n = 6000$ using the P-value defined in 4.7, the new algorithm yields the conservative confidence interval $(-0.20, 10.875)$ in 0.38s.

4.7 Discussion

This chapter presents an efficient method for constructing exact or conservative Monte Carlo confidence intervals from exact or conservative Monte Carlo tests. The method uses a single set of Monte Carlo samples. That solves two problems: it reduces the computational burden of testing all possible null values of the parameter, and it ensures that the problem of finding where the P-value crosses α is well posed, in the sense that it involves finding where a fixed function crosses a threshold rather than involving different realizations for each function evaluation. For problems with real-valued parameters, if the P-value is quasiconcave in the parameter, a minor modification of the bisection algorithm quickly finds conservative confidence bounds to any desired degree of precision. Additional computational savings are possible for common test statistics in the one-sample and two-sample problem by exploiting the relationship between values of the test statistics for different values of the parameter.

How many Monte Carlo samples?

The tests and confidence intervals we consider are exact or conservative regardless of how many or how few Monte Carlo samples n are used, so computation time can be reduced by using fewer samples. However, the highest attainable non-trivial confidence level is $n/(n+1)$. Increasing the number of samples also reduces the variability of results from seed to seed, and tends to approximate more precisely the bounds that would be obtained by examining all possible samples, permutations, or allocations.

The PRNG matters

Many common PRNGs—including the Mersenne Twister, the default PRNG in R, Python, MATLAB, SAS, and STATA (version 14 and later)—have state spaces that are too small for large problems [Stark and Ottoboni, 2018]. Depending on the problem size, they cannot even generate all samples or all permutations, much less generate them with equal probability. Linear congruential generators (LCGs) are especially limited; even a 128-bit LCG can generate only about 0.03% of the possible samples of size 25 from a set of 500 items. The Mersenne Twister cannot in principle generate all permutations of set of 2084 items, and can generate less than 2×10^{-6} of the possible samples of size 1,000 from 10^8 items. The choice of algorithms for generating samples or permutations from the PRNG also matters: a common algorithm for generating a sample—assign a pseudorandom number to each item, sort on that number, and take the first k items to be the sample—requires a much higher-quality PRNG than algorithms that generate the sample more directly [Stark and Ottoboni, 2018]. For large problems, a cryptographic quality PRNG may be needed. A Python implementation of a PRNG that uses the SHA256 cryptographic hash function is available at

https://github.com/statlab/cryptorandom.[14]

Where is the computational cost?

In our examples for the one-sample and two-sample problems, the bulk of the CPU time is in generating the (single) Monte Carlo sample, especially when using the (expensive) `cryptorandom` SHA256 PRNG. The bisection stage of the algorithm is extremely fast; increasing e from 10^{-8} to 10^{-6} has a trivial effect on runtimes.

Confidence sets for percentiles of an effect in the two-sample problem

In the two-sample problem, the null hypothesis $\theta = \eta$ is that the shift for *every* unit is η. Caughey et al. [2017] show that this hypothesis and the corresponding confidence set can be interpreted as a hypothesis test and confidence set for the maximum and minimum of the shift even when the shift varies from unit to unit, and how to use similar calculations to find confidence bounds for percentiles of the shift.

Nuisance parameters

Sometimes θ is multidimensional but we are only interested in a functional f of θ, e.g., its weighted mean, a component, a contrast, or some other linear or nonlinear functional. There are three general strategies to obtain conservative confidence bounds for $f(\theta)$ when there are nuisance parameters:

- Define the P-value to be the maximum P-value over all values of the nuisance parameters that correspond to the hypothesized value of the parameter of interest, in effect decomposing a composite null into a union of simple nulls and rejecting the composite iff every simple null is rejected [Dufour, 2006, Neyman and Pearson, 1933, Ottoboni et al., 2018c].
- Define the P-value for the composite hypothesis to be the maximum P-value over a confidence set for the nuisance parameters [Berger and Boos, 1994].
- Construct a confidence set for the entire parameter θ, then define the endpoints of the confidence interval for $f(\theta)$ as the supremum and infimum of f over the confidence set. This is sometimes called *strict bounds* [Evans and Stark, 2002, Stark, 1992].

The general strategy of re-using the Monte Carlo sample to test different hypothesized values of the parameter can be used with any of these.

4.7.1 Software

We are aware of only a few packages or repositories with code to compute confidence intervals for the one- or two-sample problem. The `CIPerm` R package constructs intervals for

[14] See `https://statlab.github.io/cryptorandom/`. The package is on PyPI and can be installed with `pip`.

the two-sample problem based on the algorithm in Nguyen [2009]. While this implementation uses a single set of permutations, it uses a brute-force search for the endpoints, which Nguyen [2009] note is time-consuming. Another R package, `Perm.CI`, only works for binary outcomes and uses a similarly inefficient brute force method. Other implementations, such as that in Caughey et al. [2017][15] use the function `uniroot`, an implementation of Brent's method, which incorrectly assumes the P-value is continuous in η.

Open-source Python software implementing the new methods presented in this paper is available at `https://github.com/akglazer/monte-carlo-ci`.

[15] See `https://github.com/li-xinran/RIQITE/blob/main/R/RI_bound_20220919.R`

Chapter 5

Bayesian Audits are Average but Risk-Limiting Audits are Above Average[1]

5.1 Introduction

The 2016 U.S. Presidential election was attacked by Russian hackers, and U.S. Intelligence agencies warn that several nation-states are already mounting attacks on the 2020 election [Zetter, 2018, 2019a,b, Select Committee on Intelligence, 2019]. Almost every U.S. jurisdiction uses computers to count votes; many use computers to record votes. All computerized systems are vulnerable to bugs, misconfiguration, and hacking [Stark and Wagner, 2012]. Voters, poll workers, and election officials are also bound to make mistakes [National Academies of Sciences, Engineering, and Medicine, 2018a]. Enough error from any source—innocent or malicious—could cause a losing candidate to appear to win.

The reported tallies will almost certainly be off by at least a little. Were the tallies accurate enough to ensure that the reported winner(s) really won—that the *reported outcome* is correct?

An election is *evidence-based* [Stark and Wagner, 2012] if it provides convincing public evidence that the reported winners really won. The only federally certified technology that can provide such evidence is trustworthy paper ballots kept demonstrably secure throughout the election and canvass, then audited manually [Appel and Stark, 2020]. However:

- 14% of registered voters live in in jurisdictions using Direct Recording Electronic (DRE) Systems for all voters. DREs do not retain a paper ballot [Verified Voting, 2020].
- Some paper ballots are not trustworthy. For instance, touchscreen voting machines and ballot-marking devices are vulnerable to bugs, hacking, and misconfiguration that can cause them to print the wrong votes [Appel et al., 2020, Bernhard et al., 2020].
- Rules for securing cast ballots and for ensuring the paper trail remains trusworthy are uneven and generally inadequate.

[1]This chapter comprises a publication [Glazer et al., 2020] co-authored by Jacob V. Spertus and Philip B. Stark.

Nonetheless, to focus on statistical issues, we assume here that elections produce a trustworthy collection of paper ballots containing voters' expressed preferences [Appel et al., 2020, Appel and Stark, 2020, Lindeman and Stark, 2012, Stark and Wagner, 2012]. A trustworthy paper trail allows audits to check whether errors, bugs, or malfeasance altered the reported outcome. ("Outcome" means who won, not the exact vote tallies.) For instance, we could tabulate the votes on all the cast ballots by hand, as some recount laws require. But full manual recounts are expensive, contentious, and rare: according to Richie and Smith [2015], only 27 statewide U.S. elections between 2000 and 2015 were manually recounted; three of the recounts overturned the original outcomes (11%).

Some states conduct tabulation audits that involve manually reading votes from some ballots. For instance, California law requires manually tabulating the votes on ballots in 1% of precincts selected at random.[2] Such audits typically do not ensure that outcome-changing errors will (probably) be detected, much less corrected. In contrast, risk-limiting audits (RLAs) [Stark, 2008a, Lindeman and Stark, 2012] have a known minimum chance of correcting the reported outcome if the reported outcome is wrong (but never alter correct outcomes). RLAs stop without a full hand count only if there is sufficiently strong evidence that a full hand count would find the same winners, i.e., if the P-value of the hypothesis that the reported outcome is wrong is sufficiently small.

RLAs have been endorsed by the National Academies of Science, Engineering, and Medicine [National Academies of Sciences, Engineering, and Medicine, 2018a], the American Statistical Association [American Statistical Association, 2010], and many other organizations concerned with election integrity. There have been roughly 60 pilot RLAs in 15 U.S. states and Denmark. Currently 10 U.S. states require or specifically allow RLAs. There have been statewide RLAs or pilot RLAs in five U.S. states: Alaska[3], Colorado [Colorado Secretary of State, 2020], Kansas[4], Rhode Island [Brennan Center for Justice, Rhode Island RLA Working Group, 2019], and Wyoming[3], and a pilot RLA in Michigan in which 80 of 83 counties participated [Michigan Secretary of State, 2020].

Bayesian audits (BAs, [Rivest and Shen, 2012, Rivest, 2018b]) have been proposed as an alternative to RLAs. BAs stop without a full hand count only if the "upset probability"—the posterior probability that the reported winner(s) actually lost, for a particular prior π, given the audit sample—is below a pre-specified threshold. They have been piloted in several states.

Bayesian and frequentist interpretations of probability are quite different. Frequentist probability is the long-run limiting relative frequency with which an event occurs in repeated trials. Bayesian probability quantifies the degree to which the subject believes an event will occur. A prior probability distribution quantifies beliefs before the data are collected; after the data are observed, Bayes' rule says how to update the prior using the data to obtain the posterior probability distribution.

Bayesian methods, including BAs, require stronger assumptions than frequentist methods, including RLAs. In particular, BAs require assuming that votes are random and follow a known "prior" probability distribution π.

[2]The law is a bit more complicated, including provisions to ensure that every contest gets some scrutiny and options for sampling vote-by-mail ballots (including not sampling them if they arrive after election day).

[3]Organized by J. Morrell; one of us (PBS) provided software and support.

[4]J. Morrell, personal communication, 2020

Both RLAs and BAs rely on manually interpreting randomly selected ballots. In principle, both can use a wide range of sampling plans to accommodate differences in how jurisdictions handle and store ballots and variations in election laws and regulations. (To the best of our knowledge, BAs have been conducted only using "ballot polling" [Howard et al., 2019].) RLA methods have been developed to use individual ballots or groups of ballots as the sampling unit, to sample with or without replacement or to use Bernoulli sampling, to sample with and without stratification, and to sample uniformly or with unequal probabilities (see, e.g., Stark [2008a,b, 2020], Lindeman and Stark [2012], Ottoboni et al. [2018b,a]).

The manual interpretations can be used in two ways: *comparison audits* look at differences between the manual interpretation and the machine interpretation and tabulation, while *polling audits* just use the manual interpretation. (The two strategies can be combined in a single audit; see, e.g., Ottoboni et al. [2018b], Stark [2020].) Comparison audits require more of the voting system and require more preparation than polling audits, but for a given size sampling unit, they generally require smaller samples. (The sample size scales like the reciprocal of the margin for comparison audits, and like the square of the reciprocal of the margin for polling audits.) Below, we focus on polling audits that use individual ballots as the sampling unit: *ballot-polling audits*. These are the simplest conceptually and require the least of the voting system: just the reported winner(s), but no other data export.

Both RLAs and BAs lead to a full hand count if sampling does not provide sufficiently strong evidence that the reported outcome is correct. If they lead to a full hand count, that hand count replaces the reported results. Thus, they might confirm a wrong outcome, but they never overturn a correct outcome. They make different assumptions, use different standards of evidence, and offer different assurances, as we shall explain.

```
while (!(full handcount) && !(strong evidence outcome is correct)) {
    audit more
}
if (strong evidence outcome is correct) {
    reported result is final
}
if (full handcount) {
    handcount result is final
}
```

Figure 5.1: Pseudo code for sequential auditing procedures

5.2 Risk

The *risk* of an auditing procedure, given a trustworthy set of cast ballots and a reported outcome, is zero if the reported outcome is correct and is the chance that the procedure will not correct the reported outcome if the reported outcome is wrong. Formally, let θ denote a set of cast votes. For example, in a contest between (only) Alice and Bob in which n ballots were cast, all containing valid votes, θ is an element of $\{\text{Alice}, \text{Bob}\}^n$. (For sampling with

replacement, we could also parametrize the cast votes as the fraction of votes for Alice; see Figure 5.2.)

RLAs treat θ as fixed but unknown. The only probability in RLAs is the probability involved in sampling ballots at random—a probability that exists by fiat and is known to the auditor, because the auditor designs the sampling protocol.

In contrast, BAs treat θ—the cast votes—as random rather than simply unknown. The probability in BAs comes not only from the sampling but also from the assumption that votes are random and follow a probability distribution π known to (or believed by) the auditor.

Let $f(\cdot)$ be the social choice function that maps a set of cast votes to the contest winner(s). Then

$$\text{risk}(\theta) \equiv \begin{cases} \mathbb{P}(\text{audit confirms reported outcome}), & \text{reported winner} \neq f(\theta) \\ 0, & \text{reported winner} = f(\theta). \end{cases}$$

RLAs ensure that the risk does not exceed a pre-specified limit (denoted α), no matter what votes were actually cast. Because θ is fixed, probabilities in RLAs come only from the random sampling of ballots.

BAs control a weighted average of the risk rather than the maximum risk (whence the title of this chapter). The weights come from the prior probability distribution on θ. In symbols:

$$\text{risk}_{\text{RLA}} = \max_{\theta} \ \text{risk}(\theta)$$
$$\text{risk}_{\text{BA}} = \frac{1}{c} \sum_{\theta} \ \text{risk}(\theta)\pi(\theta)$$

where $\pi(\theta)$ is the prior on θ and $c = \sum_{\theta:\text{reported winner} \neq f(\theta)} \pi(\theta)$ makes the weights sum to 1.

BAs can have a large chance of correcting some wrong outcomes and a small chance of correcting others, depending on the prior π. If π assigns much probability to wrong outcomes where it is easy to tell there was a problem (e.g., a reported loser really won by a wide margin) the average risk (the upset probability) can be much lower than the risk for the actual set of ballots cast in the election.

An RLA with risk limit α automatically limits the upset probability to α for any prior, but the converse is not true in general. (The average of a function cannot exceed the maximum of that function, but the maximum exceeds the average unless the function is constant.) Below, we demonstrate that the upset probability can be much smaller than the true risk using simulations based on close historical elections.

5.3 Choosing the Prior for a BA

In a BA, the prior quantifies beliefs about the cast votes and the correctness of the reported outcome before the audit commences. Beliefs differ across the electorate. To address this, Rivest and Shen [2012] considered a "bring your own prior" BA: the audit continues until everyone's upset probability is sufficiently small (see Figure 2A). Of course, if

anyone's prior implies that a reported loser is virtually certain to have won, the audit won't stop without a full hand count.

Ultimately, Rivest and Shen [2012] and Rivest [2018b] recommend using a single "non-partisan" prior. A nonpartisan prior is one for which every candidate is equally likely to win, i.e., a prior that is invariant under permutations of the candidates' names (see Figure 2B). We doubt this captures anyone's beliefs about any particular election. Beliefs about whether the reported winner really won may depend on many things, including pre-election polls and exit polls, the reported margin, reports of polling-place problems, news reports of election interference, etc.

For instance, it seems less plausible that the reported winner actually lost if the reported margin is 60% than if the reported margin is 0.6%: producing an erroneous 60% margin would require much more error or manipulation than producing an erroneous 0.6% margin if the reported winner really lost. On the other hand, when the *true* margin is small, it is easier for error or manipulation to cause the wrong candidate to appear to win. Moreover, a tight contest might be a more attractive target for manipulation.

If every audit is to be conducted using the same prior, that prior arguably should put more weight on narrow margins. Taken to the extreme, the prior would concentrate the probability of wrong outcomes at the wrong outcome with the narrowest margin: a tie or one-vote win for a reported loser.

Indeed, Vora [2019] and Morin et al. [2020] show that in a two-candidate plurality contest with no invalid votes, a ballot-polling BA using a prior that assigns probability $1/2$ to a tie (or one-vote win for the reported loser) and probability $1/2$ to correct outcomes is in fact a RLA (see Figure 2C): the upset probability equals the risk.

Constructing priors that make BAs risk-limiting for more complicated elections (e.g., elections with more than two candidates, elections in which ballots may contain invalid votes, social choice functions other than plurality, and audit sampling designs other than simple random samples of individual ballots or random samples of individual ballots with replacement) is an open problem.[5]

5.4 Empirical Comparison

How are risk and upset probability related? The upset probability is never larger than the risk, but the risk is often much larger than the upset probability for BAs with non-partisan priors, as we show using data from three recent close U.S. elections: the 2017 House of Delegates contest in Virginia's 94th district, the 2018 Congressional contest in Maine's 2nd district, and the 2018 Georgia Governor contest. The simulations, summarized in Table 1, treat the reported vote shares as correct, but re-label the reported winner as the reported loser. "Simulated Risk" is the estimated probability that a BA with 5% upset probability

[5]This is related to the problem of constructing *least-favorable priors* in statistical decision problems. There is a deep duality between Bayesian and frequentist procedures: under mild regularity conditions the Bayes risk for a *least-favorable prior* is equal to the *minimax risk* [Bickel and Doksum, 2006]. (Here, risk is a term of art, a measure of the performance of the procedure.) That is to say, for a particular choice of prior, the Bayesian procedure is in fact the frequentist procedure that does best in the worst case. The least-favorable prior is generally not "flat" or "uninformative."

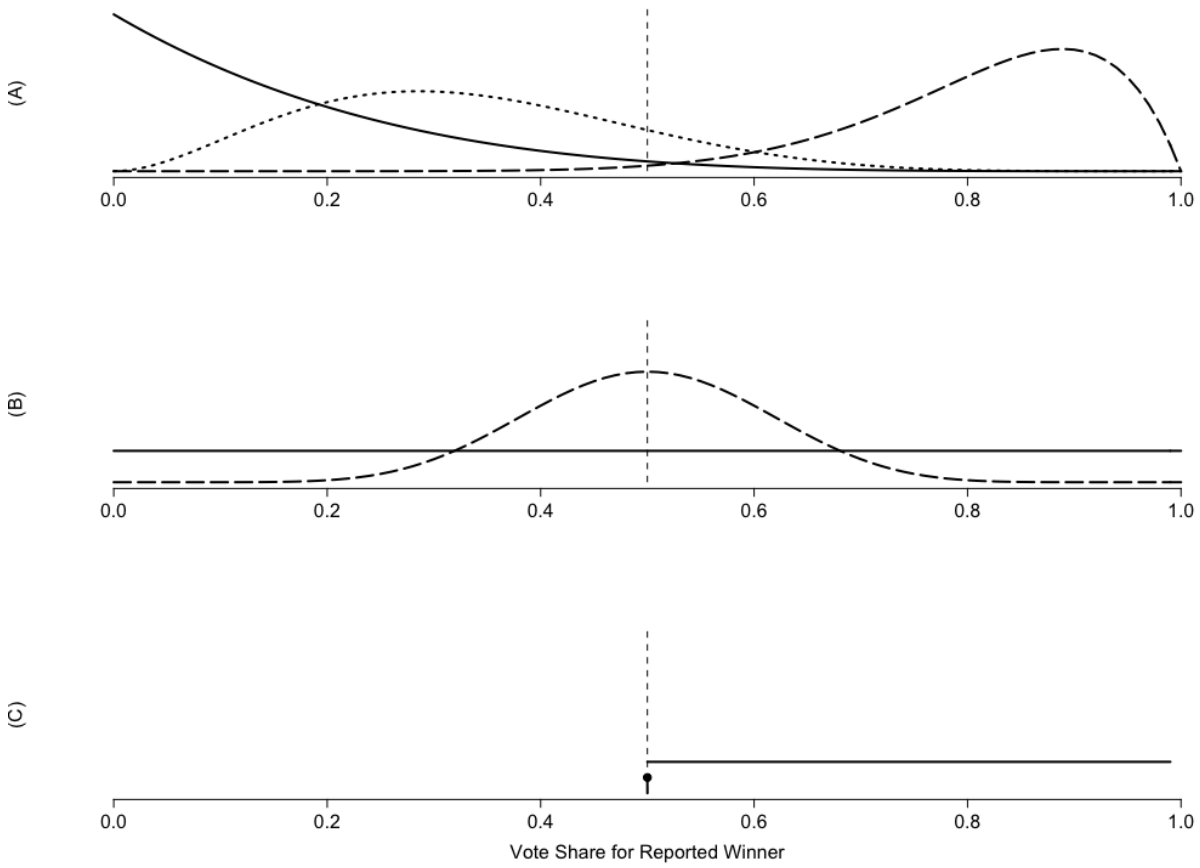

Figure 5.2: *Exemplar priors for the true vote share for the reported winner in a two-candidate election. Values to the right of the vertical dotted line (at 1/2) correspond to correct reported outcomes: the winner got more than 50% of the valid votes. (A) plots three possible partisan priors. For BAs that allow observers to bring their own prior, a BA would stop only when all three posteriors give a sufficiently low probability to all outcomes where the reported winner actually lost: values less than or equal to 1/2. (B) plots two nonpartisan priors (the priors are symmetric around 1/2 and thus invariant under exchanging the candidates' names) including the flat prior recommended by Rivest and Shen [2012]. The flat prior gives equal weight to all possible vote shares. (C) plots a least-favorable prior, a prior for which a BA is an RLA with risk limit equal to the upset probability. It assigns probability 1/2 to a tie, the wrong outcome that is most difficult to detect. The rest of the probability is spread (arbitrarily) across vote shares for which the reported outcome is correct. In this illustration, that probability is uniform. That choice affects the efficiency but not the risk.*

corrects the reported outcome. The simulations use the nonpartisan prior recommended by [Rivest, 2018b], with initial "pseudo-counts" of 0.5. Each audit begins with a sample of 25 ballots. Each step of each audit simulates 1,000 draws from the posterior distribution to estimate the upset probability. If the upset probability is above 5%, then the sample is increased by 20%, and the upset probability is estimated again. Each audit stops when the upset probability falls below 5%, or all ballots have been audited. We simulate 10,000 ballot-polling BAs for each scenario. Code for the simulations is available at `https://github.com/akglazer/BRLA-Comparison`.

A recount of the 2017 Virginia 94th district contest gave a 1-vote win for Simonds over Yancey. (A three-judge panel later determined that a vote counted as an overvote should be attributed to Yancey; the winner was determined by drawing a name from a bowl [McCammon, 2018].) The 2018 Maine Congressional election used ranked-choice voting (RCV/IRV). While there are methods for conducting RLAs of IRV contests [Blom et al., 2019, Stark, 2020], we treat the contest as if it were a plurality contest between the last two standing candidates, Golden and Poliquin, a "final-round margin" of 3,509 votes.[6]

	Number of Votes Cast	Margin	BA Risk (simulated)
Virginia 94th	23,215 votes	1 vote (0.004%)	43%
Maine 2nd	281,371 votes	3509 votes (1.25%)	23%
Georgia Governor	3,902,093 votes	54,723 votes (1.4%)	22%

Table 5.1: Simulated risk of a Bayesian Audit using 5% upset probability with a "nonpartisan" prior for the 2017 Virginia House of Delegates District 94 contest, the 2018 Maine 2nd Congressional District contest, and the 2018 Georgia gubernatorial contest. Column 2: the margin for each election in number of votes and percentage. Column 3: risk of the BA, i.e., the estimated probability that the BA audit will fail to correct the outcome.

In these experiments, the actual risk of the BA is 4 to 9 times larger than the upset probability, 5%. For example, in the Virginia 94th District contest, the BA failed to correct the outcome 43% of the time, 8.6 times the upset probability. This results from the fact that the upset probability averages the risk over all possible losing margins (with equal weight), while the actual losing margin was small. Figure 5.3 shows the simulated risk of a BA with a nonpartisan prior and initial pseudo-counts of 0.5 for an election with 1,000,000 total votes cast. The risk is plotted as a function of the vote share for the winner. The empirical risk for a BA is very high for small margins, where auditing is especially important. As far as we know, there are situations where the risk can be an arbitrarily large multiple of the upset probability, depending on the actual cast votes, the social choice function, the prior, and details of the BA implementation (such as its rule for expanding the sample).

5.5 Conclusion

Elections are audited in part to rule out the possibility that voter errors, pollworker errors, procedural errors, reporting errors, misconfiguration, miscalibration, malfunction,

[6]The final-round margin of an IRV contest is an upper bound on the true margin.

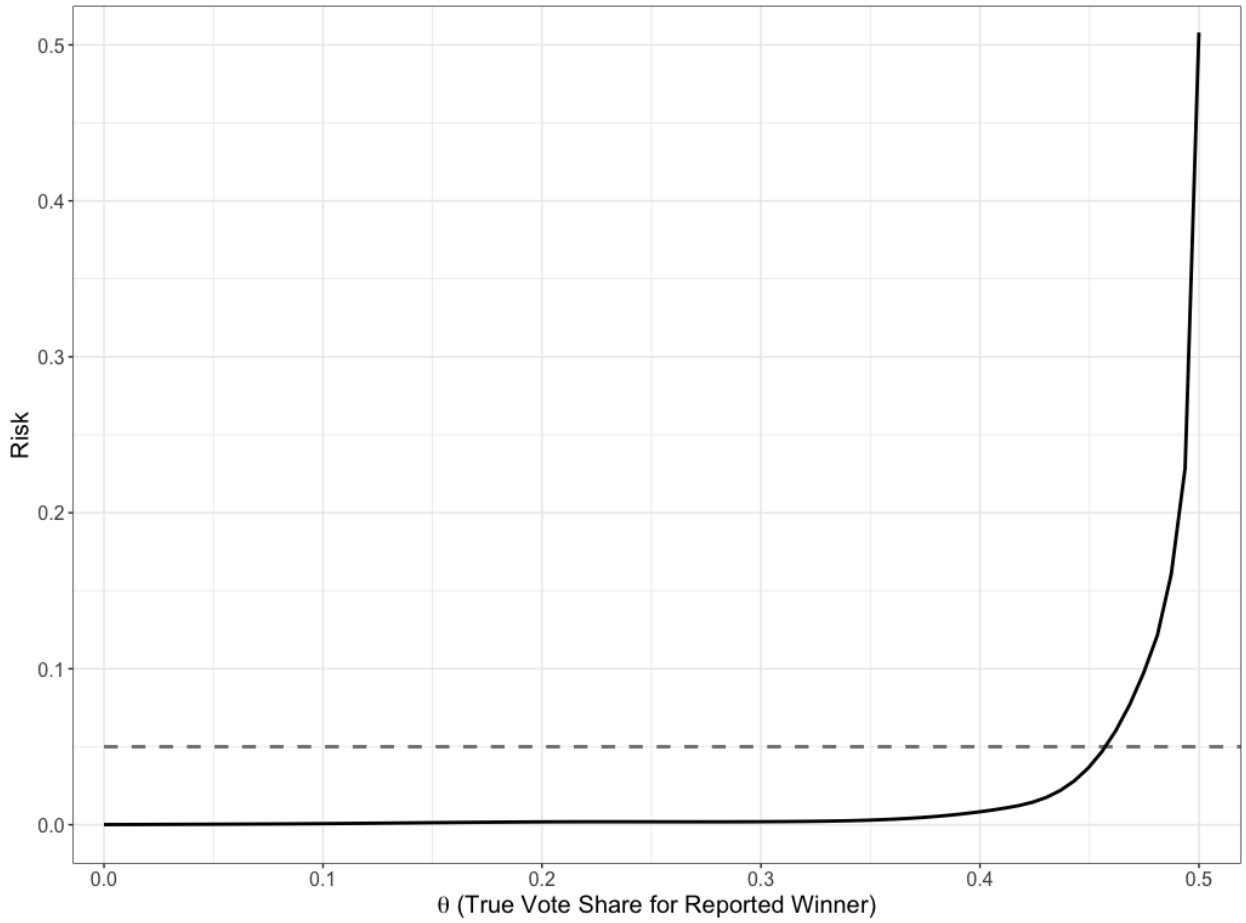

Figure 5.3: Simulated risk (solid line) of a BA with nonpartisan prior for a two-candidate election with 1,000,000 total votes cast and no invalid votes. The x-axis is θ, the actual vote share for the reported winner. The reported winner really won if $\theta > 0.5$ and lost if $\theta < 0.5$. The y-axis is the actual risk, computed for $\theta < 0.5$ as the number of times the BA confirms the outcome over the total number of simulated audits. If $\theta > 0.5$ then the risk is 0. The dashed grey line at Risk = 0.05 is the upset probability threshold for the BA, and also the maximum risk for a RLA with risk limit 0.05.

bugs, hacking, or other errors or malfeasance made losing candidates appear to win. We believe that controlling the probability that the reported outcome will not be corrected when it is wrong—the risk—should be the minimal goal of a post-election audit. RLAs control that risk; BAs control the upset probability, which can be much smaller than the risk.

Both RLAs and BAs require a trustworthy paper trail of voter intent. RLAs use the paper trail to protect against the worst case: they control the chance of certifying the reported outcome if it is wrong, no matter why it is wrong.

BAs protect against an *average* over hypothetical sets of cast votes (rather than the worst case); the weights in the average come from the *prior*.

The priors that have been proposed for BAs do not seem to correspond to beliefs about

voter preferences, nor do they take into account the chance of error or manipulation. Moreover, BAs do not condition on a number of things that bear on whether the reported outcome is likely to be wrong, such as the reported margin and the political consequences. As Vora [2019] shows, some BAs are RLAs if the prior is chosen suitably. Bayesian upset probabilities can never be larger than the maximum risk, but it seems that they can be arbitrarily smaller. Conversely, Huang et al. [2020] discuss finding a threshold for the upset probability in a BA (for any fixed prior) in a two-candidate, no invalid-vote contest so that using that threshold as a limit on the upset probability yields an RLA (with risk limit larger than the upset probability).

Sequential RLAs stop as soon as there is strong evidence that the reported result is correct. When the outcome is correct by a wide margin, they generally inspect relatively few ballots. Thus, even though RLAs protect against the worst case, they are relatively efficient when outcomes are correct. (When outcomes are incorrect, they are intended to lead to a full hand tabulation.)

Partisanship, foreign interference, vendor misrepresentations [Zetter, 2018], and suspicious results [Ottoboni and Stark, 2019] all threaten public trust in elections, potentially destabilizing our democracy. Conducting elections primarily on hand-marked paper ballots (with accessible options for voters with disabilities), routine compliance audits, and RLAs can help ensure that elections deserve public trust.

Chapter 6

More style, less work: card-style data decrease risk-limiting audit sample sizes[1]

> Style is a way to say who you are without having to speak.
>
> – Rachel Zoe

6.1 Introduction

The principle of *evidence-based elections* is that elections should provide convincing evidence that the reported winners really won [Stark and Wagner, 2012]. Evidence-based elections require a trustworthy record of the votes. Generally, that means hand-marked paper ballots (with an accessible marking option for voters who require accommodations) kept demonstrably secure throughout the canvass [Appel and Stark, 2020, Appel et al., 2020]. Crucially, the reported results must be checked against that trustworthy paper trail. *Risk-limiting audits* (RLAs) provide a rigorous way to perform that check so that there is a large chance of correcting the reported outcome if it is wrong. The probability that an RLA does not correct the reported outcome if the reported outcome is wrong is less than the risk limit [Stark, 2008a, Lindeman and Stark, 2012].

U.S. elections generally include several to dozens of contests that widely vary in size: each voter is eligible to vote in a subset of those contests. For example, a voter may be eligible to vote for president, senators, governor, representatives, mayor, judges, school board seats, local tax measures, etc. Because ballots often contain so many contests, they generally comprise more than one *card* or *sheet* or *page*, each containing some of the contests.

The term *ballot style* generally refers to the set of contests on a given voter's ballot. (Ballot style can also encode precinct information, i.e., even if voters in two different precincts are eligible to vote in the same set of contests, ballots for the two precincts are considered to be of two different styles.) Here, we use *card style* to refer to the set of contests on a given ballot card, and CSD to refer to card-style data for an election.

RLAs generally involve inspecting a random sample of ballot cards. States that perform RLAs have so far drawn the sample of ballot cards to inspect from all the cards cast in the

[1] This chapter comprises a publication [Glazer et al., 2021] co-authored by Jacob V. Spertus and Philip B. Stark.

election, without regard for the contests those cards contain. When a contest under audit appears on only a small fraction of ballot cards, this can make the audit unnecessarily costly. Especially when the margin[2] is small, an RLA based on a sample drawn from all ballot cards requires manually inspecting far more cards than an RLA that only samples cards that contain the contest. Targeting the sample to just those cards is possible if card-style data are available, that is, a listing of the contests each cast card contains.

A similar issue arises for large contests—even jurisdiction-wide contests—when ballots consist of multiple cards: if the sample can be drawn just from cards that contain those contests, a much smaller sample may suffice to confirm the outcome than if the sample is drawn indiscriminately from all cast cards [Lindeman et al., 2018]. As a rule of thumb, if a contest is on a fraction f of ballot cards cast in the contest, the sample size required to confirm the contest outcome will be roughly $1/f$ times larger than if the sample can be drawn just from cards that contain the contest.

Here, we show that CSD–i.e., keeping track of which cards contain which contests—can reduce audit sample sizes by orders of magnitude in typical elections. Section 6.2 provides additional context and defines key terms. Section 6.3 examines the simplest case with a ballot-level comparison audit: auditing two contests in an election with a one-card ballot. One contest is on every card; the other is on only some of the cards. Section 6.4 considers auditing two contests in an election with multi-card ballots. Section 6.5 extends our analysis to ballot-polling audits. Section 6.6 presents case studies of hypothetical audits in two California counties of different sizes. Section 6.7 sketches how to implement an audit that takes advantage of CSD using *consistent sampling*. Section 6.8 presents conclusions and recommendations.

6.2 Background

6.2.1 Ballots, cards, ballot manifests, and card styles

A *ballot* is what the voter receives and casts; a *ballot card* is an individual page of a ballot. In the U.S., ballots often consist of more than one card. The ballot cards that together comprise a ballot generally do not stay together once they are cast. RLAs generally draw ballot *cards* at random—not "whole" ballots.

To conduct an RLA, an upper bound on the number of validly cast ballot cards must be known before the audit begins. The bound could come from manually keeping track of the paper, or from other information available to the election official, such as the number of voters eligible to vote in each contest, the number of pollbook signatures, or the number of ballots sent to polling places, mailed to voters, and returned by voters [Bañuelos and Stark, 2012].

RLAs generally rely on *ballot manifests* to draw a random sample of ballot cards. A ballot manifest describes how the physical ballot cards are stored. It is the *sampling frame*

[2]Technically, the *diluted margin* [Stark, 2010] drives sample sizes for ballot-level comparison audits, as described below. The diluted margin is the margin in votes divided by the total number of cards in the population from which the sample is drawn. It is generally less than the conventional margin, which is the margin in votes divided by the number of valid votes in the contest, excluding undervotes and overvotes.

for the audit. This chapter explains how it can be beneficial to augment the ballot manifest with information about the *style* of each card, i.e., the particular contests the card contains—card-style data (CSD). Until recently,[3] CSD has not been used in RLAs. Figures 6.1 and 6.2 respectively display examples of CSD for single-card and multi-card ballots.

Cart	**Tray**	**Position in Tray**	**Governor**	**Mayor of Irvine**	...
1	4	96	Yes	No	...
5	1	12	Yes	No	...
2	2	72	Yes	No	...
...	...	...	...	...	...
3	5	50	Yes	Yes	...

Figure 6.1: A hypothetical example of card-style data (CSD) for an election in Orange County California with one-card ballots. Ballot cards are uniquely identified by their position (cart, tray, position in tray). CSD further identifies contests that each ballot card is supposed to contain (truncated to two contests here), appearing as columns. Here we display the records for the county-wide Governor's race and the race for Mayor of Irvine (a city within Orange County). More storage-efficient CSD might associate an unstructured list to each ballot card with numeric identifiers of the contests it contains (e.g., $\{1, 4, 10, 12\}$). The public should be able to check that there are N lines in the CSD, where N is the number of ballots (and ballot cards) cast in the county. If N_S ballots were cast in contest S, there should be N_S lines with "yes" in the column corresponding to contest S in the CSD.

Cart	**Tray**	**Position in Tray**	**Governor**	**Mayor of Irvine**	...
2	6	3	No	Yes	...
5	1	12	No	No	...
2	5	64	Yes	No	...
...	...	...	...	...	...
1	2	8	Yes	No	...

Figure 6.2: An example of card-style data (CSD) for a hypothetical election in Orange County with multi-card ballots. The Governor's race and the race for Mayor of Irvine appear on different cards. Because neither contest can appear on a card in this example (e.g., line 2), there must be at least 3 card styles.

There are two principal ways to generate CSD: (1) physically sort the ballot cards according to the contests they contain, or (2) rely on the voting system for that information—even though it might not be accurate. Precinct-based voting partially sorts ballot cards: If every voter in the precinct is eligible to vote in the same contest(s) and the ballot has only one card, then each precinct's ballot cards have a single style. Knowing which precinct a bundle of ballot cards came from then tells us the contests on each card. This does not work for multi-card ballots. Some jurisdictions sort vote-by-mail ballots by precinct before scanning, which also partially sorts the cards. Vote centers make sorting ballots more difficult because

[3]In November 2019, a pilot RLA was conducted in San Francisco that used CSD [Blom et al., 2020].

each center receives ballot styles from more than one precinct, and ballots cast in vote centers generally are not sorted before they are scanned.

Some modern vote tabulation systems record a cast-vote record (CVR, a record of how the voting system interpreted the selections on the ballot card) for each ballot card in a way that allows the corresponding physical card to be identified and retrieved, and vice versa. Such systems are amenable to efficient RLAs and they also contain (possibly inaccurate) CSD: the contests on a card can be inferred from its CVR, as long as the CVR encodes "no selection" for contests in which the voter did not express a preference, according to the voting system. CSD derived from CVRs rely on the voting system, so they could be wrong: CSD might show that a card contains a contest it does not contain, or vice versa. Such errors can be accounted for rigorously in the audit using the "manifest phantoms to evil zombies" approach [Bañuelos and Stark, 2012, Stark, 2020], described below. The same approach can accommodate errors in CSD uncovered in auditing manually or machine-sorted ballot cards.

Because physically sorting cards is expensive, we expect that CSD generally will not be available unless the jurisdiction has a voting system that can produce CVRs linked to physical ballots. If such CVRs are available, *ballot-level comparison* RLAs are possible. Such RLAs are especially efficient, so we emphasize them below.

6.2.2 Ballot-polling and ballot-level comparison audits

There are two common approaches to auditing: *comparison* and *ballot-polling* [Lindeman and Stark, 2012] (there are also audits that combine the two approaches; see, e.g., Ottoboni et al. [2018b], Stark [2020]). Comparison audits involve comparing manual tabulation of physically identifiable sets of ballot cards to the machine tabulation of the same ballot cards. The efficiency of comparison audits increases as the size of the sets decreases. The most efficient comparison audits compare human interpretation of individual ballot cards to the machine interpretation of individual ballot cards, CVRs. Such audits are called *ballot-level comparison audits* (in contrast to *batch-level comparison audits*, which compare the manual and electronic tabulation of batches of ballots, such as ballots cast in person in a particular precinct). Ballot-level comparison audits are possible only if the voting system produces CVRs that can be linked to the corresponding physical ballot card.[4] Legacy voting systems generally do not, but many newer systems do.

Ballot-polling audits check the outcome but do not check or rely on the machine tabulation. They do not require CVRs or machine subtotals: all they require is paper ballots, organized well enough to sample cards at random.

Here, we focus on ballot-level comparison audits but we briefly address ballot-polling audits. We do not address batch-level comparison audits.

Ballot-level comparison audits and ballot-polling audits sample individual physical ballot cards. We refer to the act of sampling a single ballot card as a *draw*. Retrieving and inspecting ballot cards is labor intensive, so when the reported outcome is correct we want to minimize the number of draws (i.e., the sample size). We show below that knowing which cards contain which contests can dramatically reduce the number of draws required

[4]There are also *transitive* ballot-level comparison audits, which involve re-scanning the ballot cards using an unofficial system.

to confirm correct outcomes. (When the outcome is incorrect, we *want* the audit to inspect every ballot, in order to determine the correct outcome.)

6.2.3 Super-simple simultaneous single-ballot RLAs

In this chapter we use the super-simple simultaneous single-ballot (S4) RLA of Stark [2010] to illustrate workloads. Although S4 is not the most efficient RLA method, in part because it relies on sampling with replacement, it allows simple workload computations without the need for simulation. Moreover, if the sample size is small relative to the number of ballots cast and the number of observed discrepancies is small, the method is close to the best known. (If the sample size is an appreciable fraction of the population, other methods can be far more efficient. See, e.g., [Stark, 2020].) We expect that the savings in workload afforded by CSD will be substantial for all RLA methods.

The number of draws S4 needs to confirm results depends on the diluted margin and the number and nature of discrepancies the sample uncovers.[5] The initial sample size can be written as a constant (denoted ρ)[6] divided by the "diluted margin." With CSD, there are two relevant "diluted margins," as we shall see. The *partially diluted margin* is the margin in votes divided by the number of *cards* that contain the contest, including cards with undervotes or no valid vote in the contest. This differs from how margins are often reported, where the denominator is only the *valid votes* in the contest, not the number of cards cast in the contest. The *fully diluted margin* is the margin in votes divided by the number of cards in the population of cards from which the audit sample is drawn. When the sample is drawn only from cards that contain the contest, the partially diluted margin and the fully diluted margin are equal; otherwise, the fully diluted margin is smaller. If CSD are unavailable, the number of cards in that population is the number of cards cast in the jurisdiction. If CSD are available, the number of cards in the population can be reduced to the number of cards that contain the contest. The availability of CSD drives the sample size through the difference between the partially and fully diluted margins.

6.3 One-Card Ballots

Consider an election with a one-card ballot. We want to audit two contests: a jurisdiction-wide contest B (for "big") listed on every card and a smaller contest S (for "small") listed on some of the cards. For example, B might be a countywide contest such as sheriff and S might be a mayoral race.

There are N ballots cast in the jurisdiction, of which $N_B = N$ contain contest B and $N_S = pN < N$ contain contest S, where $p \in (0, 1)$.[7]

[5] The S4 method has one tuning parameter, γ, which does not affect the risk limit but does affect the workload. To estimate the final workload, we assume that some fraction of the ballots in the sample will reveal 1-vote overstatements (errors that inflated a reported winner's margin over a reported loser by 1 vote). See Stark [2010].

[6] In general, $\rho = -\log(\alpha)/[\frac{1}{2\gamma} + \lambda \log(1 - \frac{1}{2\gamma})]$, where γ is an error inflation factor and λ is the anticipated rate of one-vote overstatements in the initial sample as a percentage of the diluted margin [Stark, 2010]. We define γ and λ as in `https://www.stat.berkeley.edu/~stark/Vote/auditTools.htm`.

[7] The fraction p can be quite small in real elections. For instance, a 2018 ballot measure in San Mateo

The reported margin of contest B is M_B votes and the reported margin of contest S is M_S votes. Let $m_B \equiv M_b/N_B$ and $m_S \equiv M_S/N_S$ be the two partially diluted margins. We assume that the ballot manifest and the CVR both indicate that N ballots were cast overall, and that ballot cards are where the manifest says they are: there are no "phantoms" in the terminology of Bañuelos and Stark [2012]. We find sample sizes for a risk limit of 0.05 on the assumption that the rate of one-vote overstatements will be 0.001. We assume that other types of errors did not occur.

Absent CSD, the sample for auditing contest S would be drawn from the entire population of N ballots. Contest S is on pN cards, so, in addition to the cards that contain undervotes or invalid votes in contest S, there are in effect another $(1-p)N$ cards with non-votes in contest S. The fully diluted margin for contest S is thus $M_S/N = pm_S$. For simplicity, we use the same risk limit for both contests (5% in our numerical examples). The number of cards we need to sample will be the larger of the sample size for contest B (with fully diluted margin m_B and risk limit α), and for contest S (with fully diluted margin pm_S and risk limit α). If $m_B \leq pm_S$ then we will be done auditing both contests when the audit of contest B is complete (assuming the number of discrepancies was not larger than anticipated). However, if $m_B > pm_S$ we must sample more ballots to finish the audit of contest S.

Here's an example. Suppose $N = 10{,}000$, $p = 0.1$, and $m_B = 0.1 = m_S$. For contest B, there are 5,500 reported votes for the winner and 4,500 reported votes for the loser; for contest S there are 550 votes for the winner and 450 votes for the loser. Contest B has an initial sample size of 64 cards, and contest S has an initial sample size of 721 cards. Therefore, we will need to sample 721 cards. In general, for contest S, we would need to sample approximately $1/p$ times more cards than if we sampled just from the cards that contain contest S. When m_S and p are small, this can lead to very large samples. For example, if $p = 0.01$ and n is the number of cards we would need to inspect if we were sampling just from cards that contained the contest with partially diluted margin m_S, then we could need to sample approximately $100n$ cards if we sampled from all cards (not just those that contain contest S).

If there are CSD, the sample can be substantially smaller: We could first sample from all the cards until the audit of contest B is complete. Then we can use CSD to draw additional cards that contain contest S.

Consider the same example as before with $N = 10{,}000$, $p = 0.1$, and $m_B = 0.1 = m_S$, but suppose we have CSD. To audit B we still need to draw 64 cards. We expect that $64p = 6.4 \approx 6$ of those cards will also contain contest S and $64(1-p) \approx 58$ will not. To finish the audit of S, we expect to need to draw 58 more cards, all containing contest S, for a total of 122 cards. CSD reduces the number of audited cards from 721 to about 122, a factor of almost 6.

Figure 6.3 displays total ballot cards needed to audit contests B and S, with and without CSD, across a range of partially diluted margins $m_B = m_S$ and proportions p of cards on which contest S appears.

Figure 6.4 plots the number of cards needed without CSD as a percentage of the number needed with CSD. Without CSD we need to inspect substantially more ballots when p is small. For very small p, an RLA using S4 is inefficient: a full hand count will be less work

County, California, had 687 eligible voters out of 399,591 registered voters, i.e., $p \approx 0.0017$.

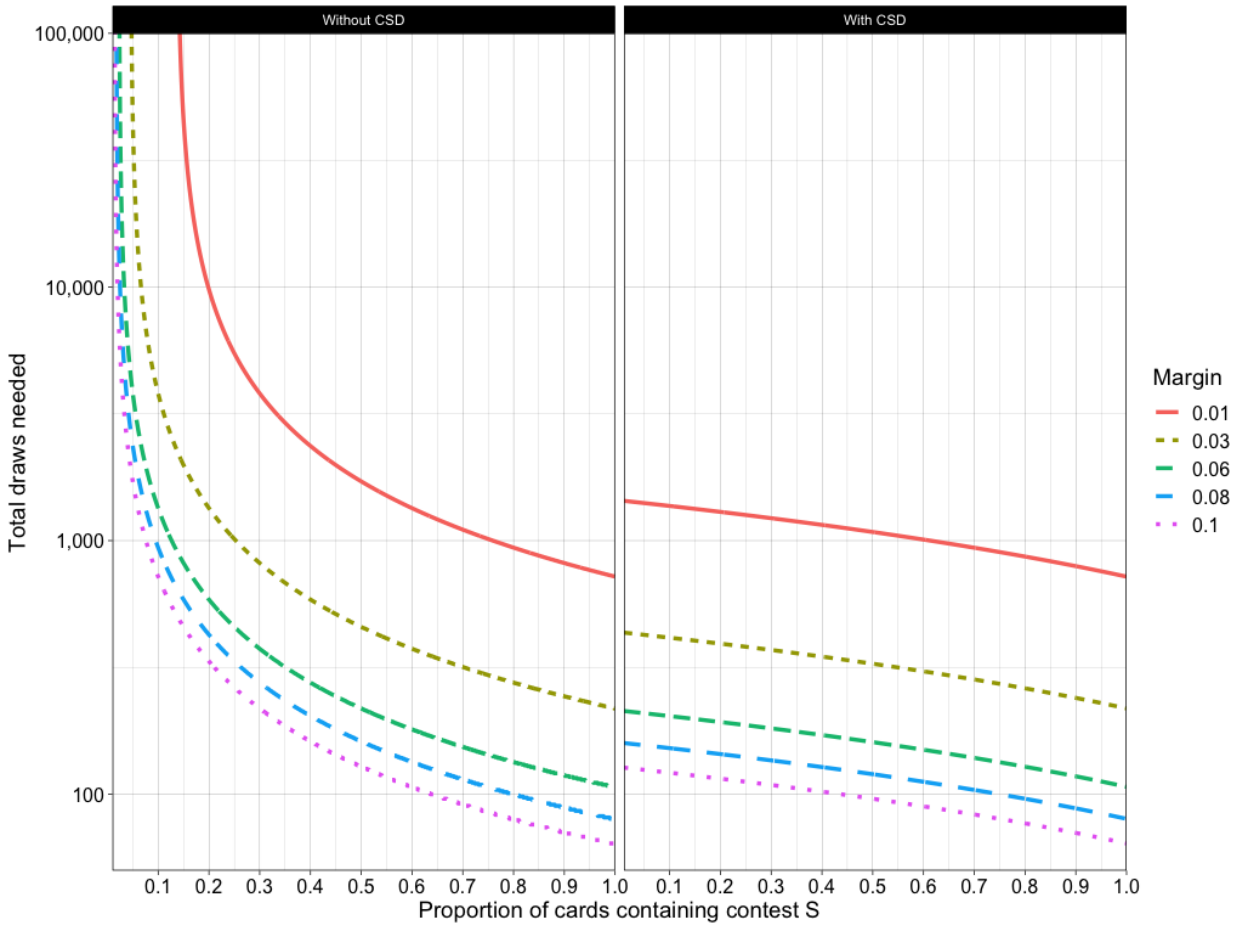

Figure 6.3: Expected draws needed to confirm the outcomes in both contest B and contest S when the reported outcomes are correct with partially diluted margin $m_B = m_S$ given in the legend. An error rate of 0.001 1-vote overstatements is assumed. The y-axis is the expected number of draws (i.e., the sample size) needed on the $\log_{10}$ scale and is truncated at 100,000 draws. The left panel gives the expected number of draws if CSD are unavailable, while the right panel gives the expected number of draws with CSD. The x-axis ranges over a grid of proportions p of ballot cards containing the small contest, from 1 in 100 ($p = 0.01$) to every ballot card ($p = 1.0$). When contest S is on every card ($p = 1$), the workload is the same with or without CSD.

than sampling.

Figure 6.5 is similar to Figure 6.4 except m_S is set to 0.1 and m_B varies from 0.01 to 0.1. When $pm_S > m_B$, there is no penalty for not having CSD.

Recall that the number of cards that must be audited is ρ divided by the fully diluted margin, so the audit of contest B will examine $\frac{\rho}{m_B}$ cards and the audit of contest S will (on average) examine $\frac{\rho}{m_S} - \frac{p\rho}{m_B}$ additional cards, for a total of $(1-p)\frac{\rho}{m_B} + \frac{\rho}{m_S}$ cards. On the other hand, without CSD we would have had to examine $\frac{\rho}{pm_S}$ ballots. If $m_B \geq pm_S$, CSD

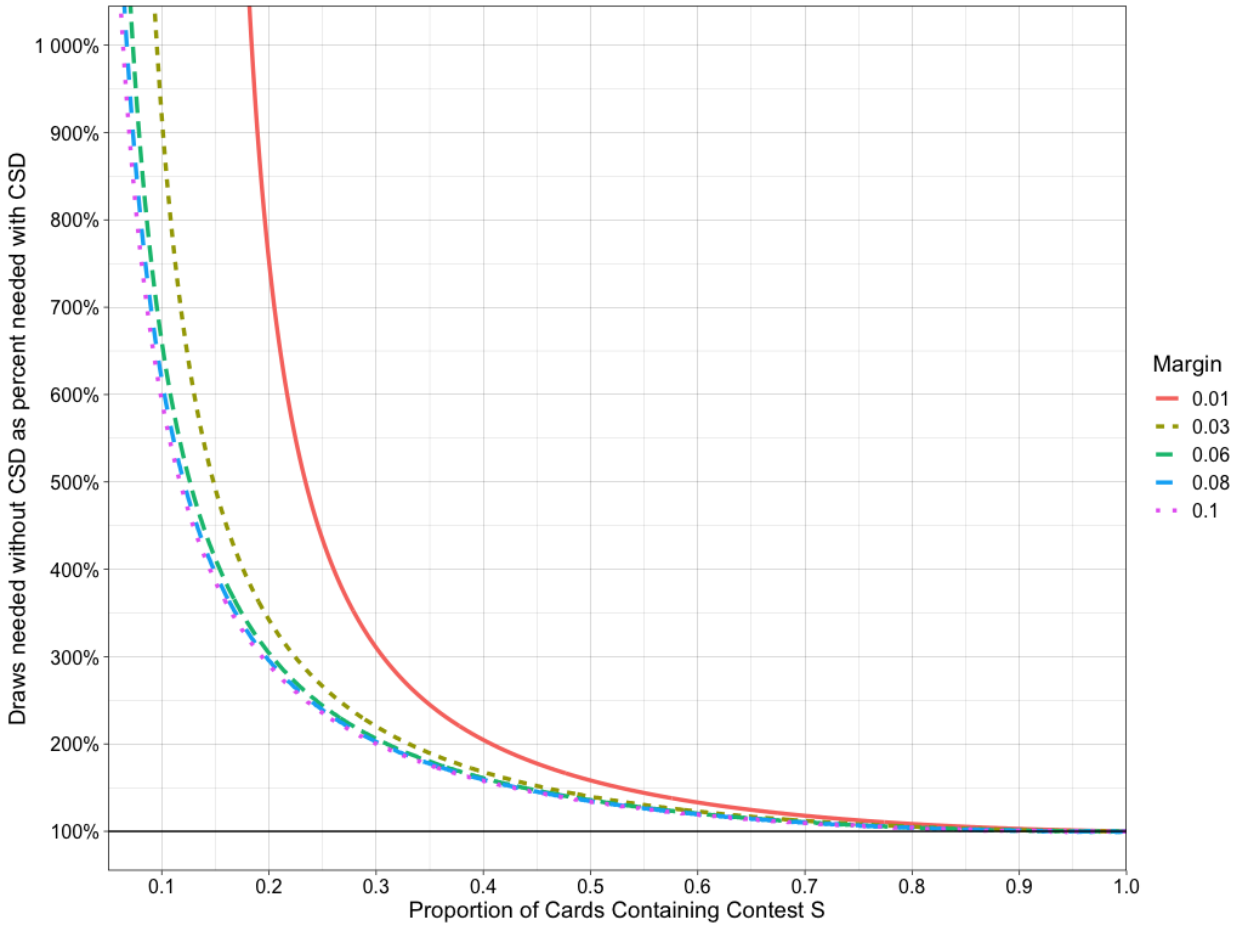

Figure 6.4: Percentage increase in the number of draws required for an RLA without CSD compared to an RLA with CSD. Partially diluted margins with $m_B = m_S$ are given in the legend. Risk limit is 5%; the audit method is S4. The y-axis shows expected draws without CSD as a multiple of the draws needed with CSD. For example, for a partially diluted margin of 0.03, an audit without CSD will require inspecting about 5 times as many ballots as an audit that uses CSD if the small contest is on 15% of the ballot cards. The y-axis is truncated at 1,000%. The x-axis ranges over a grid of proportions p of ballot cards containing the small contest, from 1 in 10 ($p = 0.1$) to every ballot card ($p = 1.0$).

reduces the workload by

$$\frac{\rho}{pm_S} - \left((1-p)\frac{\rho}{m_B} + \frac{\rho}{m_S} \right)$$

on average.[8]

[8]The sample size for auditing contest B is fixed, as is the sample size for auditing contest S, but the overlap of the two samples is random. The expected overlap is $p\rho/m_b$.

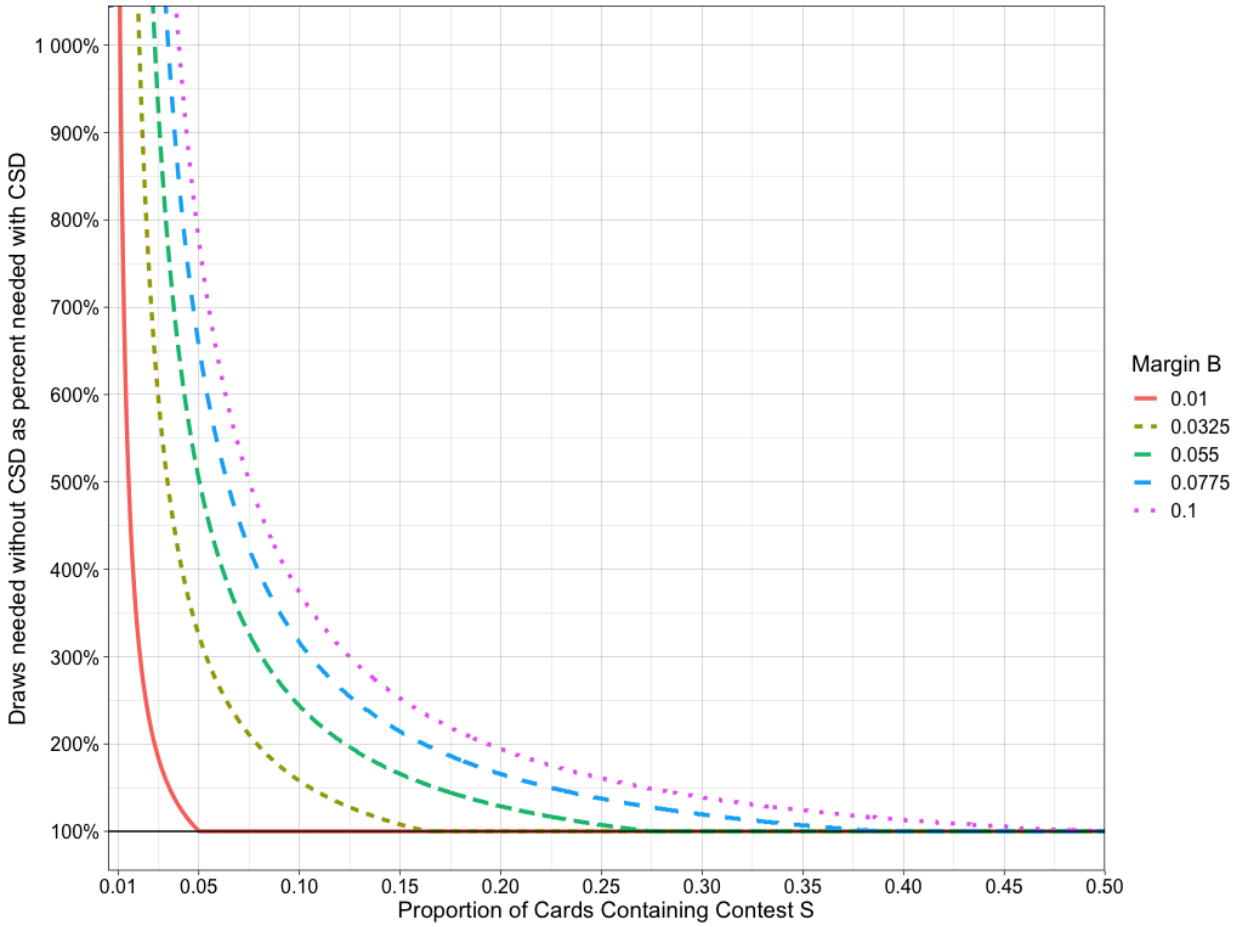

Figure 6.5: Draws needed to confirm the outcomes in both contest B and contest S when the reported outcomes are correct, as a function of the partially diluted margin in contest B. Results are for the S4 auditing method and a risk limit of 5%, on the assumption that the rate of 1-vote overstatement errors is 0.001. The partially diluted margin for contest S is fixed at 0.2. The y-axis gives the expected number of draws that are needed when we do not have card-style data (CSD) as a multiple of the number of draws needed when we do. For example, for a partially diluted margin of 0.055, when the smaller contest is on 5% of the ballot cards, 5 times as many draws are needed if we do not have CSD. The y-axis is truncated at 1,000%. The x axis ranges over a grid of proportions p of ballot cards containing contest S, from 1 in 100 to 1 in 2.

6.4 Multi-Card Ballots

Now suppose that each ballot consists of $c > 1$ cards. For simplicity, suppose that every voter casts all c cards of their ballot. Contest B is on all N *ballots* and on N of the Nc *cards*. Contest S is on Np of the Nc cards.

Suppose there are $N = 10{,}000$ ballots, $p = 0.1$, and $m_B = 0.1 = m_S$, the risk limit $\alpha = 0.05$, and we assume that the sample will reveal 1-vote overstatement errors at a rate of 0.001, as before. Recall that for $c = 1$, we had to sample 721 cards without CSD and 122

with CSD.

For $c = 2$, without CSD the fully diluted margins are

$$\frac{M_B}{cN} = \frac{1}{c} m_B = 0.1/2 = 0.05$$

for contest B and

$$\frac{M_S}{cN} = \frac{p}{c} m_s = 0.1 \times 0.1/2 = 0.005$$

for contest S, so the audit will examine $\rho/0.005 = 1{,}712$ cards.

For $c = 5$, the fully diluted margins become

$$\frac{1}{c} m_B = 0.1/5 = 0.02$$

and

$$\frac{p}{c} m_S = 0.1 \times 0.1/5 = 0.002$$

without CSD, so the audit will examine 9,775 cards.

If we had CSD, we would only need to sample 122 cards as before, no matter how large c is, if every card that contains contest S also contains B. If contests B and S are on different cards then with CSD we would need to sample $64 + 64 = 128$ ballot cards, because no card that contains S also contains B.

As the number c of cards per ballot increases, the sample size without CSD grows in proportion, but the sample size with CSD stays constant: having CSD saves more work the more cards per ballot there are.

If we do not have CSD we need to examine

$$\max\left(\frac{c\rho}{m_B}, \frac{c\rho}{pm_S}\right)$$

cards. Suppose $m_B > pm_S$, so we need to examine $\frac{c\rho}{pm_S}$ ballot cards if we do not have CSD. With CSD the audit of contest B will examine $\frac{\rho}{m_B}$ cards and the audit of contest S will examine either an additional $\frac{\rho}{m_S}$ ballot cards if it is on a different ballot card from contest B or $\frac{\rho}{m_S} - p\frac{\rho}{m_B}$ (on average) cards if every card that contains S also contains B. This results in the following expression for the difference in the number of draws with and without CSD:

$$\begin{aligned}
\left(\frac{c\rho}{pm_S}\right) - \left(\frac{\rho}{m_B} + \frac{\rho}{m_S}\right) &\quad \text{if contests } B \text{ and } S \text{ are on different cards.} \\
\left(\frac{c\rho}{pm_S}\right) - \left((1-p)\frac{\rho}{m_B} + \frac{\rho}{m_S}\right) &\quad \text{if contests } B \text{ and } S \text{ are on the same cards.}
\end{aligned}$$

Figure 6.6 plots the sample size needed without CSD as the percentage needed with CSD in an election with a multi-card ballot. The number of cards per ballot, c, ranges from 1 to 5, while the partially diluted margins are fixed at $m_b = m_S$. Contests B and S appear on the same card.

Figure 6.7 plots the sample size needed under the same set-up as Figure 6.6, except that B and S appear on different cards.

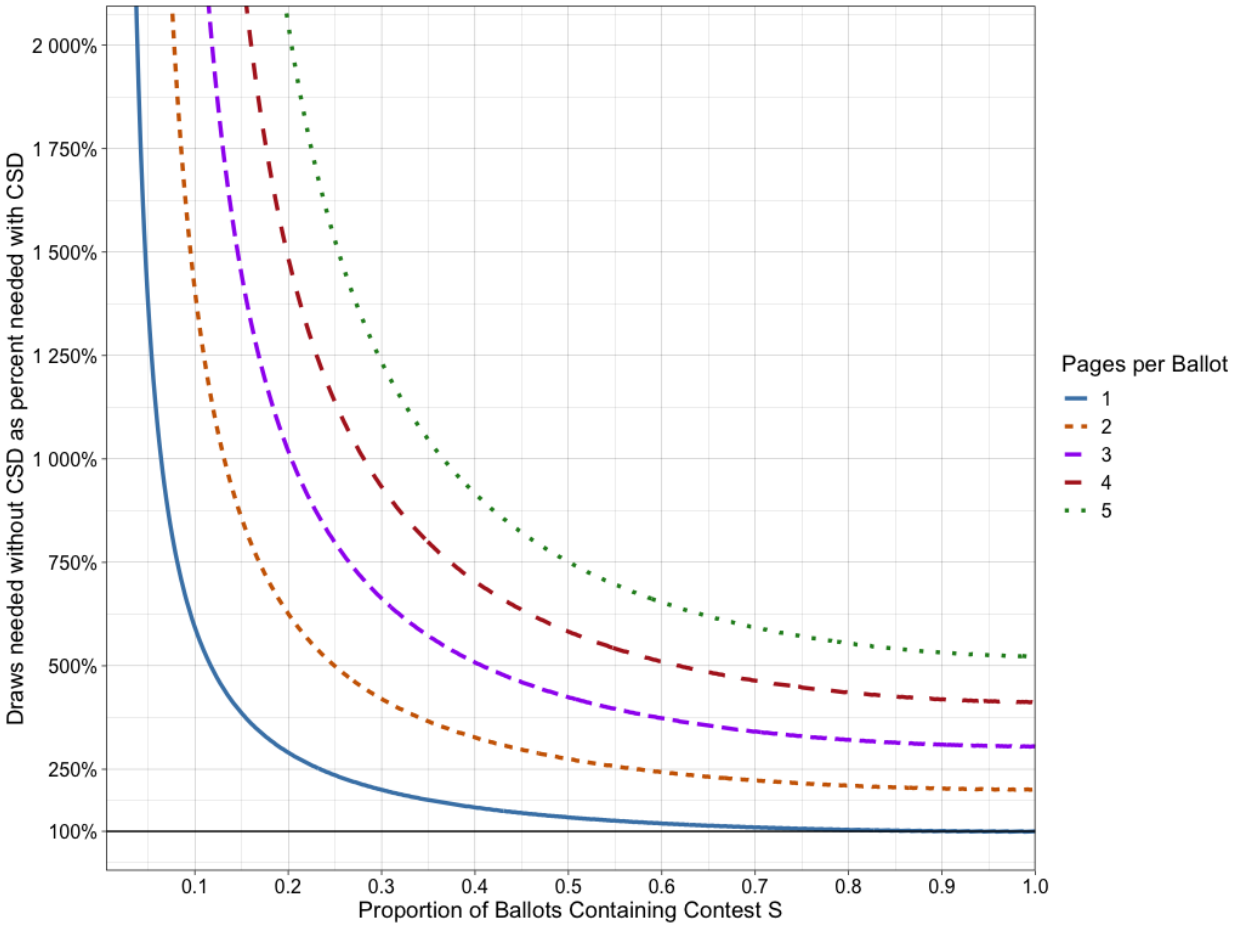

Figure 6.6: Draws needed to confirm the outcome in both contests (B and S) using the S4 method at risk limit $\alpha = 0.05$ without CSD as a multiple of the number of draws needed with CSD (y-axis), truncated at 2,000%. The contests appear on the *same* ballot-card and both partially diluted margins are fixed at 0.1. The lines indicate multiple needed; different lines correspond to different numbers of cards per ballot. The x-axis ranges over the proportion of ballots (not ballot cards) containing contest S, from 1 in 100 ($p = 0.01$) to all ballots ($p = 1.0$).

6.5 Ballot-polling audits

We have focused on comparison audits because we expect CSD will be derived from CVRs; when the voting system can produce CVRs linked to physical ballot cards, ballot-level comparison audits are possible and save a substantial amount of work compared to ballot-polling audits.

Jurisdictions that cannot conduct ballot-level comparison audits can still conduct ballot-polling audits. We show here that CSD (derived, for instance, by physically sorting ballot cards) can yield similar savings for ballot-polling audits. Whether the savings in audit effort is worth the effort of sorting the cards will depend on the jurisdiction's logistics, details of the contests under audit (including the partially diluted margins), and the number of cards

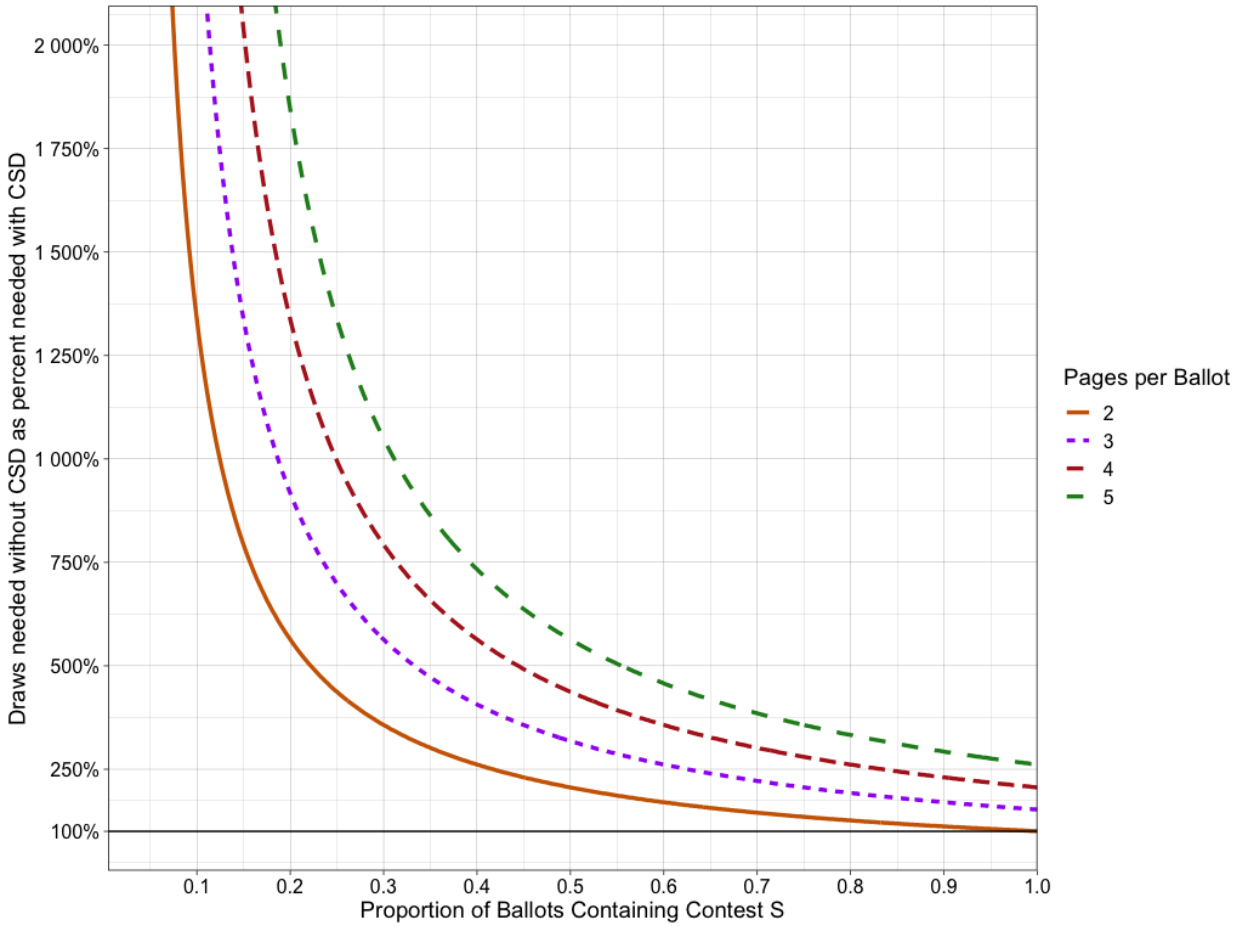

Figure 6.7: Draws needed to confirm outcome in both contests (B and S) when we do not have CSD expressed as a percentage of draws needed with CSD (y-axis), truncated at 2,000%. The contests appear on *different* pages, while both partially diluted margins are fixed at 0.1. The lines indicate percentage of draws needed, colored by the number of pages in the ballot. The x-axis ranges over the proportion of ballots (not ballot cards) containing contest S, from 1 in 100 ($p = 0.01$) to all ballots ($p = 1.0$).

per ballot.

For ballot-polling audits, the expected sample size scales like the square of the reciprocal of the margin and linearly in the reciprocal of the fraction of ballot cards that contain the contest. In other words, without CSD the sample sizes expected for ballot-polling audits scale linearly in $1/p$, the proportion of cards containing the smaller contest, and linearly in c, the number of cards per ballot—just as comparison audits do.

Consider again auditing contests B and S with margins M_B and M_S (in votes), N ballots cast each consisting of c cards, contest B on N of the Nc cards and S is on pN of the cards. For simplicity, assume that the contests are two-candidate plurality contests with no invalid votes; with small changes to the notation, the result can be generalized.

For the BRAVO method for ballot polling [Lindeman et al., 2012], the expected sample size is approximately $\frac{2\ln(1/\alpha)}{m^2}$ where m is the margin as a fraction (the margin in votes,

divided by the number of votes cast for the winner or the loser).

Suppose we know which *ballots* contain S but not which particular *cards* contain S, and that the c cards comprising each ballot are kept in the same container. This is an idealization of precinct-based voting where each voter in a precinct gets the same ballot style and casts all c cards of the ballot. If voters in that precinct are eligible to vote in S, a fraction $1/c$ of the cards in the container will have contest S; otherwise, none of the cards in the container will have S.

This lets us target the sampling for auditing contest S, but only partially: we can reduce the sampling universe for contest S from the original population of Nc cards to a smaller population of pNc cards, of which pN actually contain contest S. The sample for B would be drawn from all Nc cards, of which N contain contest B. The difference in expected sample sizes compared to "blind" ballot polling with no CSD is

$$2c\ln(1/\alpha)\left(\frac{1}{pm_S^2}-\frac{1}{m_B^2}-\frac{1}{m_S^2}\right) \text{ if contests } B \text{ and } S \text{ are on different cards.}$$

$$2c\ln(1/\alpha)\left(\frac{1}{pm_S^2}-(1-p)\frac{1}{m_B^2}-\frac{1}{m_S^2}\right) \text{ if contests } B \text{ and } S \text{ are on the same card.}$$

For a risk limit of 5% and a margin of 10%, using BRAVO, we would expect to sample 608 cards if we could target the sample. Consider the example $N = 10{,}000$, $p = 0.3$, $M_B = 0.1 = M_S$. If $c = 2$, absent CSD we would expect to sample $(2/0.3) \times 608 = 4{,}053$ cards. With partial CSD, but no information about which contests are contained on an individual card, we would expect to sample $(1-0.3)2 \times 608 + 2 \times 608 = 2{,}067$ cards if the contests are on the same card and $2 \times 608 + 2 \times 608 = 2{,}432$ ballot cards if the contests are on different cards. In either case, using CSD reduces the sample size by roughly half.

Thus, information about which containers have which card styles—even without information about which cards contain which contests—can still yield substantial efficiency gains for ballot-polling audits.

6.6 Case studies

We will give numerical examples for Inyo County, California (a small county), and Orange County, California (a large county), both of which have conducted several RLA pilots or binding RLAs.

6.6.1 Inyo County, California

Inyo County (population 18,546 according to the 2010 U.S. Census) is in eastern California. The county conducted a pilot RLA of an April 2018 special election using the S4 method Foote [2018] (it has conducted other RLAs as well, including RLAs of 7 contests in the 2020 general election).

In the June 2018 election each ballot consisted of two cards ($c = 2$). We consider two contests on this ballot: Supervisor District 1 (contest S) and U.S. Senator (contest B). (We shall pretend that the Senate contest was entirely contained in Inyo County.) Of the 5,919 ballots cast in the June 2018 election, 1,435 ($p = 0.24$) contained the Supervisor District 1

contest. This contest had a partially diluted margin of $m_S = 36/1{,}435 = 0.025$. In California primaries, the top *two* candidates with the most votes (from either party) advance to the general election, so the margin between the 2nd place candidate and the 3rd place candidate drives the audit effort. All voters in Inyo County were eligible to vote in the U.S. Senate contest. Diane Feinstein received the most votes, 1,555, followed by James Bradley with 639, and Paul Taylor with 517. We consider confirming that Feinstein and Bradley received more votes than Taylor. The partially diluted margin for this contest is

$$m_B = (639 - 517)/5{,}919 = 0.02 > 0.006 = pm_S.$$

First, consider a comparison audit. Without CSD, we would expect to sample 3,734 cards. The Supervisor District 1 and U.S. Senate contests were not on the same card, so with CSD we would expect to sample 588 cards, smaller by about $3{,}734 - 588 = 3{,}146$ cards than an audit without CSD.

Suppose instead we audit using BRAVO, a ballot-polling method. Without CSD we would expect to audit $2{,}796 \times 2$ ballot cards to verify contest B, but to confirm the outcome of contest S would essentially require a full hand count of the votes on all 11,838 cards. If we knew which containers had cards that include contest S, we would expect to audit $2{,}796 \times 2 + 1{,}435 \times 2 = 8{,}462$ cards. (While this is a substantial reduction, it is probably more efficient to conduct a full hand count than to examine a majority of ballot cards selected randomly.)

6.6.2 Orange County, California

Orange County, in Southern California, is much larger (3.017 million residents according to the 2010 U.S. Census). The ballots in their 2018 election consisted of $c = 2$ cards. A BRAVO RLA of three of the five countywide contests in 2018 was conducted Singer and McBurnett [2018].

Suppose we wanted to audit the Senate (Feinstein v De Leon) and 45th District Congressional (Porter v Walters) contests in Orange County from the November 2018 election. (As before, we shall pretend for the purpose of illustration that the senatorial contest is entirely contained in Orange County.) All voters were eligible to vote in the Senate contest. Of the 1,106,729 ballots cast, 312,700 (28.25%) included the 45th District Congressional contest. In this case $p \approx 0.2825$, $m_S = 0.04$, $m_B = 0.073$, and $c = 2$, so

$$m_B = 0.073 > 0.012 = pm_S.$$

First consider a ballot-level comparison audit. Without CSD, we would expect to have to sample 1,452 cards. With CSD, we would expect to have to sample 249 cards if the contests were on different cards, and 225 if they were on the same card. These two contests were in fact on the same ballot card, so using CSD would decrease the expected number of audited cards by $1{,}452 - 225 = 1{,}227$.

Suppose instead we audit using the BRAVO ballot-polling method. If we did not have CSD, we would expect to have to sample $2 * 3{,}671/0.2825 = 25{,}989$ ballot cards. If ballots were organized by ballot style (but not card style), we would expect to sample $(1 - 0.2825) \times 2 \times 948 + 2 \times 3{,}671 = 8{,}702$ cards.

Table 6.1 summarizes the results of this section, the expected percentage reduction in number of ballot cards drawn with CSD versus without CSD.

	Comparison Audit	Ballot-Polling Audit
Inyo County	84%	29%
Orange County	85%	67%

Table 6.1: Expected percentage reduction in the sample size required with CSD versus without CSD to audit the contests described in Section 6.6.

6.7 Implementation

To minimize the number of cards a risk-limiting audit of multiple contests needs to inspect, we would like to be able to use any card in the audit sample to audit every contest that card contains. Standard methods for ballot-level comparison audits or ballot-polling audits then require that the intersection of the sample with the cards that contain each contest is a uniform random sample from that contest (either with or without replacement, as appropriate for the auditing method). When the contests are on some of the same cards, this requires particular care. Figure 6.8 sketches the possibilities.

One method for ensuring uniformity on overlapping subsets is called *consistent sampling*. Consistent sampling can be done with replacement [Rivest, 2018a] or without replacement [Broder, 1997, Broder et al., 1997]. It is simpler without replacement, which also leads to more efficient RLAs [Stark, 2020]. Figure 6.9 illustrates consistent sampling with a toy example of auditing two contests in a small election.

Here we show how to audit K contests of different sizes using CSD and consistent sampling. We assume that (1) there has been a *compliance audit* to establish that the paper trail is trustworthy [Benaloh et al., 2011, Stark and Wagner, 2012, Stark, 2018, Appel and Stark, 2020]; (2) there is an upper bound on the number of cards that contain each contest under audit, obtained, for example, from pollbook signatures and other administrative records; and (3) the underlying auditing method (comparison vs ballot-polling, and the "risk function"), sampling method, and risk limits ($\{\alpha_1, ..\alpha_K\}$) have been chosen, along with rules for picking the initial sample size(s) $\{S_1, ..., S_K\}$ and for increasing the sample size if the audit does not confirm the outcome with the initial sample.

1. If there are more CVRs that contain any contest than the upper bound on the number of cards that contain the contest, stop: the contest outcome cannot be confirmed.

2. If the upper bound on the number of cards that contain a contest is greater than the number of CVRs that contain the contest, create a corresponding set of "phantom" CVRs as described in section 3.4 of Stark [2020].

3. If the upper bound on the number of cards that contain a contest is greater than the number of physical cards whose locations are known, create enough "phantom" cards to make up the difference.

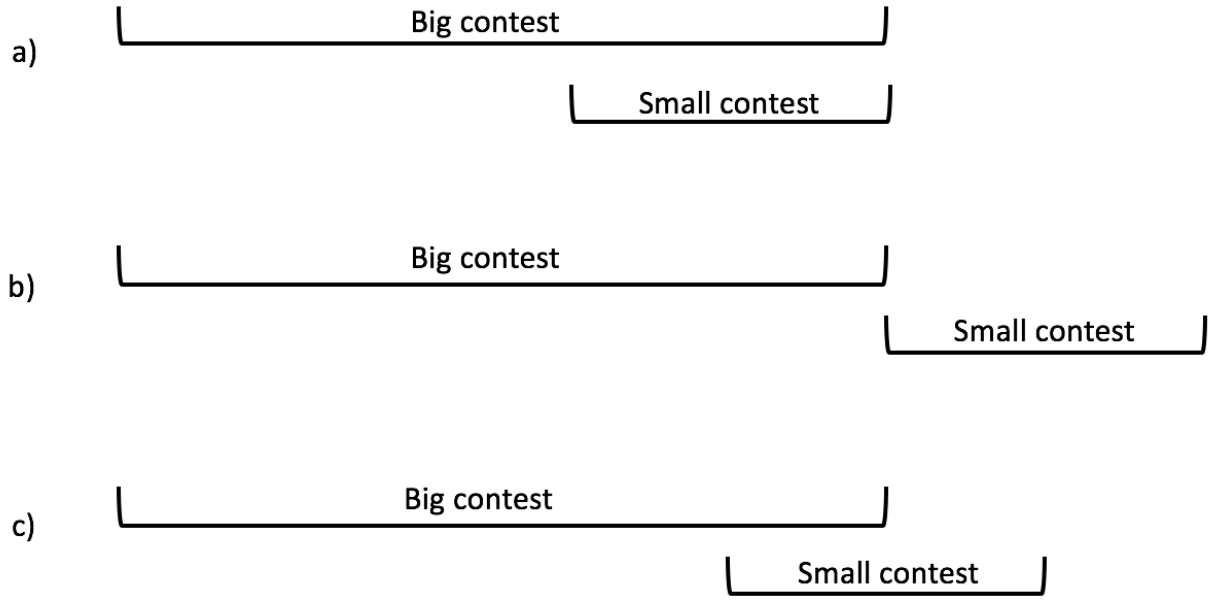

Figure 6.8: Overlap of card styles in 3 hypothetical elections with 2 contests (one big and one small). Case a represents complete nesting, where uniform draws from the big contest yield uniform draws from the small contest. If all ballots consist of a single ballot card and the big contest is on all ballots, the situation is case a. Case b represents a scenario where ballot cards that contain the big contest do not contain the small contest. Auditing the big contest tells us nothing about the small contest. In case c, the small contest sometimes appears on the card for the big contest and sometimes not. Drawing ballot cards uniformly from the big contest does not sample from the entire small contest.

4. Assign a $U[0,1]$ pseudo-random number to every ballot card that contains one or more contests under audit (including "phantom" cards), using a high-quality PRNG [Ottoboni and Stark, 2019]. For sampling without replacement, assign one number to each card; for sampling with replacement, assign several, as described in Rivest [2018a].

5. Initialize $\mathcal{A}$ to be the set of contests under audit: $\mathcal{A} \leftarrow \{1, \ldots, K\}$.

6. While $\mathcal{A}$ is not empty:

   (a) Pick the sample sizes $\{S_k\}$ for $k \in \mathcal{A}$ for this round of sampling.

   (b) Choose thresholds $\{t_k\}_{k \in \mathcal{A}}$ so that S_k ballot cards containing contest k have numbers less than or equal to t_k.

   (c) Retrieve any of the corresponding ballot cards that have not yet been audited and inspect them manually. If there is no CVR for the ballot, treat the CVR as if it recorded a non-vote in every contest still under audit. If a ballot card cannot be found or if it is a phantom card, treat it in the way that casts the most doubt on

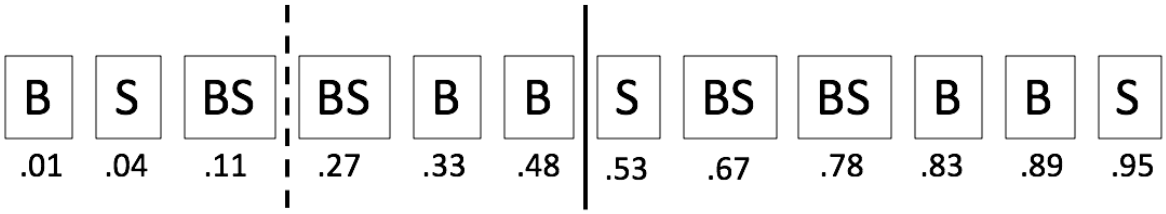

Figure 6.9: Illustration of consistent random sampling without replacement to audit 2 contests, B and S. Contest B appears on 9 cards and S appears on 7 cards; 12 cards were cast in all. There is partial overlap (case c of Figure 6.8): contest B appears alone on 5 cards (labelled "B"), the contests appear together on 4 cards (labelled "BS"), and contest S appears alone on 3 cards (labelled "S"). To perform consistent sampling, each card is assigned a random number in $[0, 1]$ independently across cards. The number appears below each card; without loss of generality, we have sorted the cards by increasing order of the random number. Suppose the audit requires inspecting 5 cards containing B and 2 cards containing S. The 5 sampled cards containing B are the cards containing B that were assigned the 5 smallest numbers, i.e., all cards containing B to the left of the second (solid) vertical bar. The 2 sampled cards containing S are the cards containing S that were assigned the 2 smallest random numbers, i.e., all cards containing S to the left of the first (dashed) vertical bar. This approach ensures that if we look at a card to audit one contest and the other contest is also on the card, it can be used in the audit of that contest, too: the sample is uniformly distributed on subsets of the population. This approach can be generalized to sampling with replacement; see Rivest [2018a].

the outcome of every audited contest it was supposed to contain (see Section 3.4 of Stark [2020]).[9]

(d) Use the data from the previous step to update the measured risk for every contest $k \in \mathcal{A}$.

(e) Remove from $\mathcal{A}$ all k that have met their risk limits.

6.8 Conclusions

Card-style data (CSD) can dramatically increase the efficiency of risk-limiting audits. For the super-simple simultaneous single-ballot audit of Stark [2010] and for ballot-polling audits using BRAVO [Lindeman et al., 2012], the expected reduction in sample size can easily be several orders of magnitude, depending on the range of sizes and margins of the contests under audit and the number of cards per ballot.

When a contest is on only a small fraction of the cards cast in an election and the sample

[9]Some cards may be selected for auditing more than one contest. If the sample is drawn with replacement, the same card may be selected more than once. Such cards are only manually inspected once, even though their data might be re-used.

is drawn from all cast cards, auditing the contest can be nearly as much work as a full manual tally involving *all* cast cards, not just those that contain the contest—even if the margin in the contest is large. In contrast, if there are CSD so that the sample can be drawn from just those cards that (reportedly) contain the contest, auditing small contests can be quite efficient.

Jurisdictions that perform RLAs should consider using CSD to reduce the workload of auditing small contests and contests on multi-card ballots: the savings can be substantial. Jurisdictions that can conduct ballot-level comparison audits (i.e., jurisdictions with voting systems that can export CVRs linked to physical ballots) can construct CSD with no additional effort, because the CVRs contain CSD—albeit, possibly erroneous. However, to ensure that errors in the CVRs resulting in errors in CSD do not compromise the risk limit, jurisdictions also need an upper bound on the total number of ballot cards that contain each contest, information that can be derived from pollbooks and related administrative voter records.

Chapter 7

Stylish Risk-Limiting Audits in Practice[1]

7.1 Introduction

Risk-limiting audits (RLAs) manually inspect ballots from a trustworthy record of the votes[2] to provide affirmative evidence that electoral outcomes (i.e., who won, not the exact vote counts) are correct if they are indeed correct, and (with a prespecified minimum probability) to correct any outcomes that are wrong. The maximum chance that an RLA does not correct a result that is wrong is the *risk limit*. For example, an RLA with a risk limit of 5% guarantees that if the reported outcome is wrong, the audit has at least a 95% chance of catching and correcting the reported outcome before it is certified. When the outcome is correct, RLAs may inspect only a small fraction of all ballot cards, saving considerable labor compared to a full manual recount.

According to the 2018 National Academies of Science, Engineering, and Medicine report *Securing the Vote: Protecting American Democracy* [National Academies of Sciences, Engineering, and Medicine, 2018b, Recommendation 5.8]:

> States should mandate risk-limiting audits prior to the certification of election results. With current technology, this requires the use of paper ballots. States and local jurisdictions should implement risk-limiting audits within a decade. They should begin with pilot programs and work toward full implementation. Risk-limiting audits should be conducted for all federal and state election contests, and for local contests where feasible.

No jurisdiction currently mandates RLAs of every contest in every election, or even every federal and statewide contest. For example, Georgia law only requires auditing one contest every two years, and Colorado law requires auditing two contests in each election. While some officials claim that such sparse or infrequent auditing shows that their voting systems work

[1] This chapter comprises a publication [Glazer et al., 2023b] co-authored by Jacob V. Spertus and Philip B. Stark.

[2] Not all paper vote records are trustworthy. See, e.g., Appel et al. [2020], Appel and Stark [2020]. Absent a trustworthy record of the vote, no audit can provide affirmative evidence that the reported winners really won.

flawlessly,[3] auditing one reported outcome says nothing about whether any other reported outcome: every contest should get some scrutiny (or at least have a high probability of being audited).

Historically, auditing local contests together with jurisdiction-wide contests using a single audit sample has been infeasible. Indeed, when some contests are small and others are jurisdiction-wide, RLA methods that sample ballots uniformly at random would require a full hand count throughout the jurisdiction, even when every margin (as a percentage of votes in the contest) is large.

However, Glazer et al. [2021] presented an approach to RLA sampling that allows many contests of different sizes to be audited efficiently using the same sample. Instead of sampling cards uniformly at random, the method uses *card-style data* (CSD) and *consistent sampling* to ensure that each contest gets the scrutiny it needs, without entailing unnecessary scrutiny of other contests. They illustrated their method with simplified examples involving only two contests, but in the U.S., there can be hundreds of contests in a single election.

We incorporated the Glazer et al. [2021] method into the SHANGRLA Python RLA library,[4] leveraging recent developments in formulating RLAs as hypothesis tests about the means of bounded, finite lists of numbers Stark [2020] and efficiently measuring risk using test supermartingales Spertus [2023], Stark [2023b], Waudby-Smith et al. [2021]. To illustrate the practical implications of CSD, we applied the method to historical data from the 2020 and 2022 general elections in Orange County, CA, which comprised 181 contests and 214 contests, respectively. In both elections, standard RLA methods would have required a full hand count to audit every contest to a risk limit of 5%. The new method reduces the estimated audit workload by more than 99% for the 2020 election and by 97% for the 2022 election.

The next section reviews terminology, describes the problem, and summarizes the building blocks, including simultaneous card-level comparison audits of multi-style elections. Section 7.3 provides a high level description of our software. Section 7.4 describes the 2020 and 2022 Orange County elections, gives an overview of our implementation, and presents sample size estimates for RLAs with and without CSD. Code that produced our results is available at `https://github.com/pbstark/SHANGRLA`. Section 7.5, discusses ramifications for real-world RLAs and provides recommendations for practice.

7.2 Background

In the U.S., a *ballot* consists of one or more *cards*, individual pieces of paper. (U.S. elections often contain more contests than can be printed on a single piece of paper in a readable font.) Each card has a *style*, which for our purposes is the collection of contests on that card. Because ballot boxes are generally designed so that the cards do not land in

[3] See, e.g., Georgia Secretary of State Brad Raffensperger's claims that the audit of one contest in 2020 "reaffirmed that the state's new secure paper ballot voting system accurately counted and reported results." `https://sos.ga.gov/news/historic-first-statewide-audit-paper-ballots-upholds-result-presidential-race` (last visited 2 May 2023) and that the audit of one contest in 2022 "shows that our system works and that our county election officials conducted a secure, accurate election." `https://sos.ga.gov/news/georgias-2022-statewide-risk-limiting-audit-confirms-results` (last visited 2 May 2023)

[4] `https://github.com/pbstark/SHANGRLA`

the order in which they were cast, it is typically impossible to reassemble a ballot from its component cards once it has been cast. Thus cards, not ballots, are the atomic sampling unit in RLAs.

When ballots have multiple cards, no contest is on more than half the cards. Contests that are not jurisdiction-wide are on even fewer cards. Following the terminology of Stark [2023b], the *sampling domain* of a contest is the population from which cards are sampled in an RLA. For the RLA to be valid, the sampling domain for a given contest generally must include every card that contains that contest. In practice, the sampling domain for RLAs has been either all cards cast in the election, or just the cards containing a particular contest. When the sampling domain for a contest includes cards that do not contain the contest, the audit generally needs to examine more ballots (when the outcome is correct) than it would if the sampling domain were limited to cards containing the contest.

In particular, audits that directly check the voting system's interpretation and tabulation of votes are more efficient when the sampling domain is limited to cards that contain the contest because the error rate (per card) required to alter the outcome is smaller the larger the denominator (the sampling domain) is. Testing whether the error rate is below a small threshold requires more data than testing whether it is below a larger threshold.

The *diluted margin*, the margin in votes divided by the number of cards in the sampling domain for the contest, captures this phenomenon. Smaller diluted margins lead to larger audit sample sizes; expanding the sampling domain increases the "dilution," reducing the diluted margin.

7.2.1 Card-level Comparison Audits and Card-Style Data

RLAs can use data from voting systems and from manually inspected cards in a number of ways. RLAs that check for error by comparing ballots to their machine interpretations are called *comparison* audits; those that check outcomes without relying on the voting system's interpretations are *polling* audits. Furthermore, RLAs may sample and check vote totals for *batches* of ballot cards—typically machines or precincts—or individual cards. Adopting the terminology of Stark [2023 (in press], we refer to audits that sample individual cards and compare a human reading of the votes on each sampled card to the CVR for that card as *card-level* audits. The literature sometimes often calles these *ballot-level*, but card-level is more accurate nomenclature because CVRs are generally for individual cards, and ballots comprise more than one card in many jurisdictions.

All else equal, card-level audits are more efficient than batch-level audits; comparison audits are more efficient than polling audits; and *card-level comparison audits* are the most efficient approach. In a card-level comparison audit, the estimated sample size scales with the reciprocal of the diluted margin.

To conduct a card-level comparison audit, the voting system must produce *cast-vote records* (CVRs)—the system's record of the votes on each card. (However, see Stark [2023 (in press], which shows how to conduct a card-level comparison audit using "overstatement-net-equivalent" CVRs derived from batch-level results.) Moreover, there must be a known 1:1 mapping from the physical card to its particular CVR; some voting systems cannot provide such a link, or cannot provide a link without compromising voter privacy. That link might be provided by exporting CVRs in the order in which the cards were scanned and keeping

the cards in that order, or by imprinting a number on each card before or as it is scanned, and including the imprinted number in the CVR for that ballot or as part of the filename of the CVR. Glazer et al. [2021] showed that an audit can rely on CVRs to infer CSD: consider a card to contain a contest if the CVR for that card contains the contest. Even though the CVRs might be inaccurate or incomplete (otherwise, no audit would be needed), their method ensures that errors in CSD derived from CVRs do not compromise the risk limit. CSD makes it possible to minimize the sampling domain (maximizing the diluted margin) for each contest, considerably lowering audit workloads when contest outcomes are correct.

Glazer et al. [2021] also showed how to combine CSD with *consistent sampling* Rivest [2018a], which ensures that cards drawn for the purpose of auditing a given contest can also be used in the audit of other contests that appear on the sampled cards. Exploiting such overlap further reduces the estimated workload. If the voting system does not provide CVRs linked to cards, CSD can be derived by manually sorting the cards (a very labor intensive alternative), or by processing them in homogeneous batches in the first place. That is straightforward for precinct-based voting systems, but many jurisdictions do not sort cards by style before scanning them.

When CSD are derived from CVRs, the RLA can also use those CVRs for card-level comparison auditing, which is much more efficient than ballot-polling or batch-level comparison audits. For this reason, the audits we describe in the remainder of this paper are card-level comparison audits, but the software also supports ballot-polling audits with CSD. We now describe how the software is implemented to run a card-level comparison audit.

In broad brush, the procedure imports audit parameters (such as risk limits, risk-measuring functions, and the strategy for estimating the initial sample size), election data (including the reported winners, the CVRs, and upper bounds on the number of cards cast in each contest[5]), and contest-specific data (such as candidate names, and the social choice function).

CSD is inferred from the CVRs. The CVRs are checked for consistency with the other inputs. An initial sample size is determined for each contest, which implies a sampling probability for each card that contains the contest. The probability that a given card is sampled is the largest sampling probability for each audited contest on the card. Summing those maximum probabilities across cards is an estimate of the total initial sample size. A sample is drawn using *consistent sampling*; the corresponding cards are retrieved and interpreted manually; the resulting *manual vote records* (MVRs) are imported; the attained risk for each audited contest is calculated; and any contests for which the risk limit has been attained or for which there has been a full hand count are removed from future auditing rounds. If every audited contest has been removed, the audit stops; otherwise, a next-round sample size is determined for the remaining contests, and the process repeats.

In more detail, the algorithm is as follows (adapted from Glazer et al. [2021]):

1. Set up the audit
    (a) Read contest descriptors, candidate names, social choice functions, upper bounds on the number of cards that contain each contest, and reported winners. Let

[5]All RLAs require information that a good canvass should generate routinely, including upper bounds on the number of validly cast cards that contain each contest, which can be derived from registration data, voter participation data, and physical inventories of ballot cards. Absent that information, even a "full" hand count is meaningless, since there is no way to know whether the count includes every validly cast ballot.

N_c denote the upper bound on the number of cards that contain contest c, $c = 1, \ldots, C$.

(b) Read audit parameters (risk limit for each contest, risk-measuring function to use for each contest, assumptions about errors for computing initial sample sizes), and seed for pseudo-random sampling.

(c) Read ballot manifest.

(d) Read CVRs.

2. Pre-processing and consistency checks

(a) Check that the winners according to the CVRs are the reported winners.

(b) If there are more CVRs that contain any contest than the upper bound on the number of cards that contain the contest, stop: something is seriously wrong.

(c) If the upper bound on the number of cards that contain any contest is greater than the number of CVRs that contain the contest, create a corresponding set of "phantom" CVRs as described in section 3.4 of Stark [2020]. The phantom CVRs are generated separately for each contest: each phantom card contains only one contest.

(d) If the upper bound N_c on the number of cards that contain contest c is greater than the number of physical cards whose locations are known, create enough "phantom" cards to make up the difference.

3. Prepare for sampling

(a) Generate a set of SHANGRLA Stark [2020] assertions $\mathcal{A}_c$ for every contest c under audit.

(b) Initialize $\mathcal{A} \leftarrow \cup_{c=1}^{C} \mathcal{A}_c$ and $\mathcal{C} \leftarrow \{1, \ldots, C\}$.

(c) Assign independent uniform pseudo-random numbers to CVRs that contain one or more contests under audit (including "phantom" CVRs), using a high-quality PRNG Ottoboni and Stark [2019]. (The code uses cryptographic quality pseudo-random integers uniformly distributed on $0, \ldots, 2^{256} - 1$.) Let u_i denote the number assigned to CVR i.

4. Main audit loop. While $\mathcal{A}$ is not empty:

(a) Pick the (cumulative) sample sizes $\{S_c\}$ for $c \in \mathcal{C}$ to attain by the end of this round of sampling. The software offers several options for picking $\{S_c\}$, including some based on simulation. The desired sampling fraction $f_c := S_c/N_c$ for contest c is the sampling probability for each card that contains contest k, treating cards already in the sample as having sampling probability 1. The probability p_i that previously unsampled card i is sampled in the next round is the largest of those probabilities: $p_i := \max_{c \in \mathcal{C} \cap \mathcal{C}_i} f_c$, where $\mathcal{C}_i$ denotes the contests on card i.

(b) Estimate the total sample size to be $\sum_i p_i$, where the sum is across all cards except phantom cards.

(c) Choose thresholds $\{t_c\}_{c \in \mathcal{C}}$ so that S_c ballot cards containing contest c have a sample number u_i less than or equal to t_c.

(d) Retrieve any of the corresponding ballot cards that have not yet been audited and inspect them manually to generate MVRs.

(e) Import the MVRs.

(f) For each MVR i:
For each $c \in \mathcal{C}$:
If $u_i \leq t_c$, then for each $a \in \mathcal{A}_c \cap \mathcal{A}$:

- If the ith CVR is a phantom, define $a(\text{CVR}_i) := 1/2$.
- If card i cannot be found or if it is a phantom, define $a(\text{MVR}_i) := 0$.
- Find the overstatement of assertion a for CVR i, $a(\text{CVR}_i) - a(\text{MVR}_i)$.

(g) Use the overstatement data from the previous step to update the measured risk for every assertion $a \in \mathcal{A}$.

(h) Optionally, conduct a full hand count of one or more contests, for instance, if the audit data suggest the outcome is wrong or if the auditors think a hand count will be less work than continuing to sample.

(i) Remove from $\mathcal{A}$ all assertions a that have met their risk limits or that are for contests for which there has been a full hand count. (The audits of those assertions are complete.)

(j) Remove from $\mathcal{C}$ all contests c for which $\mathcal{A}_c \cap \mathcal{A} = \emptyset$ (the audits of those contests are complete).

5. Replace the reported outcomes of any contests that were fully hand counted by the outcomes according to those hand counts.

7.3 Software

The software can read Dominion Democracy Suite® and Hart InterCivic Verity CVRs and manifest files. Because file sizes in large jurisdictions can be unwieldy, the software can read compressed CVR files (.zip format containing XML records).

Figure 7.1 sketches the workflow to audit a collection of contests using CSD derived from CVRs. The user specifies parameters of the audit and the contests to be audited, including paths to data and output files, a trustworthy upper bound on the number of cards cast (e.g., a bound from participation records, ballot accounting, pollbook reconcilation, etc.—not the voting system's own reported number of cards), contest information, risk limits, risk measuring functions and their tuning parameters (defaults are available), information used to estimate initial sample sizes (defaults are available), and whether to use CSD.

The software then constructs SHANGRLA assertions (or reads RAIRE assertions in json for IRV contests), reads CVRs and manifests, constructs "phantom" CVRs to account for missing cards if necessary, sets margins for overstatement assorters, estimates initial sample sizes, draws random ballots by consistent sampling, and returns their locations to the auditors.

The auditors retrieve the indicated cards, manually read the votes from those cards, and input the MVRs, which the audit software subsequently reads from a file. The software uses the specified risk-measuring function(s), the CVRs, and the MVRs to compute a P-value

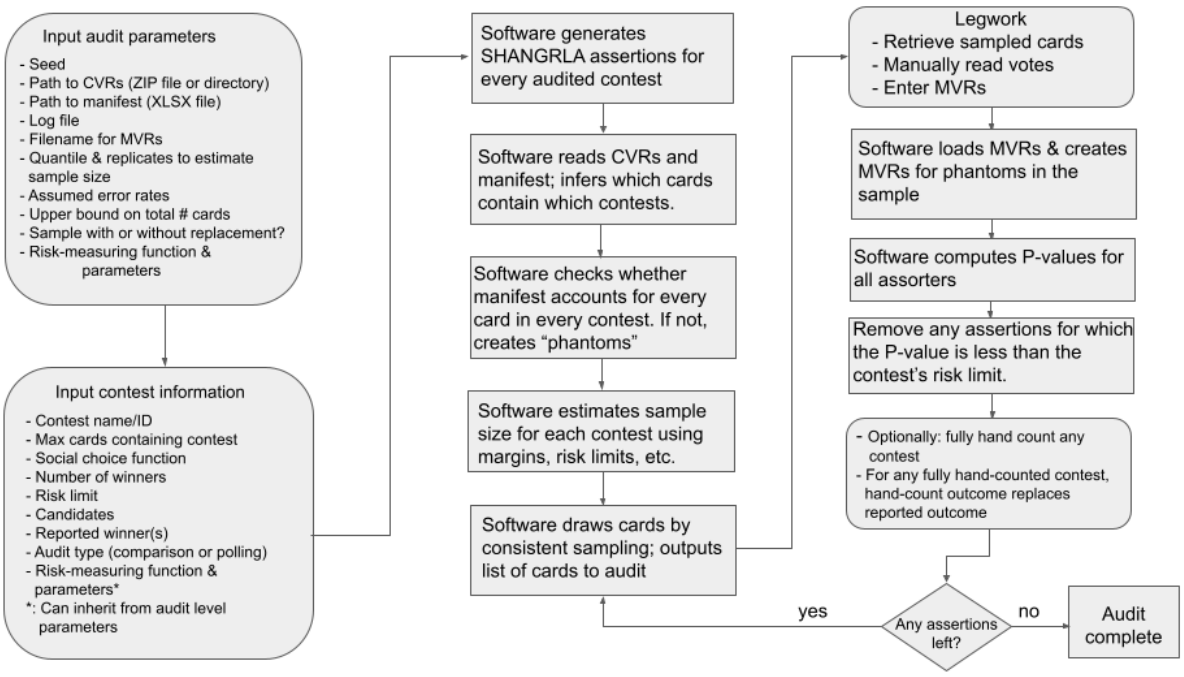

Figure 7.1: Workflow for simultaneous card-level comparison audit using SHANGRLA software with CSD and consistent sampling. Boxes with rounded corners involve inputs from the auditors.

for every assertion. All assertions with P-values below the risk limit for their corresponding contests are considered "confirmed." If any assertions remain unconfirmed, the software will estimate the number of additional cards to examine to confirm those assertions, draw a sample of that size, and export the identifiers of the ballot cards for the auditors to retrieve and interpret.

This process repeats until every assertion has been confirmed or there has been a full hand count of the contest. At any point, the auditors can choose to stop sampling at random and simply tabulate the rest of the votes in one or more contests (e.g., if they judge that that would be more efficient, or if the audit sample indicates that the reported outcome is in fact wrong).

7.4 Orange County Election Audits

Orange County, CA, is the third most populous county in California (3.19 million as of the 2020 census, with over 1.81 million active voters[6]). It has more registered voters than 24 U.S. states, and is the country's fifth-largest election jurisdiction, after Los Angeles, CA; Maricopa, AZ; Harris, TX; and San Diego, CA.[7] As of this writing, Orange County has

[6] `https://ocvote.gov/datacentral/`, last visited 8 June 2022

[7] 2022 Election Administration and Voting Survey (EAVS), U.S. Election Assistance Commission, `https://www.eac.gov/sites/default/files/2023-06/2022_EAVS_for_Public_Release_V1.xlsx`, released 29 June 2023. Last visited 9 July 2023.

2204 precincts and approximately 181 voting centers. Orange County uses the Hart Intercivic Verity system. The county first piloted an RLA in 2011,[8] conducted two additional pilots in 2018[9] and seven pilots between 2020 and 2022 mandated by California Elec. Code, §§ 15365–15367.

In this paper, we use data from the November 2020 and 2022 General Elections. We estimate the number of cards that would need to be inspected for an RLA with 5% risk limit, with and without style information. (While many states that require or authorize RLAs do not specify a risk limit in statute, 5% is a common value in practice. It was the statutory requirement in California's pilot program, Cal. Elec. Code § 15367. Sample sizes for RLAs generally scale approximately like the log of the risk limit, so sample sizes for a risk limit of 1% would be about $\log(0.01)/\log(0.05) - 1 \approx 54\%$ larger.) Table 7.1 summarizes these elections and the results of our calculations for all contests, and for all contests with margins greater than 0.1%, 0.5%, or 1%. Section 7.4.1 that because of automatic manual recount laws in various states, it may make sense to omit contests with small margins from the workload estimates.

To audit cross-jurisdictional contests requires sampling from all cards cast in the contest, not just those cast in one jurisdiction. Since we did not have access to CVRs for other counties, the sample size estimates we report treat every contest in both elections as if it were entirely contained in Orange County. In particular, the estimates take the margins of statewide contests to be the margins within Orange County alone, and ignore the fact that the resulting audit burden would be shared across all jurisdictions with voters eligible to vote in those contests. The results still give a reasonable estimate of the workload required to audit a large number of (partially) overlapping contests simultaneously, and it is generally the *smaller* contests that drive audit workload for audits that do not use CSD, for reasons explained above. In particular, statewide contests appear on every ballot in each jurisdiction and on approximately the same fraction of cards in each jurisdiction (depending on the number of local contests in each jurisdiction). Moreover, because Orange County has more registered voters than 24 U.S. states, it is a reasonable proxy for many statewide audits.

The actual sample size depends on the luck of the draw—which particular cards end up in the sample—and on the errors in the CVRs for those cards. We estimate sample sizes using two assumed error rates: no error at all, and one 1-vote overstatement per 1,000 cards, i.e., a rate of 10^{-3}. (A one-vote overstatement occurs if the CVR has an error that increased the margin of a reported winner over a reported loser by one vote, e.g., if the card shows a valid vote for a loser but the CVR shows an undervote or overvote, or if the card shows an overvote, but the CVR shows a valid vote for a reported winner.) When CVRs are error-free, the sample size is deterministic. For the assumed rate of 10^{-3}, we generated artificial data that reflects a one-vote overstatement error every 1,000 ballots, starting with an error in the first CVR in the sample. The risk-measuring function is the ALPHA supermartingale Stark [2023b], with the `optimal_comparison` estimator of Spertus [2023]. That estimator depends on an assumed rate of 2-vote overstatement errors in the CVRs, which we set to 10^{-4}.

[8]See California Secretary of State Report to the US Election Assistance Commission, `https://admin.cdn.sos.ca.gov/reports/2011/post-election-audit-report-20111130.pdf`, last visited 15 May 2023.

[9]See `https://verifiedvoting.org/wp-content/uploads/2020/08/2018-RLA-Report-Orange-County-CA.pdf`, last visited 15 May 2023.

	Year	2020		2022	
1	**Turnout**	1,546,570		994,227	
2	**Cards cast**	3,094,308		1,989,416	
3	**Total contests**	181		214	
4	**Exact ties**	1		0	
5	**Margins below 0.1%**	1		3	
6	**Margins below 0.5%**	4		9	
7	**Margins below 1.0%**	5		14	
	Sample sizes	**rate of 1-vote overstatements**			
		0	10^{-3}	0	10^{-3}
8	**all contests**	20,112	37,996	62,251	119,814
9		(0.6%)	(1.2%)	(3.1%)	(6.0%)
10	**omit margins ≤0.1%**	15,964	33,852	22,110	33,215
11		(0.5%)	(1.1%)	(1.1%)	(1.7%)
12	**omit margins ≤0.5%**	9,228	11,347	11,053	14,125
13		(0.3%)	(0.4%)	(0.6%)	(0.7%)
14	**omit margins ≤1%**	7,827	9,634	8,123	9,980
15		(0.3%)	(0.3%)	(0.4%)	(0.5%)

Table 7.1: Summary of the 2020 and 2022 General Elections in Orange County, CA. Row 4 is the number of contests reported to be tied. Rows 5–7 are the number of contests with reported margins below 0.1%, 0.5%, and 1%, respectively. Rows 8–15 are the sample sizes to confirm all contests to a risk limit of 5%, expressed as the number of cards (rows 8, 10, 12, 14) or the percentage of all cards (in parentheses, rows 9, 11, 13, 15), when the audit finds no errors, or when the rate of one-vote overstatement errors is 1 in 1,000 CVRs. (A one-vote overstatement occurs if correcting the error reduces the margin between a reported winner and a reported loser by one vote, e.g., if the CVR erroneously counts a vote for the reported loser as an undervote.) When there is no error, the sample size is deterministic. When there are errors, the sample size depends not only on their rate, but on the order in which they occur. To simplify the calculations, we estimate the sample size by assuming that the first CVR shows an error, and thereafter errors are equispaced, one every 1,000 ballots. Rows 8 and 9 are for all contests, including tied contests. Rows 10 and 11 exclude contests with reported margins less than or equal to 0.1%, a threshold some states use for automatic recounts (see section 7.4.1). Rows 12 and 13 exclude contests with margins less than or equal to 0.5%, a common threshold for automatic recounts. Rows 14 and 15 exclude contests with margins less than or equal to 1%, another common automatic recount threshold. For the purpose of illustration, the calculations assume that every contest (including statewide contests) is entirely contained in Orange County.

7.4.1 Audits and Recounts

If a jurisdiction conducts a *manual* recount of the ballots after a robust canvass, there is no need for an RLA (some states allow machine recounts). Many U.S. states conduct automatic recounts for contests with small reported margins. Alabama, Arizona, Colorado,

Connecticut, Delaware, Florida, New York, Ohio, Pennsylvania, and Washington recount contests with margins less than 0.5% (possibly with exceptions).[10] Hawaii automatically recount if the margin is below 0.25%. Nebraska and Wyoming have automatic recounts if the margin is less than 1% of the winner's tally. New Mexico and North Dakota automatically recount elections with margins less than 1%, 0.5%, or 0.25%, depending on the office. Ohio has thresholds of 0.5% and 0.25%, depending on the office. Oregon has a threshold of 0.2%. South Carolina has a 1% threshold. Alaska, Montana, South Dakota, Texas, and Vermont automatically recount tied elections. Some states have a recount threshold based on the number of votes rather than the percentage margin; for instance, Michigan has automatic recounts for statewide contests with margins below 2,000 votes.

To understand how automatic recounts affect audit workloads, we estimate the number of ballots to inspect to audit all contests regardless of their margins, and all contests with reported margins greater than 0.1%, 0.5%, and 1%.

7.4.2 November 2020

In the November 2020 general election in Orange County, a total of 1,546,570 ballots (and 3,094,308 ballot cards) were cast in 181 contests. One contest was reported to be a tie, a margin of 0 votes: in the contest for Brea Olinda Unified School District Governing Board Member, Trustee Area 5, both Lauren Barnes and Gail Lyons were reported to receive $1,805$ votes. Because the reported result was a tie, auditing this contest requires a full hand count. If there were no way to identify which cards contain this contest without manually inspecting the cards, a full hand count of that single contest would entail manually inspecting all 3,094,308 cards cast in the election. In all, 27 contests have diluted margins so small (with respect to all cards cast) that auditing each of them would require examining more than 99% of the cast cards, unless CSD are used.

But with CSD, auditing a contest never requires inspecting more cards than contain that contest. This reduces the workload substantially: the estimated workload to audit all 181 contests to a risk limit of 5% is only 20,112 cards in all, a reduction of more than 99%. Without the contest with margin less than 0.1%, the estimated sample size drops to 15,964 cards. Without the four contests with margins less than 0.5%, the estimated sample size drops to 9,228 cards. Without the five contests with margins less than 1%, the estimated sample size further drops to 7,827 cards. Table 7.2 lists the contests with margins under 1%, along with their margins and estimated sample size for that contest for a 5% risk limit RLA using CSD.

Figure 7.2 shows the proportion of ballot cards containing each contest that we would expect the RLA to inspect, versus the number of cards the contest appears on (both on log scale). In general, the sampling fraction decreases as the number of cards the contest appears on increases.

[10] `https://ballotpedia.org/Election_recount_laws_and_procedures_in_the_50_states`, last visited 2 July 2023. Washington's automatic recounts are machine recounts, not hand recounts, so they do not obviate the need for an RLA.

Contest	Cards Cast	Diluted Margin	Sample Size
Brea Olinda Unified School District Governing Board Member, Trustee Area 5	4,164	0	4,164
City of Irvine, City Council	129,948	0.2%	3,930
City of Lake Forest Member, City Council, District 1	10,042	0.2%	2,843
Proposition 17	1,546,210	0.4%	1,488
City of Laguna Beach Member, City Council	16,661	0.8%	746
South Coast Water District Director	22,046	0.9%	696
Member of the State Assembly 74th District	277,516	0.9%	652

Table 7.2: Contests with margins under 1% in the General Election in Orange County, CA, November 2020, number of cards cast, reported diluted margin, and estimated sample size to audit each of them to a risk limit of 5%, on the assumption that the CVRs have no errors.

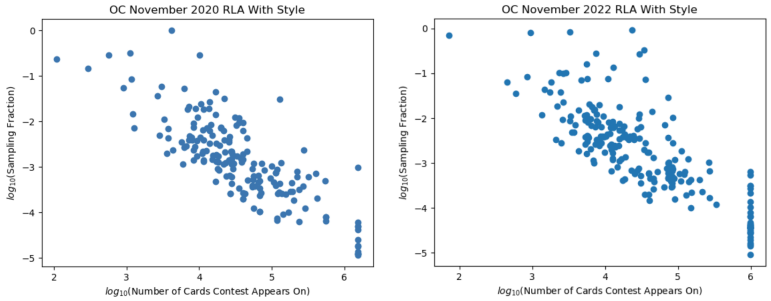

Figure 7.2: Log of the sampling fraction (cards in the sample that contain the contest, divided by cards that contain the contest) versus the log number of cards the contest appears on, for a 5% risk limit RLA using CSD information, for General Elections in Orange County, CA, USA, in November 2020 (left panel, 181 contests) and 2022 (right panel, 214 contests). In general, for a given margin, larger contests with correct outcomes can be confirmed by examining a smaller fraction of the cards that contain the contest. The vertical set of points at the right edge of the plots are county-wide and statewide contests, which appear on the maximum possible number of cards. In 2020, all but one had a sampling fraction less than 1 in 10,000; the smallest was less than 1 in 100,000. In 2022, sampling fractions for the largest contests ranged from 1 in 100,000 to about 1 in 1,000.

7.4.3 November 2022

In the November 2022 general election in Orange County, 994,227 ballots (comprising 1,989,416 cards) were cast in 214 contests. Several contests had small margins. For instance, in the vote-for-three Fountain Valley School District, Governing Board Member contest, the margin between the winner with the fewest votes, Phu Nguyen, and the loser with the most votes, Megan Irvine, was 0.02%. The City of Villa Park, City Council Member contest was also multiwinner plurality with three winners; the margin between the winner with the fewest votes (Jordan Wu) and the loser with the most votes (Donna Buxton) was 0.09%. The margin for Measure K in Costa Mesa, a simple majority contest, was 0.06%.

The estimated sample size to audit all 214 contests to a risk limit of 5% without using CSD is 1,989,415 ballot cards—essentially every card. Indeed, there are 33 contests which, if each had been the *only* contest audited, would have required inspecting more than 99% of all cast ballot cards if CSD were not used to target the sample.

Using CSD reduces the estimated workload by 97%: the estimated sample size to audit all 214 contests to a risk limit of 5% is 62,251 ballot cards, about 3.1% of the cards cast. As mentioned above, state laws for automatic recounts typically have threshold margins of 1%, 0.5%, 0.25%, or 0.1%. Table 7.3 lists the contests with margins of 1% or less, their sizes, margins, and estimated sample size for a CSD RLA of each, at 5% risk limit, computed on the assumption that the CVRs are accurate. The right panel of Figure 7.2 plots sample sizes versus contest sizes for the 214 contests.

Contest	**Cards Cast**	**Diluted Margin**	**Sample Size**
Fountain Valley Sch Dist Governing Board Member	23,512	0.03%	21,772
K-City of Costa Mesa	34,626	0.06%	11,354
City of San Clemente Member, City Council	29,670	0.08%	7,999
City of Villa Park Member, City Council	3,260	0.1%	2,715
Ocean View Sch Dist Governing Board Member	35,990	0.2%	2,634
Orange Unif Sch Dist Governing Board Member, Trustee Area 4	73,665	0.3%	2,088
City of Westminster Member, City Council, District 1	7,467	0.3%	2,064
La Habra City Sch Dist Governing Board Member	12,915	0.3%	1,738
City of Los Alamitos Member, City Council, District 5	946	0.4%	750
Rossmoor Community Services District Director	5,540	0.6%	897
Member of the State Assembly 71st District	85,911	0.7%	873
City of Anaheim Member, City Council, District 2	10,997	0.7%	835
United States Senator Full Term	994,227	0.97%	626
City of Orange Mayor	43,813	0.99%	612

Table 7.3: Contests with margins under 1% in the General Election in Orange County, CA, November 2022, number of cards cast, reported diluted margin, and estimated sample size to audit each of them to a risk limit of 5%, on the assumption that the CVRs have no errors. Unif Sch Dist = Unified School District. The sample size does not decrease monotonically as the margin grows because the sample is drawn without replacement: the sampling fraction matters, too.

Without the three contests with margins below 0.1%, the estimated sample size would be 22,110 cards. Without the nine contests with margins below 0.5%, the estimated sample size would be 11,053 cards. Without the 14 contests with margins less than 1%, the estimated sample size would be 8,123 cards.

7.5 Discussion

It is prudent to give every contest outcome some audit scrutiny: auditing some contests has little bearing on whether the outcomes of other contests are correct. But conducting an RLA of a large number of partially overlapping contests with a wide range of sizes has been thought to be logistically infeasible. By using CSD, the method of Glazer et al. [2021] makes it practical to audit every contest in an election, which we illustrate using data from the 2020 and 2022 general elections in Orange County, California, the fifth largest election jurisdiction in the U.S., with more voters than 24 entire U.S. states. (With previous methods, auditing every contest in an election is generally more challenging in larger jurisdictions than in smaller ones, because larger jurisdictions have more contests and because the small contests are on a smaller fraction of the cards cast in the jurisdiction.)

CSD sampling would reduce the workload of a 5% risk limit RLA by more than 99% for the 2020 election and by 97% for the 2020 election compared to previous approaches. These estimates treat every contest in both elections as if they are entirely contained in Orange County. While this is not true for statewide contests, the estimates still give an idea of the workload to audit many overlapping contests simultaneously. These sample size estimates also assume that a card-level comparison audit could be conducted using all validly cast cards in Orange County. In reality, card-level comparison audits of cards cast in vote centers and polling places might require additional work, e.g., re-scanning cards centrally to create CVRs that are uniquely associated with individual cards. Non(c)esuch Stark [2023a] could be used to avoid such re-scanning, but would require changes to the voting equipment to imprint nonces on cards as they are scanned. CSD-based sampling can also be used with ONEAudit Stark [2023 (in press] without re-scanning or changing the voting system, albeit with some increase in sample size. Future work will investigate the magnitude of that increase.

California law requires auditing the votes in approximately 1% of precincts. In 2020, Orange County's statutory audit tabulated votes on 53,163 cards[11] and in 2022, it tabulated votes on 51,346 cards.[12] While it is easier to count the votes on all the cards in a precinct than to count the votes on the same number of cards selected at random, (i) the statutory 1% audit does not provide evidence that outcomes are correct, (ii) a CSD 5% risk-limit RLA would have involved examining fewer ballots in all in 2020, (iii) the CSD RLA generally involves transcribing data from fewer contests per audited card, and (iv) hand-counting teams generally comprise four people to tabulate votes on a single card, while comparison-audit teams generally comprise only two people.

CSD makes the recommendation of the 2018 National Academies report National Academies of Sciences, Engineering, and Medicine [2018b] practical: jurisdictions can efficiently audit every federal and state election contest as well as all local contests using samples that will generally comprise only a modest fraction of cards cast when reported outcomes are correct. An open-source Python implementation of the method is available at `https://github.com/pbstark/SHANGRLA/tree/main/shangrla`.

[11] `https://elections.cdn.sos.ca.gov/manual-tally/2020-general/orange.pdf` last visited 8 September 2023.

[12] `https://elections.cdn.sos.ca.gov/manual-tally/2022-general/orange.pdf`, last visited 8 September 2023.

Chapter 8

Robust inference for matching under rolling enrollment[1]

8.1 Introduction

Matching methods attempt to estimate average causal effects by grouping each treated unit with one or more otherwise similar controls and using paired individuals to approximate the missing potential outcomes. Assuming that paired individuals are sufficiently similar on observed attributes and that no important unobserved attributes confound the comparison, the difference in outcomes approximates the impact of treatment for individuals in the pair [Stuart, 2010]. Despite matching's transparency and intuitive appeal, it faces complications in datasets containing repeated measures for the same individuals over time. When only a single time of treatment is present, the primary challenge is deciding how to construct matching distances from pre-treatment repeated measures and assess outcomes using post-treatment repeated measures [Haviland et al., 2008]. The situation is more complex under rolling enrollment, or staggered adoption, when individuals opt into treatment at different times [Ben-Michael et al., 2021]. Several authors [Li et al., 2001, Lu, 2005, Witman et al., 2019, Imai et al., 2020] proceed by matching each treated unit to the version of the control unit present in the data at the time of treatment. For example, in Imai et al. [2020]'s reanalysis of data from Acemoglu et al. [2019] on the impact of democratization on economic growth, countries undergoing democratizing political reforms are matched to similar control countries not undergoing such reforms in the same year.

Although this method is logical whenever strong time trends are present, in other cases it may overemphasize similarity on time at the expense of other variables. Bohl et al. [2010] study the impact of serious falls on subsequent healthcare expenditures for elderly adults using patient data from a large healthcare system. While patients who fall could be matched to patients who appear similar based on recent health history on the calendar date of the fall, the degree of similarity in health histories is likely much more important than the similarity of the exact date at which each patient is measured. Following this idea, the GroupMatch algorithm [Pimentel et al., 2020] constructs matches optimally across time, prioritizing matching on important covariates over ensuring that units are compared at the

[1]This chapter comprises a publication [Glazer and Pimentel, 2023] co-authored by Samuel D. Pimentel.

same point in time.

Another example where rolling enrollment arises is in major league baseball (MLB). Quantifying the impact of injury on player performance in professional sports is important for both managers and players themselves. Increasingly, players are valued and compensated in a manner driven by quantitative metrics of past performance, but injuries have potential to disrupt the continuity between past and future performance [Begly et al., 2018, Conte et al., 2016, Frangiamore et al., 2018, Wasserman et al., 2015]. One way to quantify impact in this setting is as the difference between the value of a performance metric the player would have achieved in the absence of injury and the value of the same metric achieved after a given injury. This quantity can be estimated using matching. However, players do not all get injured at the same time, so GroupMatch is a natural fit here. It allows us to match injured and non-injured players flexibly across time, because we likely do not care whether a player injured on June 1 is matched to a non-injured player on exactly the same day or a couple weeks earlier, e.g., May 15, so long as those players are sufficiently similar on other covariates such as recent performance.

Several challenges remain outstanding for matching methods under rolling enrollment. GroupMatch's flexible approach relies heavily on a strong assumption that time itself is not a confounder, and discussion of checking this assumption has been minimal so far. Even when flexible matching is warranted, the presence of multiple copies of the same control individual necessitates a constraint to ensure that a treated unit is not simply paired to multiple slightly different copies of the same control; several choices of this constraint exist permitting varying degrees of flexibility, and users must choose among them. Most importantly, for both GroupMatch and methods that match exactly in time there is substantial ambiguity about how to conduct valid inference. When multiple copies of a control individual are forbidden from appearing in the matched design, randomization inference may be used [Lu, 2005, Pimentel et al., 2020] but no strong guarantees exist outside this special case.

In what follows, we present several innovations that greatly enhance the toolkit for matching and treatment effect evaluation under rolling enrollment. First, we introduce a new matched design called GroupMatch with instance replacement, which has computational, analytical, and statistical advantages over existing designs in many common settings. Second, we give a comprehensive characterization of a new block-bootstrap-based method for inference that applies broadly across existing methods for matching under rolling enrollment, including our new design. The block-bootstrap approach was originally suggested by Imai et al. [2020] and is based on related work in the cross-sectional case by Otsu and Rai [2017], but until now has not carried any formal guarantee. Finally, we introduce a falsification test to partially check the assumption of timepoint agnosticism underpinning GroupMatch's validity, empowering investigators to extract evidence from the data about this key assumption prior to matching. We prove the validity of our bootstrap method under the most relevant set of constraints on reuse of controls, and we demonstrate the effectiveness of both the placebo test and the bootstrap inference approach through simulations and an analysis of injury data in major league baseball. In particular, the bootstrap method shows improved performance over linear-regression-based approaches to inference often applied in similar settings, while making much weaker assumptions.

This chapter is organized as follows. Section 8.2 presents the basic statistical framework and reviews the GroupMatch framework, inference approaches for matching designs, and

other related literature. In Section 8.3 we introduce a new constraint for use of controls in GroupMatch designs, leading to a new design called GroupMatch with instance replacement. Section 8.4 presents a block bootstrap inference approach for matching under rolling enrollment, and Section 8.5 evaluates it via simulation. In Section 8.6 we present a falsification test for the assumption that time is not a confounder. In Section 8.7 we apply our methods to evaluate whether minor injuries impact short-term MLB performance. Section 8.8 concludes.

8.2 Statistical framework

8.2.1 Setting and notation

We observe n subjects. For each subject i in the study, we observe repeated measures $(Y_{i,t}, \mathbf{X}_{i,t})$ for timepoints $t = 1, \ldots, T$, where $Y_{i,t}$ is an outcome of interestand $\mathbf{X}_{i,t}$ is a vector of covariates. We also observe a time of treatment initiation T_i for each subject, with $T_i \in \{1, \ldots, T\}$ for subjects who receive treatment at some point and $T_i = \infty$ for those who remain controls at all observed timepoints. We specify a burn-in period of length $L - 1$ during which no individuals are treated, i.e. $T_i \geq L$ for all i (or allow treatment at $t = 1$ by setting $L = 1$). We denote the collection of repeated measures for each subject i, along with T_i, as the trajectory O_i.

For clarity, we focus on "instantaneous" effects of treatment, with outcomes measured immediately following treatment at the same time when treatment is first initiated. Let $Y_{i,t}(0)$ be the potential outcome for unit i that would have been observed at time t if $T_i > t$, and let $Y_{i,t}(1)$ be the potential outcome that would have been observed if $T_i = t$. The finite sample average effect of treatment on the treated (ATT) is denoted by Δ:

$$\begin{aligned}\Delta &= \frac{1}{N_1} \sum_{i=1}^{N} \sum_{t=1}^{T} 1\{t = T_i\}[Y_{i,t}(1) - Y_{i,t}(0)] \\ &= \frac{1}{N_1} \sum_{i=1}^{N} D_i \left[Y_{i,t=T_i}(1) - Y_{i,t=T_i}(0)\right]\end{aligned}$$

Here the D_i variable is introduced as a convenient shorthand to indicate whether $T_i < \infty$. We assume that trajectories O_i are sampled independently from some infinite population, although we do not assume independence of observations within the same trajectory. Defining expectation $E(\cdot)$ with respect to sampling from this population, we define the population ATT as $\Delta_{pop} = E(\Delta)$. For future convenience, we also introduce a concise notation for conditional expectation (again, over the sampling distribution) of potential outcomes given no treatment through time t and the covariates observed in the previous L timepoints:

$$\mu_0^t(\mathbf{X}) = E[Y_{i,t}(0) | \{X_{i,t'}\}_{t'=t-L+1}^{t'=t} = \mathbf{X}, T_i > t]$$

Throughout, we abuse notation slightly by writing $\mu_0(\mathbf{X}_{i,t})$ to indicate conditional expectation given the L lagged values of $\mathbf{X}_i$ directly preceding time t, inclusive.

The potential outcomes framework adopted here represents one of many possible framings for studies with rolling enrollment. Pimentel et al. [2020] define potential outcomes as

functions of the length of time since treatment initiation, while both Ben-Michael et al. [2021] and Athey and Imbens [2022] define them as functions of the time of treatment initiation for the subject in question. In principle these alternate constructions are much richer than ours, allowing for much more general and complicated patterns of effects, but in practice all these authors use simplifying assumptions or focus on estimands that reduce attention to at most two potential outcomes of interest for each individual at each timepoint. For example, both Pimentel et al. [2020] and Ben-Michael et al. [2021] allow for treatment effects to be measured at some follow-up time postdating the time of treatment rather than focusing on instantaneous effects, but since the length of follow-up is fixed in advance only two potential outcomes ever need to be considered for each unit. Similarly, the "no anticipation" assumption of Ben-Michael et al. [2021] and Athey and Imbens [2022] ensures that there is exactly one potential outcome of interest associated with the control condition for each individual, and the "invariance to history" assumption of Athey and Imbens [2022] collapses distinctions among potential outcomes under treatment. As such, the results we present below extend easily to all the potential outcomes frameworks just described, making the appropriate substitutions for our $Y_{i,t}(1)$ and $Y_{i,t}(0)$. While the potential outcomes framework of Imai et al. [2020] is much more general than all the previously-mentioned works in allowing subjects to revert from treatment back to control, our framing also extends easily to it in the special case when no exit from treatment is allowed.

8.2.2 Identification assumptions

Pimentel et al. [2020] studied the following difference-in-means estimator in designs where each treated unit is matched to C control observations. $M_{it,jt'}$ is an indicator for whether subject i at time t has been matched to subject j at time t':

$$\hat{\Delta} = \frac{1}{N_1}\sum_{i=1}^{n} D_i[Y_{i,t=T_i} - \frac{1}{C}\sum_{j=1}^{N}\sum_{t'=1}^{T} M_{iT_i,jt'}Y_{j,t'}]$$

Pimentel et al. [2020] show that this estimator is unbiased for the population ATT under the following conditions:

1. Exact matching: matched units share identical values for covariates in the L timepoints preceding treatment.

2. L-ignorability: conditional on the covariate history over the previous L timepoints and the absence of treatment prior to baseline, an individual's potential outcome at a given time is independent of the individual's overall treatment status. Formally,

$$\{T_i < \infty\} \perp\!\!\!\perp Y_{i,t}(0) | T_i > t-1, \{X_{i,s}\}_{s=t-L+1}^{t}, \forall i.$$

   Intuitively, this assumption prevents unobserved confounding that makes potential outcomes for treated subjects systematically different from those that remain controls even after accounting for information from a baseline period.

3. Timepoint agnosticism: mean potential outcomes under control do not differ for any instances with identical covariate histories at different timepoints. Formally, for any set of L covariate values $\mathbf{X}$,

$$\mu_0^t(\mathbf{X}) = \mu_0^{t'}(\mathbf{X}) = \mu_0(\mathbf{X}) \text{ for any } 1 \leq t, t' \leq T.$$

   This assumption ensures that matching across time is reasonable by ruling out time trends other than those captured by time-varying covariates. For clarity we drop the t superscript when discussing the conditional expectation $\mu_0(\mathbf{X})$ in what follows, with the exception of Section 8.6 where we temporarily consider failures of this assumption.

4. Covariate L-exogeneity: future covariates do not encode information about the potential outcome at time t given covariates and absence of treatment over the previous L timepoints. Formally,

$$(X_{i,1}, ..., X_{i,T}) \perp\!\!\!\perp Y_{i,t}(0) | T_i > t-1, \{X_{i,s}\}_{s=t-L+1}^{t}), \forall i.$$

   Like time agnosticism, covariate L-exogeneity is important to justify considering past and future instances from a control trajectory as part of the matching procedure. If future instances' covariates include or are correlated with past instances' outcomes, then we may indirectly match on study outcomes introducing bias into our estimation step [Rosenbaum, 1984, Hansen, 2008]. Covariate L-exogeneity ensures that future covariates are safe to consider during the design stage.

5. Overlap: given that a unit is not yet treated at time $t-1 \geq L$, the probability of entering treatment at the next time point is neither 0 nor 1 for any choice of covariates over the L timepoints at and preceding t.

$$0 < P(T_i = t \mid T_i > t-1, X_i^t, \ldots, X_i^{t-L+1}) < 1 \qquad \forall t > L$$

   While not stated explicitly in Pimentel et al. [2020], we note that the authors rely on an overlap assumption of this type in the proof of their main result.

The exact matching assumption is no longer needed for asymptotic identification of the population ATT if we modify the estimator by adding in a bias correction term. As in Otsu and Rai [2017] and Abadie and Imbens [2011], we first estimate the conditional mean function $\mu_0(\mathbf{X})$ of the potential outcomes and use this outcome regression to adjust each matched pair for residual differences in covariates not addressed by matching. As outlined in Abadie and Imbens [2011], bias correction leads to asymptotic consistency under regularity conditions on the potential outcome mean estimator $\widehat{\mu}(\cdot)$ (for further discussion of regularity assumptions on $\widehat{\mu}_0(\cdot)$ see the proof of Theorem 2 in Appendix B.1). Many authors have also documented benefits from adjusting matched designs using outcome models [Rubin, 1979, Antonelli et al., 2018]. The specific form of our bias-corrected (i.e., model-adjusted) estimator is as follows:

$$\hat{\Delta}_{adj} = \frac{1}{N_1} \sum_{i=1}^{n} D_i[(Y_{i,t=T_i} - \hat{\mu}_0(\mathbf{X}_{i,T_i})) - \frac{1}{C} \sum_{j=1}^{N} \sum_{t'=1}^{T} M_{iT_i,jt'}(Y_{j,t'} - \hat{\mu}_0(\mathbf{X}_{i,t'}))]$$

Datasets with many variables, especially continuous variables or variables with many categories, ensure that exact matching is rarely possible in practice, and in light of this we focus primarily on estimator $\widehat{\Delta}_{adj}$ in what follows.

As discussed by Imai et al. [2020] for settings without rolling enrollment, identification is possible under weaker assumptions if a difference-in-differences estimator is used instead of the difference-in-means. While we focus primarily on the simpler bias-corrected difference-in-means estimator for clarity of exposition, the difference-in-differences approach also offers advantages for our setting, and the new matching and inference strategies we propose extend naturally to such estimators. We provide further discussion in Section 8.4.3.

8.3 GroupMatch with instance replacement

Before discussing our method for inference in general matched designs under rolling enrollment, we introduce a new type of GroupMatch design. Pimentel et al. [2020] described two different designs produced by GroupMatch denoted Problems A and B, designs we refer to as GroupMatch without replacement and GroupMatch with trajectory replacement respectively.

1. **GroupMatch without replacement**: each control unit can be matched to at most one treated unit. If a treated unit is matched to an instance of a control unit, no other treated unit can match to (any instance) of that control unit.

2. **GroupMatch with trajectory replacement**: each control *instance* can be matched to at most one treated unit. Each treated unit can match to no more than one instance from the same control trajectory. However, different treated units can match to different instances of the same control trajectory, so a single control trajectory can contribute multiple distinct instances to the design.

As our chosen names for these designs suggest, their relative costs and benefits reflect the choice between matching without and with replacement in cross-sectional settings. As discussed by Hansen [2004], matching without replacement (in which each control may appear in at most one matched set), leads to less similar matches compared to matching with replacement (in which controls can reappear in many matched sets) since in cases where two treated units both share the same nearest control only one can use it. On the other hand, matching without replacement frequently leads to estimators with lower variance than those from matching with replacement, where an individual control unit may appear in many matched sets, making the estimator more sensitive to random fluctuations in its response. Thus, one aspect of choosing between these designs is a choice about how to strike a bias-variance tradeoff. The other important aspect distinguishing these designs is that randomization inference, which is based on permuting treatment assignments in each matched set independently of others, generally requires matching without replacement. Specifically, when multiple controls may be matched to each treated unit and replacement is allowed, the resulting configuration of treated and control units no longer resembles the design of a blocked or matched experiment.

These same dynamics play out in comparing GroupMatch without replacement and GroupMatch with trajectory replacement. GroupMatch without replacement ensures that responses in distinct matched sets are statistically independent (under a model in which trajectories are sampled independently), allowing for randomization inference, and ensures that the total weight on observations from any one control trajectory can sum only to $1/C$, ensuring that the estimator's variance cannot be too highly inflated by a single trajectory with large weight. The resulting data configuration also resembles what might be obtained in a sequential experimental design employing matching-on-the-fly as discussed by Kapelner and Krieger [2014] and Pimentel et al. [2020], and inference may be conducted using the associated randomization distribution. On the other hand, GroupMatch with trajectory replacement leads to higher-quality matches and reduced bias in matched pairs, although overlap among the matched sets formed makes it difficult to envision a corresponding "target trial."

We suggest a third GroupMatch design which leans even further towards expanding the potential control pool and reducing bias.

3. **GroupMatch with instance replacement**: Each treated unit can match to no more than one instance from the same control unit, but control instances can be matched to more than one treated unit.

GroupMatch with instance replacement is identical to GroupMatch without trajectory replacement except that it allows repetition of individual instances within the matched design as well as non-identical instances from the same trajectory. As such, it is guaranteed to produce higher-quality matches than GroupMatch without trajectory replacement, but may lead to higher-variance estimators since individual instances may receive weights larger than $1/C$. Figure 8.1 illustrates these three GroupMatch methods with a toy example that matches injured baseball players to non-injured players based on on-base percentage (OBP).

In practice we view GroupMatch with instance replacement as a more attractive approach than GroupMatch with trajectory replacement almost without exception. One reason is that while the true variance of estimators from GroupMatch with instance replacement may often exceed that of estimators from GroupMatch with trajectory replacement by a small amount, our recommended approach for *estimating* the variance and conducting inference are not able to capture this difference. As we describe in Section 8.4, in the absence of a specific parametric model for correlations within a trajectory, inference proceeds in a conservative manner by assuming arbitrarily high correlations within a trajectory (much like the clustered standard error adjustment in linear regression). Since the variance advantage for GroupMatch with trajectory replacement arises only when correlations between instances within a trajectory are lower than one, the estimation strategy is not able to take advantage of them. This disconnect means that GroupMatch with trajectory replacement will not generally lead to narrower empirical confidence intervals, much as variance gains associated with paired randomized trials relative to less-finely-stratified randomized trials may not translate into reduced variance estimates [Imbens, 2011].

A second important advantage of GroupMatch with instance replacement is its computational and analytical tractability relative to the other GroupMatch designs. One way to implement GroupMatch with instance replacement as a network flow optimization problem

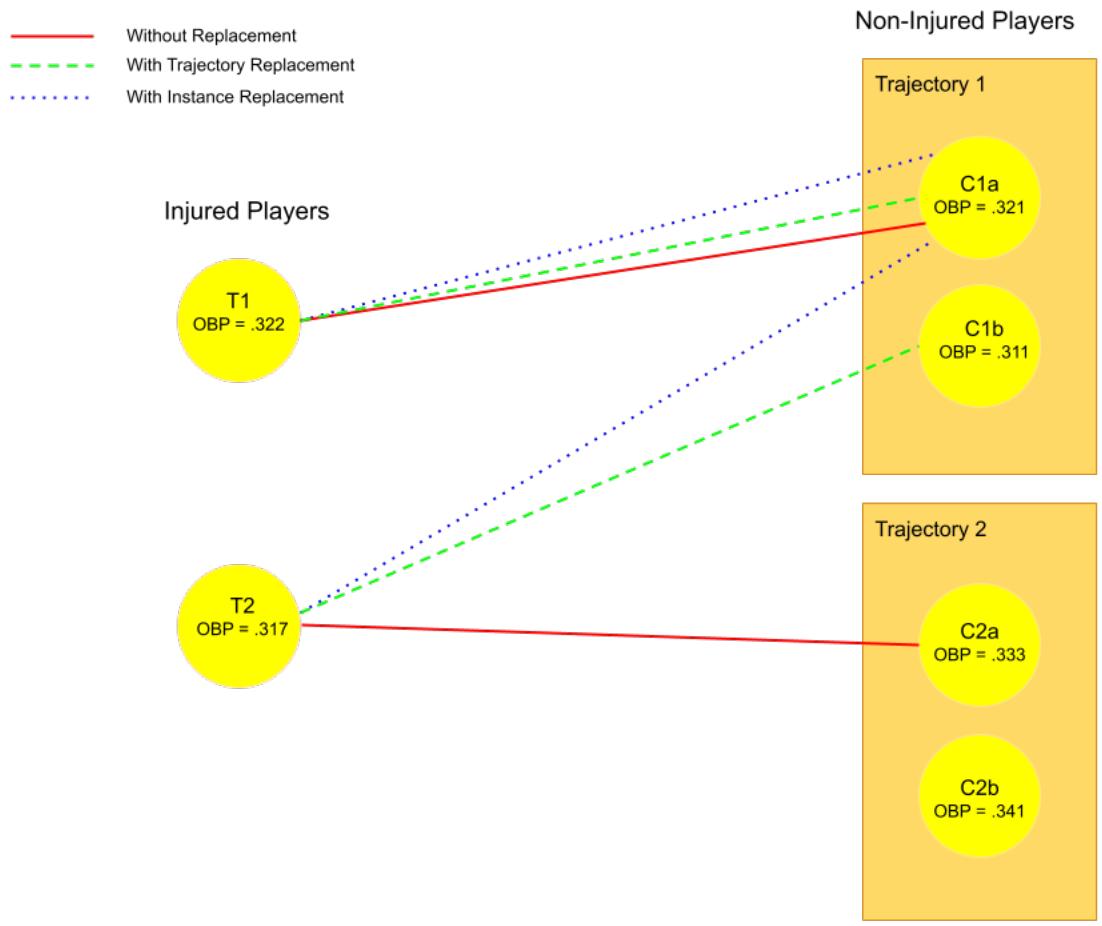

Figure 8.1: Toy example illustrating the three GroupMatch matching methods. Two injured baseball players (T1 and T2) are matched 1-1 to non-injured baseball players (C1a/b and C2a/b) based on player on-base percentage (OBP). Each non-injured player has two pseudo-injury times or instances. Under GroupMatch without replacement, T2 must match to an instance in Trajectory 2 because at most one instance from Trajectory 1 can participate in the match. Under GroupMatch with trajectory replacement, T2 can match to C1b but not to C1a, since multiple control instances can be chosen from the same trajectory as long as they are distinct. Under GroupMatch with instance replacement, both T1 and T2 are able to match to C1a. However, if each treated instance were matched to two control instances instead of one, GroupMatch with instance replacement would still forbid either T1 or T2 to match to a second instance in Trajectory 1.

is to remove a set of constraints in Pimentel et al. [2020]'s Network B (specifically the upper capacity on the directed edges connected to the sink node), and in Sections 8.5 and 8.7 we use this implementation for its convenient leveraging of the existing `groupmatch` package in R. However, much more computationally efficient algorithms are also possible. Crucially, the removal of the constraint forbidding instance replacement means that matches can be calculated for each treated instance without reference to the choices made for other treated units; the C best matches for a given treated unit are simply the C nearest neighbor instances such that no two such control instances within the matched set come from the same trajectory. In principle, this allows for complete parallelization of the matching routine. On the analytical side, this aspect of the design makes it possible to characterize the matching algorithm as a

generalized form of nearest neighbor matching, a strategy we adopt in the proof of Theorem 2 to leverage proof techniques used by Abadie and Imbens [2006] for cross-sectional nearest neighbor matching. In light of these considerations, we focus primarily on GroupMatch with instance replacement in what follows, although the methods derived appear to perform well empirically for other GroupMatch designs too.

8.4 Block Bootstrap Inference

8.4.1 Inference methods for matched designs

Broadly speaking, there are two schools of thought in conducting inference for matched designs. One approach, spearheaded by Abadie and Imbens [2006, 2008, 2011, 2012], views the raw data as samples from an infinite population and demonstrates that estimators based on matched designs (which in this framework are considered to be random variables, as functions of random data) are asymptotically normal. Inferences are based on the asymptotic distributions of matched estimators. A second approach, described in detail in Rosenbaum [2002a,b] and Fogarty [2020], adopts the perspective of randomization inference in controlled experiments. Conditional on the structure of the match and the potential outcomes, the null distribution of a test statistic over all possible values of the treatment vector is obtained by permuting values of treatment within matched sets. When matches are exact and unobserved confounding is absent, strong finite sample guarantees hold for testing sharp null hypotheses without further assumptions on outcome variables. Asymptotic guarantees for weak null hypotheses may be obtained too, assuming a sequence of successively larger finite populations [Li and Ding, 2017]. Well-developed methods of sensitivity analysis are also available.

As described in Pimentel et al. [2020], while standard methods of inference may be applied to GroupMatch without replacement, in which control individuals contribute at most one unit to any part of the match, none have been adequately developed for GroupMatch with trajectory replacement, in which distinct matched sets may contain different versions of the same control individual. For randomization inference, the barrier appears to be quite fundamental, because permutations of treatment within one matched set can no longer be considered independently for different matched sets. In GroupMatch with trajectory replacement, a treated unit receives treatment at one time and appears in a match only once; if treatment is permuted among members of a matched set so that a former control now attains treatment status, what is to be done about other versions of this control unit that are present in distinct matched sets? We note that similar issues arise when contemplating randomization inference for general cross-sectional matching designs with replacement, and we are aware of no solutions for randomization inference even in this simpler case.

In contrast, the primary issue in applying sampling-based inference to GroupMatch designs with trajectory replacement is the unknown correlation structure for repeated measures from a single control individual. The literature on matching with replacement provides estimators for pairs that are fully independent [Abadie and Imbens, 2012] and for cases in which a single observation appears identically in multiple pairs [Abadie and Imbens, 2006], but not for the intermediate case of GroupMatch with trajectory replacement where distinct but correlated observations appear in distinct matched sets. These issues extend beyond

the GroupMatch family to any matched design under rolling enrollment in which control trajectories contribute to multiple matched sets, including those of Witman et al. [2019] and Imai et al. [2020].

In what follows we give formal guarantees for a sampling-based inference method appropriate for general matching designs under rolling enrollment suggested by Imai et al. [2020], which generalizes a recent proposal of Otsu and Rai [2017] for valid sampling-based inference of cross-sectional matched studies using the bootstrap. Although the bootstrap often works well for matched designs without replacement [Austin and Small, 2014], naïve applications of the bootstrap in matched designs with replacement have been shown to produce incorrect inferences as a consequence of the failure of certain regularity conditions [Abadie and Imbens, 2008]. Intuitively, if matching is performed after bootstrapping the original data, multiple copies of a treated unit will necessarily all match to the same control unit, creating a clumping effect not present in the original data. However, Otsu and Rai [2017] arrived at an asymptotically valid bootstrap inference method for matching by bootstrapping weighted and bias-corrected functions of the original observations *after* matching rather than repeatedly matching from scratch in new bootstrap samples. We show that a similar bootstrap approach, applied to entire trajectories of repeated measures in a form of the block bootstrap, provides valid inference for matched designs under rolling enrollment. Note that in our formal results we focus on GroupMatch with instance replacement as the most difficult case, since the designs of Witman et al. [2019] and Imai et al. [2020] may be understood as restricted special cases in which matching on time is exact.

8.4.2 Block Bootstrap

In order to conduct inference under GroupMatch with trajectory or instance replacement we propose a weighted block bootstrap approach. We rearrange the GroupMatch ATT estimator from Section 8.2 as follows, letting $K_M(i,t)$ be the number of times the instance at trajectory i and time t is used as a match.

$$\begin{aligned}\hat{\Delta}_{adj} &= \frac{1}{N_1}\sum_{i=1}^{N} D_i[(Y_{i,T_i} - \hat{\mu}_0(\mathbf{X}_{i,T_i})) - \frac{1}{C}\sum_{j=1}^{N}\sum_{t'=1}^{T} M_{iT_i,jt'}(Y_{j,t'} - \hat{\mu}_0(\mathbf{X}_{i,t'}))] \\ &= \frac{1}{N_1}\sum_{i=1}^{N}\left[D_i\left(Y_{i,T_i} - \hat{\mu}_0(\mathbf{X}_{i,T_i})\right) - (1-D_i)\sum_{t=1}^{T}\frac{K_M(i,t)}{C}(Y_{i,t} - \hat{\mu}_0(\mathbf{X}_{i,t}))\right] = \frac{1}{N_1}\sum_{i=1}^{N}\hat{\Delta}_i.\end{aligned}$$

Because different instances of the same control unit are correlated, we resample the *trajectory*-level quantities $\widehat{\Delta}_i$ rather than the instance-level quantities. Since the $\widehat{\Delta}_i$ are functions of the $K_M(i,t)$ weights in the original match, we do not repeat the matching process within bootstrap samples. In particular, we proceed as follows:

1. Fit an outcome regression $\widehat{\mu}_0(\cdot)$ for outcomes based on covariates in the previous L timepoints using only control trajectories.

2. Match treated instances to control instances using GroupMatch with instance replacement. Calculate matching weights $K_M(i,t)$ equal to the number of times the instance at time t in trajectory i appears in the matched design.

3. Calculate the model-adjusted ATT estimator $\widehat{\Delta}_{adj}$.

4. Repeat B times:

   (a) Randomly sample N elements $\widehat{\Delta}_i$ with replacement from $\{\widehat{\Delta}_1, \ldots, \widehat{\Delta}_N\}$.

   (b) Calculate the bootstrap bias-corrected ATT estimator $\widehat{\Delta}^*_{adj}$ for this sample of trajectories as follows:

$$\widehat{\Delta}^*_{adj} = \frac{1}{N_1} \sum_{i=1}^{N} \widehat{\Delta}^*_i$$

5. Construct a (1 - α) confidence interval based on the $\alpha/2$ and $1 - \alpha/2$ percentile of the $\widehat{\Delta}^*_{adj}$- values calculated from the bootstrap samples.

This method is essentially a block bootstrap, very similar to the method proposed in Imai et al. [2020]. Note that while the recipe above uses the nonparametric bootstrap, it may easily be generalized to other approaches such as the wild bootstrap and the Bayesian bootstrap. In particular, consider rewriting $\widehat{\Delta}^*_{adj}$ in terms of a new set of random variables $W^*_1, \ldots, W^*_N$ that we denote the bootstrap weights:

$$\widehat{\Delta}^*_{adj} = \frac{1}{N_1} \sum_{i=1}^{N} \widehat{\Delta}^*_i = \sum_{i=1}^{N} \frac{W^*_i}{\sqrt{N_1}} \widehat{\Delta}_i \tag{8.1}$$

To recover the nonparametric bootstrap, the bootstrap weights W^*_i are chosen as $Q_i/\sqrt{N_1}$, where Q_i is the number of times subject i is selected when sampling with replacement; if the W^*_i are chosen instead by sampling from a scaled Dirichlet or scaled two-point distribution, we obtain the Bayesian bootstrap and the wild bootstrap respectively (see Otsu and Rai [2017] for specifics). To adapt the step-by-step algorithm for these approaches, we draw W^*_is rather than $\widehat{\Delta}^*$s in step 4(a) and use (8.1) to calculate $\widehat{\Delta}^*_{adj}$ in step 4b.

Our main result below shows the asymptotic validity of this approach. Several assumptions, in addition to Assumptions 2-5 in Section 8.2.2, are needed to prove this result. We summarize these assumptions verbally here, deferring formal mathematical statements to Section B.1 of the appendix. First, we require the covariates X_i to be continuous with compact and convex support and a density both bounded and bounded away from zero. Second, we require that the conditional mean functions are smooth in $\mathbf{X}$, with bounded fourth moments. In addition, we require that conditional variances of the treated potential outcomes and conditional variances of nontrivial linear combinations of control potential outcomes from the same trajectory are smooth and bounded away from zero. We also require that conditional fourth moments of potential outcomes under treatment and linear combinations of potential outcomes under control are uniformly bounded in the support of the covariates. Finally, we make additional assumptions related to the conditional outcome mean estimator $\widehat{\mu}_0(\cdot)$, specifically that the kLth derivative of the true conditional mean functions $\mu^t_1(\cdot)$ and $\mu_0(\cdot)$ exist and have finite suprema, and that the $\widehat{\mu}_(\cdot)$ converges to $\mu_0(\cdot)$ at a sufficiently fast rate. Finally, we require mild regularity conditions on the bootstrap weights W^*_i, easily satisfied by construction in the bootstrap approaches we have mentioned. To state the theorem,

we also define

$$\sqrt{N_1}U^* = \frac{1}{\sqrt{N_1}} \sum_{i=1}^{N} \left(\widehat{\Delta}_i^* - D_i \widehat{\Delta}_{adj} \right) = \sum_{i=1}^{N} W_i^* (\hat{\Delta}_i - D_i \hat{\Delta}_{adj}).$$

Theorem 2. *Under assumptions M, W, and R presented in Section B.1.1 of the appendix,*

$$sup_r | Pr\{\sqrt{N_1}U^* \leq r | (\mathbf{Y}, \mathbf{D}, \mathbf{X})\} - Pr\{\sqrt{N_1}(\hat{\Delta}_{adj} - \Delta) \leq r\} | \xrightarrow{p} 0$$

as $N \to \infty$ with fixed control:treated ratio C.

Our regularity assumptions on the data-generating process and the regression estimator $\widehat{\mu}_0(\cdot)$ are modeled closely on those of Abadie and Imbens [2006] and later Otsu and Rai [2017], and our proof technique is very similar to arguments in Otsu and Rai [2017]. Briefly, U^* is decomposed into three terms which correspond to deviations of the potential outcome variables around their conditional means, approximation errors for $\widehat{\mu}_0(\mathbf{X})$ terms as estimates of $\mu_0(\mathbf{X})$ terms, and deviations of conditional average treatment effects $\mu_1^t(\mathbf{X}) - \mu_0(\mathbf{X})$ around the population ATT Δ. Regularity conditions on the data-generating process ensure that the conditional average treatment effects converge quickly to the population ATT. Regularity assumptions on the regression estimator, combined with bounds on the largest nearest-neighbor discrepancies in $\mathbf{X}$ vectors due originally to Abadie and Imbens [2006] and adapted to the GroupMatch with instance replacement design, show that the deviation between $\widehat{\mu}_0(\cdot)$ and $\mu_0(\cdot)$ disappears at a fast rate. Finally, a central limit theorem applies to the deviations of the potential outcomes. For details, see Section B.1 of the appendix.

8.4.3 Difference-in-Differences Estimator

While we have focused so far on the difference-in-means estimator, Imai et al. [2020] recommend a difference-in-differences estimator for matched designs with rolling enrollment in the context of designs that match exactly on time. This estimator can be used under rolling enrollment as well, taking the following form under bias correction:

$$\begin{aligned}
\hat{\Delta}_{DiD} = \frac{1}{N_1} \sum_{i=1}^{N} D_i [& ((Y_{i,T_i} - \hat{\mu}_0(\mathbf{X}_{i,T_i})) - (Y_{i,T_i-1} - \hat{\mu}_0(\mathbf{X}_{i,T_i-1}))) - \\
& \frac{1}{C} \sum_{j=1}^{N} \sum_{t'=1}^{T} M_{iT_i,jt'} ((Y_{j,t'} - \hat{\mu}_0(\mathbf{X}_{i,t'})) - (Y_{j,t'-1} - \hat{\mu}_0(\mathbf{X}_{i,t'-1})))] \\
= \frac{1}{N_1} \sum_{i=1}^{N} D_i [& ((Y_{i,t=T_i} - \hat{\mu}_0(\mathbf{X}_{i,T_i})) - (Y_{i,t=T_i-1} - \hat{\mu}_0(\mathbf{X}_{i,T_i-1}))) - \\
& (1 - D_i) \sum_{t=1}^{T} \frac{K_M(i,t)}{C} ((Y_{i,t} - \hat{\mu}_0(\mathbf{X}_{i,t})) - (Y_{i,t-1} - \hat{\mu}_0(\mathbf{X}_{i,t-1})))] = \frac{1}{N_1} \sum_{i=1}^{N} \hat{\Delta}_i^{DiD}
\end{aligned}$$

This estimator requires L lags to be measured at time $T_i - 1$, so a burn-in period of length L rather than $L - 1$ is needed.

A key advantage of this bias-corrected differences-in-differences estimator is that it relies on different identification assumptions than the bias-corrected difference-in-means: essentially, any assumption previously made on the potential outcome $Y_{i,t}(0)$ must now hold instead only for the post-pre potential outcome difference $Y_{i,t}(0) - Y_{i,t-1}(0)$. The resulting assumptions tend to be substantively weaker. In particular, Imai et al. [2020] highlight how the L-ignorability assumption can be replaced by a parallel trends assumption that requires only that post-pre differences in potential outcomes be conditionally independent of treatment, allowing for different unobserved outcome intercepts for different individuals. The time-agnosticism assumption also becomes weaker when formulated for outcome differences, allowing for a constant linear trend in potential outcome means rather than requiring them to be invariant to time conditional on covariates.

We can easily adapt the results of the previous section to show that the block bootstrap gives valid inference for the difference-in-difference estimator when these identification assumptions hold. The inference procedure simply requires bootstrapping the $\hat{\Delta}_i^{DiD}$ terms in place of the $\hat{\Delta}_i$s defined above. While the regularity conditions given for Theorem 2 suffice for the new estimator, the proof (as presented in Section B.1 of the appendix) requires mild modification to work for this difference-in-differences estimator. In particular, the variance estimators include additional covariance terms. For more details, see Section B.1 of the appendix.

8.5 Simulations

We now explore the performance of weighted block bootstrap inference via simulation. In particular, we investigate coverage and length of confidence intervals compared to those obtained by conducting parametric inference for weighted least squares estimators with and without cluster-robust error adjustment for controls from the same trajectory.

8.5.1 Data Generation

We generate eight covariates, four of them uniform across time for each individual i, (i.e., they take on the same value at every timepoint):

$$\begin{pmatrix} X_{3,i} \\ X_{4,i} \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.7 \\ 0.7 & 01 \end{pmatrix} \right]$$

$X_{2,i} \sim N(0,1)$ for control units and $X_{2,i} \sim N(0.25,1)$ for treated units

X_1 is also uniform across time, but it is correlated with a time-varying covariate, X_5, so we will introduce it below. The correlations between covariates (X_3 and X_4, and X_1 and X_5) are calibrated to the correlations observed between covariates in the baseball example in Section 8.7 (i.e., height and weight have a correlation of approximately 0.7, and lag OBP and age have a correlation of approximately -0.4).

For treated units:

$$\begin{pmatrix} X_{1,i} \\ X_{5,i} \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & -0.4 \\ -0.4 & 01 \end{pmatrix}\right]$$
$$X_7, X_8 \sim N(0,1) \text{ and } X_6 \sim N(0.5,1)$$

Four of the covariates are time-varying for control units. For each control unit, three instances are generated from a random walk process to correlate their values across time. Formally, for instance t in trajectory i, covariate j is generated as follows:

$$\begin{pmatrix} X_{1,i} \\ X_{5,i,1} \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & -0.4 \\ -0.4 & 01 \end{pmatrix}\right]$$
$$X_{j,i,1} \sim N(0,1) \text{ for } j = 6,7,8$$
$$X_{j,i,t} = X_{j,i,(t-1)} + \epsilon_{j,i,(t-1)} \text{ for } t = 2,3$$
$$\epsilon_{j,i,1}, \epsilon_{j,i,2} \sim N(0, 0.5^2)$$

Fixing $a_L = log(1.25)$, $a_M = log(2)$, $a_H = log(4)$ and $a_{VH} = log(10)$, and drawing the $\epsilon_{i,t}$ terms independently from a standard normal distribution, we define out outcome as:

$$Y_{i,t} = a_L \sum_{j=1}^{4} X_{j,i,t} + a_{VH} X_{5,i,t} + a_M(X_{6,i,t} + X_{8,i,t}) + a_H(X_{7,i,t}) + \Delta D_i + \epsilon_{i,t} \tag{8.2}$$

The outcome for a unit is correlated across time as it is generated from some time-varying covariates. Each simulation consists of 400 treated and 600 control individuals. We consider 1:2 matching. The true treatment effect, Δ, is 0.25.

We consider two alternative ways of generating the continuous outcome variable besides model (8.2). First, we add correlation to the error terms within trajectories. Specifically, the $\epsilon_{i,t}$s for a given trajectory i are generated from a normal distribution with mean 0 and covariance matrix with off diagonal values of 0.8. Second, in addition to the correlated error terms, we square the $X_{2,i,t}$ term in the model, so it is no longer linear. We also run simulations with poor overlap. See Section B.3 of the appendix for results and discussion of simulation performance under poor overlap.

We compare the bias-corrected block bootstrap approach outlined in Section 8.4, using a linear outcome model and a nonparametric bootstrap, to the confidence intervals obtained from weighted least squares (WLS) regression and WLS with clustered standard errors. We focus on the nonparametric bootstrap, as opposed to alternatives such as the wild bootstrap, because of its more common prevalence in practice; however, for comparisons between the wild bootstrap and the nonparametric showing almost equivalent performance in a generally similar setting see Otsu and Rai [2017]. We choose to compare to WLS because this is commonly recommended in matching literature [Ho et al., 2007, Stuart et al., 2011]. However, Abadie and Spiess [2021] pointed out that standard errors from regression may be incorrect due to dependencies among outcomes of matched units, and identified matching with replacement as a setting in which these dependencies are particularly difficult to correct for. Our simulation results suggest that these difficulties carry over into the case of repeated measures. It is worth noting that the standard functions in R used to compute WLS with matching weights such as lm and Zelig (which calls lm), compute biased standard error estimates in most settings. See Section B.2 of the appendix for details.

8.5.2 Results

Tables 8.1 and 8.2 show the coverage and average 95% confidence interval (CI) length, respectively, of WLS regression, WLS regression with clustered standard errors, and bootstrap inference using our model-adjusted ATT estimator, for each of our three simulation settings under 10,000 simulations. As misspecification of the estimated linear outcome model increases the bootstrap method is substantially more robust (although under substantial misspecification the bias-corrected method also fails to achieve nominal coverage). While the bootstrap confidence intervals are generally slightly wider than the WLS and WLS cluster confidence intervals, this is to be expected as the wider confidence intervals lead to improved coverage. In settings where strong scientific knowledge about the exact form of the outcome model is absent, the bootstrap approach appears more reliable than its chief competitors.

Coverage	WLS	WLS Cluster	Bootstrap Bias Corrected
Linear DGP	92.6%	94.4%	94.0%
Linear DGP, Correlated Errors	89.4%	91.7%	94.4%
Nonlinear DGP, Correlated Errors	83.3%	86.2%	89.8%

Table 8.1: Coverage of the WLS, WLS cluster and bootstrap bias corrected methods of inference for our three simulation set-ups.

Average CI Length	WLS	WLS Cluster	Bootstrap Bias Corrected
Linear DGP	0.25	0.27	0.27
Linear DGP, Correlated Errors	0.25	0.27	0.30
Nonlinear DGP, Correlated Errors	0.26	0.28	0.31

Table 8.2: Average 95% confidence interval length for the WLS, WLS cluster and bootstrap bias corrected methods of inference for our three simulation set-ups.

Results in tables 8.1 and 8.2 are for GroupMatch with instance replacement, however matching with trajectory replacement performed very similarly in our simulations. Computation time was similar for GroupMatch with instance replacement and with trajectory replacement. In principle, GroupMatch with instance replacement should be substantially faster, however in its current form GroupMatch does not implement the most computationally efficient algorithm for with instance replacement. Over 100 iterations, the average matching computation time was 4.63 seconds for matching with instance replacement and 4.71 seconds for matching with trajectory replacement. The average block bootstrap computation time was 2.51 seconds. Computation time was calculated on an Apple M1 Max 10-core CPU with 3.22 GHz processor and 64 GB RAM running on macOS Monterey.

8.6 Testing for Timepoint Agnosticism

The key advantage of GroupMatch relative to other matching techniques designed for rolling enrollment settings is its ability to consider and optimize over matches between units at different timepoints, which leads to higher quality matches on lagged covariates. This

advantage comes with a price in additional assumptions, notably the assumption of timepoint agnosticism. Timepoint agnosticism means that mean potential outcomes under control for any two individual timepoints in the data should be identical; in particular, this rules out time trends of any kind in the outcome model that cannot be explained by covariates in the prior L timepoints.

While in many applications scientific intuition about the data generating process suggests this assumption may be reasonable, it is essential that we consider any information contained in the observed data about whether it holds in a particular case. Accordingly, we present a falsification test for timepoint agnosticism. Falsification tests are tests "for treatment effects in places where the analyst knows they should not exist," [Keele, 2015] and are useful in a variety of settings in observational studies [Rosenbaum, 1999]. In particular, our test is designed to detect violations of timepoint agnosticism, or "treatment effects of time" when they should be absent; rejections indicate settings in which GroupMatch is not advisable and other rolling enrollment matching techniques that do not rely on timepoint agnosticism are likely more suitable. While failure to reject may not constitute proof positive of timepoint agnosticism's validity, it rules out gross violations, thereby limiting the potential for bias.

To test the timepoint agnosticism assumption we use *control-control time matching*: matching control units at different timepoints and testing if the average difference in outcomes between the two timepoint groups, conditional on relevant covariates, is significantly different from zero using a bootstrap test. Specifically, restricting attention to trajectories i from the control group, we select two timepoints t_0 and t_1 and match each instance at one timepoint to one at the other timepoint using the GroupMatch optimization routine, based on similarity of covariate histories over the previous L timepoints. Since this match compares instances at two fixed time points, any optimal method of matching without replacement may be used. One practical issue arises: GroupMatch and related matching routines expect one group to be designated "treated," all members of which are generally retained in the match, and the other "control," some members of which will be included, but both matching groups are controls in this case. We label whichever of the two groups has fewer instances as treated; without loss of generality, we will assume there are fewer instances at time t_1 and use these instances as the reference group to be retained.

The test statistic for the falsification test is motivated by the ATT estimator in section 8.2.2. Let N_c be the total number of control units and let N_{t_1} be the number of control instances at time t_1. Let $\hat{\mu}_0^{t_0}$ be a bias correction model fit on our new control group (i.e., control instances at time t_0). In addition, let $D'_{it} = 1$ if unit i is present at time t. We define the test statistic as follows:

$$\begin{aligned}
\hat{\Delta}_{cc} &= \frac{1}{N_{t_1}} \sum_{i=1}^{N_c} D'_{it_1}((Y_{i,t=t_1} - \hat{\mu}_0^{t_0}(\mathbf{X}_{i,t=t_1})) - \sum_{j=1}^{N_c} M_{it_1,jt_0}(Y_{j,t=t_0} - \hat{\mu}_0^{t_0}(\mathbf{X}_{j,t=t_0}))) \\
&= \frac{1}{N_{t_1}} \sum_{i=1}^{N_c} [D'_{it_1}(Y_{i,t=t_1} - \hat{\mu}_0^{t_0}(\mathbf{X}_{i,t=t_1})) - D'_{it_0} K_M(i,t_0)(Y_{i,t=t_0} - \hat{\mu}_0^{t_0}(\mathbf{X}_{i,t=t_0}))] \\
&= \frac{1}{N_{t_1}} \sum_{i=1}^{N_c} \hat{\Delta}_{cc,i}
\end{aligned}$$

We use a bootstrap test to test the following null hypothesis, where $E_0^{t_1}\{\cdot\}$ indicates expectation over the distribution of the covariates in control instances at time t_1.

$$E_0^{t_1}\left\{\mu_0^{t_0}(\mathbf{X})\right\} = E_0^{t_1}\left\{\mu_0^{t_1}(\mathbf{X})\right\}$$

In words, this null hypothesis says that, accounting for differences in the covariate distribution at times 0 and 1, the difference in the average outcomes of control instances at the two timepoints is zero.

The test constructs a bootstrap confidence interval as in Section 8.3 and checks whether the interval covers 0. If the interval covers 0, the test fails to reject. In steps:

1. Label control instances from the first group of trajectories at timepoint t_1 the new "treated" units, and control instances from the second group of trajectories at timepoint t_0 the new "control" units.
2. Fit a bias correction model on the new control units.
3. Match the new treated units to the new control units and calculate the test statistic.
4. Repeat B times:
   (a) Randomly resample N_c elements $\hat{\Delta}*_{cc,i}$ with replacement from $\{\hat{\Delta}_{cc,1}, ..., \hat{\Delta}_{cc,N_c}\}$
   (b) Calculate $\hat{\Delta}_{cc}$ on the resampled data.
5. Construct a $(1-\alpha)$ confidence interval based on the $\alpha/2$ and $1-\alpha/2$ percentile of the values calculated from the boostrap samples.
6. If this confidence interval covers 0, fail to reject the null hypothesis.

We choose to use a bootstrap test here in line with with our inference methods in previous sections. However, it is worth noting that a permutation test is also feasible here.

A key consideration for the falsification test is which timepoints to choose as t_0 and t_1. The choice of timepoint comparison depends largely on what a plausible time trend would be for the problem at hand. For example, if you suspect a linear time trend, it makes sense to look at the first and last timepoints. If the trend is linear, this test should have high power to detect a problem in moderate to large samples. If one is uncertain about the specific shape of the time trend that is most likely to occur and wants to test for all possible trends, we recommend testing each sequential pair of timepoints (i.e., timepoints 1 and 2, 2 and 3, 3 and 4, and so on) and using a multiplicity adjustment.

The falsification test is subject to several common criticisms levied at falsification tests, particularly their ineffectiveness in settings with low power. One possible approach is to reconfigure the test to assume violation of timepoint agnosticism as a null hypothesis and seek evidence in the data to reject it; Hartman and Hidalgo [2018] recommend a similar change for falsification tests used to assess covariate balance, called *equivalence tests*.

The implementation of these modifications is fairly straightforward. First, we must define an equivalence range for our outcome variable: a set of values for which the difference is substantively inconsequential. Let ϵ_L and ϵ_U denote the lower and upper bounds under

which the outcome variable is deemed equivalent. Hartman and Hidalgo [2018] recommend using $\epsilon = \pm 0.36\sigma$ as a default when researchers are unsure of an appropriate equivalence region. Next, prior to step 4 in our falsification test one simply subtracts ϵ_L (and in a separate run ϵ_U) from all treated outcomes. If either one-sided test fails to reject the null, then the test fails.

See Section B.4 of the appendix for simulations illustrating this method.

8.7 Application: Baseball Injuries

We study the impact of short-term injury on hitting performance in observational data from major league baseball (MLB) during 2013-2017. Quantitative studies of major league hitting performance [Baumer, 2008] and of injury trends and impact in athletics [Conte et al., 2016] have been performed repeatedly, but only a few studies so far have evaluated the impact of injury on position players' hitting performance. These have focused on specific injury types, and have not found strong evidence that injury is associated with a decline in performance [Begly et al., 2018, Frangiamore et al., 2018, Wasserman et al., 2015].

We use GroupMatch to match baseball players injured at certain times to similar players at other points in the season that were not injured. We evaluate whether players see a decline in offensive performance immediately after their return from injury. In contrast to other studies, we pool across injury types to see if there is a more general effect of short term injury on hitter performance.

8.7.1 Data and Methodology

We use publicly-available MLB player data from Retrosheet.org and injury data scraped from ProSportsTransactions.com for the years 2013-2017. Our dataset is composed of player height, weight and age, quantities that remain constant over a single season of play, as well as on-base percentage (OBP), plate appearances (PAs) at different points in the season, and dates of short-term injuries, in which the player's team designated him for a 7-10 day stay on the team's official injured list, for each year. OBP is a common measure of hitter performance and is approximately equal to the number of times a player reaches base divided by their number of plate appearances.[2]

For each non-injured player, we generate three pseudo-injury dates evenly spaced over their PAs. In each season, we match injured players to four non-injured players. Matches were formed using GroupMatch with instance replacement, matching on age, weight, height, number of times previously injured, recent performance measured by OBP over the previous 100 PAs, and performance over the entire previous year as measured by end-of-year OBP after James-Stein shrinkage[3] We choose to shrink the OBP using James-Stein to limit the impact of sampling variability for players with a relatively small number of PAs the previous season [Efron and Morris, 1975].

[2]OBP = (Hits + Walks + Hit By Pitch) / (At Bats + Walks + Hit by Pitch + Sacrifice Flies)

[3]See `https://chris-said.io/2017/05/03/empirical-bayes-for-multiple-sample-sizes/` for discussion of James-Stein shrinkage to estimators with variable sample sizes.

Table 8.3 shows the balance for each of the covariates before and after matching. For each covariate, matching shrinks the standardized difference between the treated and control means. The balance achieved is not perfect, especially for the number of previous injuries. This underlines the importance of combining matching with bias-correction to clean up imbalances not removed by matching.

	Treated	Control Mean		Standardized Difference	
Variable	Mean	Before	After	Before	After
Height	73.7	73.1	73.4	0.26	0.14
Weight	213	209	212	0.24	0.07
2016 OBP (JS Shrunk)	.324	.328	.323	-0.09	0.02
Lag OBP	.336	.341	.338	-0.07	-0.02
Birth Year	1988	1988	1988	-0.08	-0.06
Number Previous Injuries	2.73	1.91	2.16	0.30	0.21

Table 8.3: Balance table for MLB injury analysis before and after matching each injured player to four non-injured players.

8.7.2 Results

We compare the results for bias-corrected block bootstrap inference, WLS, and WLS with clustered standard errors. The ATT estimates are positive (0.010), but the 95% confidence intervals cover zero for all methods, indicating that there is not strong evidence that short term injury impacts batter performance. We present the results for 2017 in Figure 8.2. Results from each of 2013 - 2016 were substantively the same, as were results obtained by pooling the matched data across years. The data pass the timepoint agnosticism test, comparing the first and last pseudo-injury dates. We chose to compare the first and last pseudo-injury dates, because we were most concerned about player performance degrading over the course of the entire season due to fatigue. We also perform equivalence tests using $\epsilon = \pm 0.02$, which our data also pass.

We could have also chosen to use a difference-in-differences estimator here, using the difference in performance right before and after the injury, or pseudo-injury, date as our outcome. Results are substantively the same for both estimators. This similarity is due to the close lag OBP matches GroupMatch produces for this example.

8.8 Discussion

The introduction of GroupMatch with instance replacement, a method for block bootstrap inference, and a test for timepoint agnosticism provide substantial new capabilities for matching in settings with rolling enrollment. We now discuss a number of limitations and opportunities for improvement.

Our proof of the block bootstrap approach assumes the use of GroupMatch with instance replacement. The large-sample properties of matched-pair discrepancies are substantially

Figure 8.2: Estimates and 95% confidence intervals for block bootstrap, WLS and cluster WLS inference methods for the ATT in our 2017 baseball injury analysis.

easier to analyze mathematically in this setting than GroupMatch with trajectory replacement or GroupMatch without replacement, designs in which different treated units may compete for the same control units, and the technical argument must be altered to account for this complexity. However, Abadie and Imbens [2012] successfully characterized similar large-sample properties in cross-sectional settings for matching without replacement. While beyond the scope of our work here, we believe it is likely that this approach could provide an avenue for extending Theorem 2 to cover the other two GroupMatch designs. Empirically, we have found that the block bootstrap performs well when matches are calculated using any of the three GroupMatch designs.

Setting aside the technical barriers associated with extending the theory to GroupMatch without replacement, our new approach provides a competitor method to the existing randomization inference framework described by Pimentel et al. [2020] available for GroupMatch without replacement. The randomization inference framework offers the advantage of closely-related methods of sensitivity analysis and freedom from making assumptions about the sampling distribution of the response variables; on the other hand, the block bootstrap method avoids the need to assume a sharp null hypothesis. In general these same considerations arise in choosing between sampling-based inference and randomization-based inference for a cross-sectional matched study, although such choices have received surprisingly little direct and practical attention in the literature thus far.

As described in Section 8.6, the falsification test faces several criticisms, such as ineffectiveness in lower power settings. The modifications to construct equivalence tests as in Hartman and Hidalgo [2018] address these concerns. However, even in the absence of such a change the falsification test may prove useful in concert with a sensitivity analysis. Sensitivity analysis, already widely studied in causal inference as a way to assess the role of ignorability assumptions, places a nonzero bound on the degree of violation of an assumption and reinterprets the study's results under this bound, often repeating the process for larger and larger values of the bound to gain insight. Such a procedure, which focuses primarily on assessing the impact of small or bounded violations of an assumption, naturally complements our falsification test, which can successfully rule out large violations but is more equivocal about minor violations.

Unfortunately, no sensitivity analysis appropriate for block bootstrap inference has been developed yet, either for timepoint agnosticism or other strong assumptions such as ignorability. The many existing methods for sensitivity analysis (developed primarily with ignorability assumptions in mind) are unsatisfying in our framework for a variety of reasons: some rely on randomization inference [Rosenbaum, 2002b], others focus on weighting methods rather than matching [Zhao et al., 2019, Soriano et al., 2021], and others are limited to specific outcome measures [Ding and VanderWeele, 2016] or specific test statistics [Cinelli and Hazlett, 2020]. We view the development of compelling sensitivity analysis approaches to be an especially important methodological objective for matching under rolling enrollment.

Finally, we note that in cross-sectional settings moderate imbalances like those observed after matching in the baseball study in Section 8.7 can often be removed by refining the match to include calipers [Rosenbaum and Rubin, 1985, Yu et al., 2020] or balance constraints [Rosenbaum et al., 2007, Zubizarreta, 2012, Pimentel et al., 2015] on important variables. For computational reasons these constraints are difficult to implement and use in full generality for GroupMatch designs. For example, some balance constraints rely on network flow representations of the matching problem that are not immediately compatible with the network flow representation underpinning GroupMatch. Further work to consider how calipers and balance constraints can be elegantly incorporated will enhance GroupMatch's effectiveness in practice.

Bibliography

A. Abadie and G. W. Imbens. Large sample properties of matching estimators for average treatment effects. *Econometrica*, 74:235–267, 2006.

A. Abadie and G. W. Imbens. On the failure of the bootstrap for matching estimators. *Econometrica*, 76:1537–1557, 2008.

A. Abadie and G. W. Imbens. Bias-corrected matching estimators for average treatment effects. *Journal of Business & Economic Statistics*, 29:1 – 11, 2011.

A. Abadie and G. W. Imbens. A martingale representation for matching estimators. *Journal of the American Statistical Association*, 107:833 – 843, 2012.

A. Abadie and J. Spiess. Robust post-matching inference. *Journal of the American Statistical Association*, pages 1–13, 2021.

D. Acemoglu, S. Naidu, P. Restrepo, and J. A. Robinson. Democracy does cause growth. *Journal of Political Economy*, 127(1):47–100, 2019.

American Statistical Association. American Statistical Association statement on risk-limiting post-election audits. `www.amstat.org/outreach/pdfs/Risk-Limiting_Endorsement.pdf`, 2010.

J. Antonelli, M. Cefalu, N. Palmer, and D. Agniel. Doubly robust matching estimators for high dimensional confounding adjustment. *Biometrics*, 74(4):1171–1179, 2018.

A. Appel and P. Stark. Evidence-based elections: Create a meaningful paper trail, then audit. *Georgetown Law Technology Review*, 4.2:523–541, 2020. `https://georgetownlawtechreview.org/wp-content/uploads/2020/07/4.2-p523-541-Appel-Stark.pdf`.

A. Appel, R. DeMillo, and P. Stark. Ballot-marking devices cannot assure the will of the voters. *Election Law Journal, Rules, Politics, and Policy*, 2020. `https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3375755`.

S. Athey and G. W. Imbens. Design-based analysis in difference-in-differences settings with staggered adoption. *Journal of Econometrics*, 226(1):62–79, 2022.

P. C. Austin and D. S. Small. The use of bootstrapping when using propensity-score matching without replacement: a simulation study. *Statistics in medicine*, 33(24):4306–4319, 2014.

J. Bañuelos and P. Stark. Limiting risk by turning manifest phantoms into evil zombies. Technical report, arXiv.org, 2012. URL `http://arxiv.org/abs/1207.3413`. Retrieved 17 July 2012.

C. Bardelli. *Nonparametric Confidence Intervals Based on Permutation Tests*. PhD thesis, Politecnico Di Milano, 2016.

G. Barnard. Discussion of the spectral analysis of point processes. *Journal of the Royal Statistical Society Series B*, 25, 1963.

B. S. Baumer. Why on-base percentage is a better indicator of future performance than batting average: An algebraic proof. *Journal of Quantitative Analysis in Sports*, 4, 2008.

J. P. Begly, M. S. Guss, T. S. Wolfson, S. A. Mahure, A. S. Rokito, and L. M. Jazrawi. Performance outcomes after medial ulnar collateral ligament reconstruction in major league baseball positional players. *Journal of Shoulder and Elbow Surgery*, 27:282–290, 2018.

E. Ben-Michael, A. Feller, and J. Rothstein. Synthetic controls and weighted event studies with staggered adoption. *arXiv preprint arXiv:1912.03290*, 2021.

J. Benaloh, D. Jones, E. Lazarus, M. Lindeman, and P. Stark. SOBA: Secrecy-preserving observable ballot-level audits. In *Proceedings of the 2011 Electronic Voting Technology Workshop / Workshop on Trustworthy Elections (EVT/WOTE '11)*. USENIX, 2011. URL `http://statistics.berkeley.edu/~stark/Preprints/soba11.pdf`.

R. L. Berger and D. D. Boos. P values maximized over a confidence set for the nuisance parameter. *Journal of the American Statistical Association*, 89(427):1012–1016, 1994.

M. Bernhard, A. McDonald, H. Meng, J. Hwa, N. Bajaj, K. Chang, and J. Halderman. Can voters detect malicious manipulation of ballot marking devices? *41st IEEE Symposium on Security and Privacy*, 2020. `https://jhalderm.com/pub/papers/bmd-verifiability-sp20.pdf`.

P. Bickel and K. Doksum. *Mathematical Statistics: Basic Ideas and Selected Topics*. Pearson, 2006.

Z. W. Birnbaum. *Reliability and Biometry*, chapter Computers and unconventional test-statistics, pages 441–458. Philadelphia: SIAM, 1974.

M. Blair-Loy, L. Rogers, D. Glaser, Y. A. Wong, D. Abraham, and P. Cosman. Gender in engineering departments: Are there gender differences in interruptions of academic job talks? *Social Sciences*, 6:29, 2017.

M. Blom, P. Stuckey, and V. Teague. RAIRE: Risk-limiting audits for IRV elections. `https://arxiv.org/abs/1903.08804`, 2019.

M. Blom, A. Conway, D. King, L. Sandrolini, P. Stark, P. Stuckey, and V. Teague. You can do RLAs for IRV. In R. Krimmer, M. Volkamer, B. Beckert, A. Maurer, D. Duenas-Cid, S. Glondu, I. Krivonosova, O. Kulyk, R. Küsters, B. Martin-Rozumilowicz, P. Ronne,

M. Solvak, and O. Spycher, editors, *E-VOTE-ID 2020*, pages 296–310, Tallinn, Estonia, 2020. TalTech Press.

A. A. Bohl, P. A. Fishman, M. A. Ciol, B. Williams, J. LoGerfo, and E. A. Phelan. A longitudinal analysis of total 3-year healthcare costs for older adults who experience a fall requiring medical care. *Journal of the American Geriatrics Society*, 58(5):853–860, 2010.

E. Bølviken and E. Skovlund. Confidence intervals from monte carlo tests. *Journal of the American Statistical Association*, 91(435):1071–1078, 1996.

A. Boring, K. Ottoboni, and P. Stark. Student evaluations of teaching (mostly) do not measure teaching effectiveness. *ScienceOpen Research*, 2016.

Brennan Center for Justice, Rhode Island RLA Working Group. Pilot implementation study of risk-limiting audit methods in the state of Rhode Island, 2019. URL `https://www.brennancenter.org/our-work/research-reports/pilot-implementation-study-risk-limiting-audit-methods-state-rhode-island`.

A. Broder. On the resemblance and containment of documents. In *Compression and Complexity of Sequences*, pages 21–29, Washington DC, 1997. IEEE.

A. Broder, S. Glassman, M. Manasse, and G. Zweig. Syntactic clustering of the web. In *Selected Papers from the Sixth International World Wide Web Conference*, page 1157–1166, Essex, UK, 1997. Elsevier Science Publishes Ltd.

D. L. Brunsma, D. G. Embrick, and J. H. Shin. Graduate students of color: Race, racism, and mentoring in the white waters of academia. *Sociology of Race and Ethnicity*, 3:1–13, 2017.

A. E. Budden, T. Tregenza, L. W. Aarssen, J. Koricheva, R. Leimu, and C. J. Lortie. Double-blind review favours increased representation of female authors. *Trends in ecology & evolution*, 23:4–6, 2008.

N. Caplar, S. Tacchella, and S. Birrer. Quantitative evaluation of gender bias in astronomical publications from citation counts. *Nature Astronomy*, 1:1–5, 2017.

D. Caughey, A. Dafoe, and L. Miratrix. Beyond the sharp null: Randomization inference, bounded null hypotheses, and confidence intervals for maximum effects. *arXiv preprint arXiv:1709.07339*, 2017.

CDOT. Colorado dot non-motorized monitoring program evaluation and implementation plan. Technical report, 2016. URL `https://www.codot.gov/programs/bicycleped/documents/2016-10-21-cdot-nonmotorized-monitoring-plan_low-res.pdf`.

C. Cinelli and C. Hazlett. Making sense of sensitivity: Extending omitted variable bias. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(1):39–67, 2020.

Colorado Secretary of State. Audit Center, 2020. URL https://www.sos.state.co.us/pubs/elections/auditCenter.html.

S. Conte, C. L. Camp, and J. S. Dines. Injury trends in major league baseball over 18 seasons: 1998-2015. *Am J Orthop*, 45:116–123, 2016.

J. R. A. Davenport, M. Fouesneau, E. Grand, A. Hagen, K. Poppenhaeger, and L. L. Watkins. Studying gender in conference talks–data from the 223rd meeting of the american astronomical society. *ArXiv Preprint ArXiv:1403.3091*, 2014.

A. C. Davison and D. V. Hinkley. *Bootstrap methods and their application*. Cambridge university press, 1997.

P. Ding and T. J. VanderWeele. Sensitivity analysis without assumptions. *Epidemiology (Cambridge, Mass.)*, 27(3):368, 2016.

J.-M. Dufour. Monte carlo tests with nuisance parameters: A general approach to finite-sample inference and nonstandard asymptotics. *Journal of econometrics*, 133(2):443–477, 2006.

P. Dupas, A. S. Modestino, M. Niederle, and J. Wolfers. Gender and the dynamics of economics seminars. Technical report, National Bureau of Economic Research, 2021.

M. Dwass. Modified randomization tests for nonparametric hypotheses. *The Annals of Mathematical Statistics*, pages 181–187, 1957.

B. Efron and C. Morris. Data analysis using stein's estimator and its generalizations. *Journal of the American Statistical Association*, 70:311–319, 1975.

S. N. Evans and P. B. Stark. Inverse problems as statistics. *Inverse problems*, 18(4):R55–R97, 2002.

Federal Highway Administration. Traffic monitoring guide. Technical report, 2016. URL https://www.fhwa.dot.gov/policyinformation/tmguide/tmg_fhwa_pl_17_003.pdf.

C. B. Fogarty. Studentized sensitivity analysis for the sample average treatment effect in paired observational studies. *Journal of the American Statistical Association*, 115(531): 1518–1530, 2020.

K. Foote. Report on the inyo county risk-limiting audit pilot. https://lhc.ca.gov/sites/lhc.ca.gov/files/Reports/247/WrittenTestimony/FooteJuly2018.pdf, 2018. Retrieved: November 30th, 2020.

R. V. Foutz. A method for constructing exact tests from test statistics that have unknown null distributions. *Journal of Statistical Computation and Simulation*, 10(3-4):187–193, 1980.

S. J. Frangiamore, S. Mannava, K. K. Briggs, S. McNamara, and M. J. Philippon. Career length and performance among professional baseball players returning to play after hip arthroscopy. *The American Journal of Sports Medicine*, 46:2588–2593, 2018.

P. H. Garthwaite. Confidence intervals from randomization tests. *Biometrics*, 52:1387–1393, 1996.

P. H. Garthwaite and M. C. Jones. A stochastic approximation method and its application to confidence intervals. *Journal of Computational and Graphical Statistics*, 18:184–200, 2009.

A. K. Glazer and S. D. Pimentel. Robust inference for matching under rolling enrollment. *Journal of Causal Inference*, 11(1):20220055, 2023.

A. K. Glazer, J. V. Spertus, and P. B. Stark. Bayesian audits are average but risk-limiting audits are above average. In *Electronic Voting: 5th International Joint Conference, E-Vote-ID 2020, Bregenz, Austria, October 6–9, 2020, Proceedings 5*, pages 84–94. Springer, 2020.

A. K. Glazer, J. V. Spertus, and P. B. Stark. More style, less work: card-style data decrease risk-limiting audit sample sizes. *Digital Threats: Research and Practice (DTRAP)*, 2:1–15, 2021.

A. K. Glazer, H. Luo, S. Devgon, C. Wang, X. Yao, S. S. Ye, F. McQuarrie, Z. Li, A. Palma, Q. Wan, W. Gu, A. Sen, Z. Wang, G. D. O'Connell, and P. B. Stark. Look who's talking: gender differences in academic job talks. *ScienceOpen*, 2023a.

A. K. Glazer, J. V. Spertus, and P. B. Stark. Stylish risk-limiting audits in practice. In *Proceedings of E-VOTE-ID 2023, Lecture Notes in Informatics*, 2023b.

P. I. Good. *Resampling methods*. Springer, 2006.

B. B. Hansen. Full matching in an observational study of coaching for the sat. *Journal of the American Statistical Association*, 99(467):609–618, 2004.

B. B. Hansen. The prognostic analogue of the propensity score. *Biometrika*, 95(2):481–488, 2008.

M. T. Harrison. Conservative hypothesis tests and confidence intervals using importance sampling. *Biometrika*, 99(1):57–69, 2012.

J. A. Hartigan. Using subsample values as typical values. *Journal of the American Statistical Association*, 64:1303–1317, 1969.

E. Hartman and F. D. Hidalgo. An equivalence approach to balance and placebo tests. *American Journal of Political Science*, 62(4):1000–1013, 2018. doi: 10.1111/ajps.12387. URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/ajps.12387`.

A. Haviland, D. S. Nagin, P. R. Rosenbaum, and R. E. Tremblay. Combining group-based trajectory modeling and propensity score matching for causal inferences in nonexperimental longitudinal data. *Developmental psychology*, 44(2):422–436, 2008.

J. Hemerik. On the term "randomization test". *The American Statistician*, pages 1–8, 2024.

J. Hemerik and J. Goeman. Exact testing with random permutations. *TEST*, 27:811–825, 2018. doi: 10.1007/s11749-017-0571-1.

J. Hemerik and J. J. Goeman. Another look at the lady tasting tea and differences between permutation tests and randomisation tests. *International Statistical Review*, 89(2):367–381, 2021.

J. J. Higgins. *An introduction to modern nonparametric statistics*. Brooks/Cole Pacific Grove, CA, 2004.

D. E. Ho, K. Imai, G. King, and E. A. Stuart. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political analysis*, 15:199–236, 2007.

L. Howard, R. Rivest, and P. Stark. A review of robust post-election audits: Various methods of risk-limiting audits and Bayesian audits. Technical report, Brennan Center for Justice, 2019. `https://www.brennancenter.org/sites/default/files/2019-11/2019_011_RLA_Analysis_FINAL_0.pdf`.

Z. Huang, R. Rivest, P. Stark, V. Teague, and D. Vukcevic. A unified evaluation of two-candidate ballot-polling election auditing methods. In *Proceedings of the 5th Annual Conference on Electronic Voting (E-Vote-ID '20)*, 2020.

K. Imai, I. S. Kim, and E. Wang. Matching methods for causal inference with time-series cross-section data. Technical report, Harvard University, 2020.

G. W. Imbens. Experimental design for unit and cluster randomid trials. In *Conference International Initiative for Impact Evaluation, Cuernavaca*, 2011.

G. W. Imbens and D. B. Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge university press, 2015.

K. N. Jackson, S. W. O'Brien, S. E. Searcy, and S. E. Warchol. Quality assurance and quality control processes for a large-scale bicycle and pedestrian volume data program. *Transportation research record*, 2644(1):19–29, 2017.

A. Kaatz, B. Gutierrez, and M. Carnes. Threats to objectivity in peer review: the case of gender. *Trends in pharmacological sciences*, 35:371–373, 2014.

A. Kapelner and A. Krieger. Matching on-the-fly: Sequential allocation with higher power and efficiency. *Biometrics*, 70(2):378–388, 2014.

L. Keele. The statistics of causal inference: A view from political methodology. *Political Analysis*, pages 313–335, 2015.

O. Kempthorne and T. Doerfler. The behaviour of some significance tests under experimental randomization. *Biometrika*, 56(2):231–248, 1969.

S. Kothuri, J. Broach, N. McNeil, K. Hyun, S. Mattingly, M. M. Miah, K. Nordback, and F. Proulx. Exploring data fusion techniques to estimate network-wide bicycle volumes. 2022.

E. L. Lehmann, J. P. Romano, and G. Casella. *Testing statistical hypotheses*, volume 3. New York: springer, 2005.

X. Li and P. Ding. General forms of finite population central limit theorems with applications to causal inference. *Journal of the American Statistical Association*, 112(520):1759–1769, 2017.

Y. P. Li, K. J. Propert, and P. R. Rosenbaum. Balanced risk set matching. *Journal of the American Statistical Association*, 96(455):870–882, 2001.

M. Lindeman and P. Stark. A gentle introduction to risk-limiting audits. *IEEE Security and Privacy*, 10:42–49, 2012.

M. Lindeman, P. Stark, and V. Yates. BRAVO: Ballot-polling risk-limiting audits to verify outcomes. In *Proceedings of the 2011 Electronic Voting Technology Workshop / Workshop on Trustworthy Elections (EVT/WOTE '11)*. USENIX, 2012.

M. Lindeman, N. McBurnett, K. Ottoboni, and P. Stark. Next steps for the colorado risk-limiting audit (corla) program. `https://arxiv.org/abs/1803.00698`, 2018.

G. Lindsey, S. Coll, and G. Stewart. Quality assurance methods for hourly nonmotorized traffic counts. *Transportation research record*, 2678(2):723–742, 2024.

B. Lu. Propensity score matching with time-dependent covariates. *Biometrics*, 61:721–728, 2005.

J. M. Madera, M. R. Hebl, and R. C. Martin. Gender and letters of recommendation for academia: Agentic and communal differences. *Journal of Applied Psychology*, 94:1591–1599, 2009.

S. McCammon. Virginia Republican David Yancey wins tie-breaking drawing. `https://www.npr.org/2018/01/04/573504079/virginia-republican-david-yancey-wins-tie-breaking-drawing`, 2018.

N. McNeil and K. Tufte. Biking and walking quality counts: Using "bikeped portal" counts to developdata quality checks. 2019.

MDOT. Minnesota's walking and bicycling data collection report: Annual data from 2014 to 2017. Technical report, 2018. URL `https://www.dot.state.mn.us/bike/documents/planning-research/bike-ped-report.pdf`.

Michigan Secretary of State. Pilot audit of march presidential primary results showcases security, accuracy of Michigan elections systems, 2020. URL `https://www.michigan.gov/sos/0,4670,7-127--531561--,00.html`.

S. Morin, G. McClearn, N. McBurnett, P. Vora, and F. Zagorski. A note on risk-limiting Bayesian polling audits for two-candidate elections. *Voting '20*, in press, 2020.

C. A. Moss-Racusin, J. F. Dovidio, V. L. Brescoll, M. J. Graham, and J. Handelsman. Science faculty's subtle gender biases favor male students. *Proceedings of the national academy of sciences*, 109:16474–16479, 2012.

National Academies of Sciences, Engineering, and Medicine. *Securing the Vote: Protecting American Democracy*. The National Academies Press, Washington, DC, 2018a. ISBN 978-0-309-47647-8. doi: 10.17226/25120. URL `https://www.nap.edu/catalog/25120/securing-the-vote-protecting-american-democracy`.

National Academies of Sciences, Engineering, and Medicine. *Securing the Vote: Protecting American Democracy*. The National Academies Press, Washington, DC, 2018b. ISBN 978-0-309-47647-8. doi: 10.17226/25120. URL `https://www.nap.edu/catalog/25120/securing-the-vote-protecting-american-democracy`.

J. Neyman and E. S. Pearson. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231:289–337, 1933.

M. Nguyen. *Nonparametric Inference Using Randomization and Permutation Reference Distribution and Their MonteCarlo Approximation*. PhD thesis, Portland State University, 2009.

K. Nordback, S. Kothuri, T. Petritsch, P. McLeod, E. Rose, and H. Twaddell. Exploring pedestrian counting procedures. a review and compilation of existing procedures, good practices, and recommendations. Technical report, Federal Highway Administration, 2016.

T. W. O'Gorman. Regaining confidence in confidence intervals for the mean treatment effect. *Statistics in medicine*, 33:3859–3868, 2014.

T. Otsu and Y. Rai. Bootstrap inference of matching estimators for average treatment effects. *Journal of the American Statistical Association*, 112:1720–1732, 2017.

K. Ottoboni and P. Stark. Election integrity and electronic voting machines in 2018 Georgia, USA. In *E-Vote-ID 2019 Proceedings*, 2019. Preprint: `https://ssrn.com/abstract=3426250`.

K. Ottoboni, M. Bernhard, A. Halderman, R. Rivest, and P. Stark. Bernoulli ballot polling: A manifest improvement for risk-limiting audits. In *Proceedings of the 4th Annual Workshop on Advances in Secure Electronic Voting (Voting'19)*, 2018a. Preprint: `http://arxiv.org/abs/1812.06361`.

K. Ottoboni, P. Stark, M. Lindeman, and N. McBurnett. Risk-limiting audits by stratified union-intersection tests of elections (SUITE). In *Electronic Voting. E-Vote-ID 2018. Lecture Notes in Computer Science*. Springer, 2018b. `https://link.springer.com/chapter/10.1007/978-3-030-00419-4_12`.

K. Ottoboni, P. B. Stark, M. Lindeman, and N. McBurnett. Risk-limiting audits by stratified union-intersection tests of elections (suite). In *International Joint Conference on Electronic Voting*, pages 174–188. Springer, 2018c.

A. Ozgumus, H. Rau, S. Trautmann, and C. Konig-Kersting. Gender bias in the evaluation of teaching materials. *Frontiers in Psychology*, 11:1074, 2020.

M. Pagano and D. Tritchler. On obtaining permutation distributions in polynomial time. *Journal of the American Statistical Association*, 78(382):435–440, 1983.

F. Pesarin and L. Salmaso. *Permutation Tests for Complex Data: Theory, Applications and Software*. Wiley, 2010.

S. D. Pimentel, R. R. Kelz, J. H. Silber, and P. R. Rosenbaum. Large, sparse optimal matching with refined covariate balance in an observational study of the health outcomes produced by new surgeons. *Journal of the American Statistical Association*, 110(510): 515–527, 2015.

S. D. Pimentel, L. V. Forrow, J. Gellar, and J. Li. Optimal matching approaches in health policy evaluations under rolling enrollment. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 183:1411–1435, 2020.

E. J. Pitman. Significance tests which may be applied to samples from any populations. *Supplement to the Journal of the Royal Statistical Society*, 4(1):119–130, 1937.

A. Ramdas, R. F. Barber, E. J. Candès, and R. J. Tibshirani. Permutation tests using arbitrary permutation distributions. *Sankhya A*, pages 1–22, 2023.

E. Reuben, P. Sapienza, and L. Zingales. How stereotypes impair women's careers in science. *Proceedings of the National Academy of Sciences*, 111:4403–4408, 2014.

R. Richie and H. Smith. A survey and analysis of statewide election recounts 2000–2015. `https://fairvote.app.box.com/v/recounts`, 2015.

R. Rivest. Consistent sampling with replacement. `https://arxiv.org/abs/1808.10016v1`, 2018a.

R. Rivest and E. Shen. A Bayesian method for auditing elections. In *Proceedings of the 2012 Electronic Voting Technology Workshop / Workshop on Trustworthy Elections (EVT/WOTE '12)*. USENIX, August 2012.

R. L. Rivest. Bayesian tabulation audits: Explained and extended. `https://arxiv.org/abs/1801.00528`, January 1, 2018b.

H. Robbins and S. Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, pages 400–407, 1951.

J. Roll. Nonmotorized traffic monitoring and crash analysis. Technical report, Oregon. Dept. of Transportation. Research Section, 2021.

P. R. Rosenbaum. The consequences of adjustment for a concomitant variable that has been affected by the treatment. *Journal of the Royal Statistical Society: Series A (General)*, 147(5):656–666, 1984.

P. R. Rosenbaum. Choice as an alternative to control in observational studies. *Statistical science*, pages 259–278, 1999.

P. R. Rosenbaum. Covariance adjustment in randomized experiments and observational studies. *Statistical Science*, 17(3):286–327, 2002a.

P. R. Rosenbaum. *Observational Studies*. Springer, New York, NY, 2002b.

P. R. Rosenbaum and D. B. Rubin. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, 39 (1):33–38, 1985.

P. R. Rosenbaum, R. N. Ross, and J. H. Silber. Minimum distance matched sampling with fine balance in an observational study of treatment for ovarian cancer. *Journal of the American Statistical Association*, 102(477):75–83, 2007.

D. B. Rubin. Using multivariate matched sampling and regression adjustment to control bias in observational studies. *Journal of the American Statistical Association*, pages 318–328, 1979.

P. Ryus, E. M. Ferguson, K. M. Laustsen, R. J. Schneider, F. R. Proulx, T. Hull, and L. Miranda-Moreno. *Guidebook on pedestrian and bicycle volume data collection*, volume 797. Transportation Research Board Washington, DC, 2014.

L. Salmaso and F. Pesarin. *Permutation tests for complex data: theory, applications and software*. John Wiley & Sons, 2010.

M. L. Samuels, J. A. Witmer, and A. A. Schaffner. *Statistics for the life sciences*, volume 4. Prentice Hall Upper Saddle River, NJ, 2003.

H. Sarsons. Gender differences in recognition for group work. *Harvard University*, 2015.

T. Schmader, J. Whitehead, and V. H. Wysocki. A linguistic comparison of letters of recommendation for male and female chemistry and biochemistry job applicants. *Sex roles*, 57:509–514, 2007.

Select Committee on Intelligence. Russian active measures campaigns and interference in the 2016 U.S. election. `https://www.intelligence.senate.gov/sites/default/files/documents/Report_Volume1.pdf`, 2019.

S. Singer and N. McBurnett. Orange county, ca pilot risk-limiting audit. `https://www.verifiedvoting.org/wp-content/uploads/2018/12/2018-RLA-Report-Orange-County-CA.pdf`, 2018. Retrieved: November 30, 2020.

G. Snedecor and W. Cochran. *Statistical Methods*, volume 6. Ames, Iowa: Iowa State University Press, 1967.

D. Soriano, E. Ben-Michael, P. J. Bickel, A. Feller, and S. D. Pimentel. Interpretable sensitivity analysis for balancing weights. *arXiv preprint arXiv:2102.13218*, 2021.

J. Spertus. COBRA: Comparison-optimal betting for risk-limiting audits. *The Workshop on Advances in Secure Electronic Voting*, 2023. In Press.

J. Splawa-Neyman, D. M. Dabrowska, and T. P. Speed. On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, pages 465–472, 1990.

P. Stark. Conservative statistical post-election audits. *Ann. Appl. Stat.*, 2:550–581, 2008a. URL `http://arxiv.org/abs/0807.4005`.

P. Stark. Election audits by sampling with probability proportional to an error bound: dealing with discrepancies. `https://www.stat.berkeley.edu/~stark/Preprints/ppebwrwd08.pdf`, 2008b.

P. Stark. Super-simple simultaneous single-ballot risk-limiting audits. In *Proceedings of the 2010 Electronic Voting Technology Workshop / Workshop on Trustworthy Elections (EVT/WOTE '10)*. USENIX, 2010. URL `http://www.usenix.org/events/evtwote10/tech/full_papers/Stark.pdf`.

P. Stark. An introduction to risk-limiting audits and evidence-based elections, written testimony prepared for the Little Hoover Commission. `https://www.stat.berkeley.edu/~stark/Preprints/lhc18.pdf`, 2018.

P. Stark. Sets of half-average nulls generate risk-limiting audits: SHANGRLA. *Voting '20*, in press, 2020. Preprint: `http://arxiv.org/abs/1911.10035`.

P. Stark. Non(c)esuch ballot-level risk-limiting audits for precinct-count voting systems. In S. e. a. Katsikas, editor, *Computer Security. ESORICS 2022 International Workshops. Lecture Notes in Computer Science,13785*, pages 541–554, Cham, 2023a. Springer. doi: 10.1007/978-3-031-25460-4_31.

P. Stark. ONEAudit: Overstatement-net-equivalent risk-limiting audit. In *Proceedings of the 8th Annual Workshop on Advances in Secure Electronic Voting (Voting'23)*. Springer, 2023 (in press).

P. Stark and K. Ottoboni. Random sampling: Practice makes imperfect. *arXiv preprint arXiv:1810.10985*, 2018.

P. B. Stark. Inference in infinite-dimensional inverse problems: discretization and duality. *Journal of Geophysical Research: Solid Earth*, 97(B10):14055–14082, 1992.

P. B. Stark. ALPHA: Audit that learns from previously hand-audited ballots. *Annals of Applied Statistics*, 17:641–679, 2023b. doi: 10.1214/22-AOAS1646.

P. B. Stark and D. A. Wagner. Evidence-based elections. *IEEE Security and Privacy*, 10:33–41, 2012. `https://www.stat.berkeley.edu/~stark/Preprints/evidenceVote12.pdf`.

E. Stuart. Matching methods for causal inference: A review and a look forward. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 25, 2010.

E. A. Stuart, G. King, K. Imai, and D. Ho. Matchit: nonparametric preprocessing for parametric causal inference. *Journal of statistical software*, 2011.

D. Tritchler. On inverting permutation tests. *Journal of the American Statistical Association*, 79:200–207, 1984.

S. Turner and P. Lasley. Quality counts for pedestrians and bicyclists: Quality assurance procedures for nonmotorized traffic count data. *Transportation research record*, 2339(1): 57–67, 2013.

S. Turner, P. Lasley, J. Hudson, R. Benz, et al. Guide for pedestrian and bicyclist count data submittal. Technical report, Texas Department of Transportation, 2019.

Verified Voting. The Verifier, 2020. URL `https://verifiedvoting.org/verifier/#mode/navigate/map/ppEquip/mapType/normal/year/2020`.

P. Vora. Risk-limiting Bayesian polling audits for two-candidate elections. `https://arxiv.org/abs/1902.00999`, 2019.

E. B. Wasserman, B. Abar, M. N. Shah, D. Wasserman, and J. J. Bazarian. Concussions are associated with decreased batting performance among major league baseball players. *The American Journal of Sports Medicine*, 43:1127–1133, 2015.

I. Waudby-Smith, P. Stark, and A. Ramdas. RiLACS: Risk Limiting Audits via Confidence Sequences. In R. Krimmer, M. Volkamer, D. Duenas-Cid, O. Kulyk, P. Rønne, M. Solvak, and M. Germann, editors, *Electronic Voting*, pages 124–139, Cham, 2021. Springer International Publishing. ISBN 978-3-030-86942-7.

A. Witman, C. Beadles, Y. Liu, A. Larsen, N. Kafali, S. Gandhi, P. Amico, and T. Hoerger. Comparison group selection in the presence of rolling entry for health services research: Rolling entry matching. *Health services research*, 54:492–501, 2019.

H. Witteman, M. Henricks, S. Straus, and C. Tannenbaum. Female grant applicants are equally successful when peer reviewers assess the science, but not when they assess the scientist. *Biorxiv*, 2018.

J. Wu and P. Ding. Randomization tests for weak null hypotheses in randomized experiments. *Journal of the American Statistical Association*, 116(536):1898–1913, 2021.

R. Yu, J. H. Silber, and P. R. Rosenbaum. Matching methods for observational studies derived from large administrative databases. *Statistical Science*, 35(3):338–355, 2020.

K. Zetter. The crisis of election security. *The New York Times*, 2018. `https://www.nytimes.com/2018/09/26/magazine/election-security-crisis-midterms.html`.

K. Zetter. Critical U.S. election systems have been left exposed online despite official denials. *Vice*, 2019a. https://www.vice.com/en_us/article/3kxzk9/exclusive-critical-us-election-systems-have-been-left-exposed-online-despite-official

K. Zetter. How close did Russia really come to hacking the 2016 election? *Politico*, 2019b. https://www.politico.com/news/magazine/2019/12/26/did-russia-really-hack-2016-election-088171.

Y. Zhang and Q. Zhao. What is a randomization test? *Journal of the American Statistical Association*, 118(544):2928–2942, 2023.

Q. Zhao, D. S. Small, and B. B. Bhattacharya. Sensitivity analysis for inverse probability weighting estimators via the percentile bootstrap. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2019.

J. R. Zubizarreta. Using mixed integer programming for matching in an observational study of kidney failure after surgery. *Journal of the American Statistical Association*, 107(500): 1360–1371, 2012.

Appendix A

Appendix for Chapter 3

A.1 Data Quality

We examined three sets of annotations to get a sense of data quality. The first two sets were collected towards the beginning of the annotation process: 21 videos from ME and 17 from Physics. The last set was collected towards the end: 7 additional videos from ME.

First, we looked at how often the two annotators agreed there was an audience utterance, regardless of how it was labelled. For the first ME set and the Physics annotations the agreement was 74% and 69% respectively. For the second ME set, the agreement was 74%.

We randomly sampled and reviewed several videos to understand the source of these discrepancies.

In one video, there were no discrepancies: both annotators found 6 audience utterances.

In another video, one annotator found 32 audience utterances and the other annotator found 37 audience utterances, including all 32 the first annotator found. We summarize this as

$$\frac{\#\text{ events labelled by both as an interruption}}{\#\text{ events labelled by either as an interruption}} = \frac{32}{37} = 86\%\text{ accuracy.}$$

Of the 5 times the second annotator found an interruption but the first did not, 3 were interruptions that lasted less than 2 seconds and a fourth lasted 8 seconds. The fifth discrepancy was that the first annotator missed a question from an audience member who had interrupted the presenter's response to a different question from a different audience member, coding the exchange as one interruption when it was two.

In the third video, annotators agreed on $18/26 = 69\%$ of the utterances one or both identified. There were 8 discrepancies. Six involved utterances that lasted less than 5 seconds. One annotator missed a 9-second question at the end of the video. The last resulted from one annotator coding two quick interjections (6 seconds and 4 seconds with a 3-second presenter remark in between) as one interjection.

In general, discrepancies arose from an annotator missing an audience utterance during a quick exchange, especially in videos with many audience interjections. When both annotators agreed there was an utterance, they generally agreed on how to code that utterance.

A.2 Detailed comparison with Blair-Loy et al.

A.2.1 Video Annotation

Our system of annotations differs slightly from that used by Blair-Loy et al. [2017]: They used three categories of audience utterances: acknowledged questions, follow-up questions, and unacknowledged interruptions. They label all follow-up questions "acknowledged," while we label a follow-up question "unacknowledged" if the speaker was cut off by the audience member to ask another question. We also include "attempted interruptions," which Blair-Loy et al. [2017] do not appear to include in their analysis—we think their taxonomy classifies attempted interruptions as interruptions. Blair-Loy et al. [2017] consider only the pre-Q&A period; we consider both pre-Q&A and Q&A.

We found that individual reviewers may miss some utterances, so two raters reviewed each video and a third rater resolved differences (See Appendix A.1). In the Blair-Loy et al. [2017] study, one person annotated each video. In the Dupas et al. [2021] study, some talks were annotated by two raters—who often disagreed substantially—but most were annotated by only one (untrained) rater in real time. Our raters found it necessary to rewind and review portions of the video repeatedly to accurately code rapid exchanges between the audience and the speaker, so we expect that the data quality in Dupas et al. [2021] is uneven.

A.2.2 Statistical Analysis

Blair-Loy et al. [2017] use a statistical test based on the coefficient of gender in a ZINB regression that includes data from all departments (see Section 5 of Blair-Loy et al. [2017]).

They note that ZINB is a common model for "overdispersed" count data. However, that does not justify using it as a basis for *inference*, which requires the data to have been generated by the ZINB model.

The ZINB model involves two sub-models: a model for the probability zero questions are asked (the *zero model*) and a model for the number of questions given that at least one question was asked (the *positive model*). The zero model is a logistic function of a linear combination of covariates, and the positive model is a negative binomial model in which the parameters are a function of a set of covariates, including presenter gender.

The test statistic is the gender coefficient in the positive model. We find this noteworthy because that coefficient does not capture whether male or female presenters get more questions overall; it only involves the distribution of the number of questions given that there were some.

Blair-Loy et al. [2017] translate the scientific hypothesis that there is no gender bias into the statistical hypothesis that the gender coefficient in the positive model equals zero. The P-value is computed on the assumption that the ZINB model is *true*, i.e., it is how the data were generated.

A.2.3 ZINB Test on the New Data

We fit a ZINB model to our pooled pre-Q&A data using the same covariates Blair-Loy et al. [2017] used: proportion of faculty who are women and years since the presenter received

a PhD. Tables A.1 and A.2 give the results. The resulting nominal P-values are smaller than those for our randomization test.

The parametric P-values are uninterpretable when the parametric assumptions are false, i.e., when the number of questions is not generated by a ZINB model (with the assumed functional relationship between the included covariates and the parameters of the model). Those assumptions are implausible, but one can still use the estimated coefficient of gender in the ZINB positive model to construct a valid test by calibrating the null distribution of the coefficient using randomization rather than relying on the parametric assumptions, as we now describe.

For each random assignment of the presenter gender, we fit the ZINB model and record the gender coefficient. The randomization P-value is the proportion of random assignments that yield an estimate of the gender coefficient greater than or equal to the estimate computed from the original data.

The randomization test can be performed with or without stratification by department, i.e., it can fix the number of female presenters in each department or only fix the total across departments. Because the stratified randomization test respects the number of male and female presenters in each department, it is tied more closely to the underlying data. The unstratified and stratified randomization P-values are in Tables A.1 and A.2. The randomization P-values generally are above 0.1; some are as large as 0.96.

Table A.1 gives the results corresponding to Table 4 of Blair-Loy et al. [2017] for our data. The parametric P-value for the coefficient of gender in the ZINB positive model for attempted interruptions is 0.002. If we calibrate the P-value using randomization rather than relying on the (false) parametric assumptions, the presenter gender coefficient is not significantly different from zero at level 5% in any of the models, after adjusting for multiplicity.

Response Variable	Parametric P-value	Unstratified Randomization P-value	Stratified Randomization P-value
Attempted Interruption	< 0.01	0.05	0.04
Acknowledged Question	0.93	0.94	0.95
Unacknowledged Question	0.76	0.76	0.79
Follow-up Question	0.19	0.18	0.18
Total Questions	0.60	0.56	0.56

Table A.1: P-values from randomization test using the ZINB positive model gender coefficient as test statistic. Each response variable was regressed on the presenter gender indicator variable and the proportion female faculty in that department in the ZINB model.

We also attempted to replicate Table 6 from Blair-Loy et al. [2017]. It includes the number of years since the presenter earned a PhD as a covariate in the ZINB model. Unadjusted P-values are reported in Table A.2. The smallest randomization P-value is slightly above 0.05. If the five tests were adjusted for multiplicity, the resulting P-values would be above 0.1.

Response Variable	Parametric P-value	Unstratified Randomization P-value	Stratified Randomization P-value
Attempted Interruption	< 0.01	0.07	0.05
Acknowledged Question	0.75	0.93	0.94
Unacknowledged Question	0.96	0.79	0.75
Follow-up Question	0.24	0.19	0.20
Total Questions	0.67	0.57	0.57

Table A.2: P-values from permutation test using the coefficient of presenter gender in the ZINB positive model as the test statistic. Each response variable was regressed on the presenter gender indicator variable, the proportion of female faculty in that department, and the number of years since the presenter earned a PhD in the ZINB model.

A.2.4 Testing ZINB: Negative Controls

As a further illustration that the parametric assumptions of the ZINB model may produce misleading conclusions, we use the ZINB model to estimate the effect of a variable that should not matter: whether the presenter's first name has an even or odd number of letters.

The parametric P-value associated with this variable in the ZINB positive model is less than 0.05 for two of the response variables, the number of unacknowledged questions and the number of follow up questions. On the other hand, the randomization P-values (for the same test statistic) are above 0.05 for all response variables. Results are in Table A.3.

Response Variable	Parametric P-value	Unstratified Randomization P-value	Stratified Randomization P-value
Attempted Interruption	0.36	0.36	0.34
Acknowledged Question	0.45	0.46	0.46
Unacknowledged Question	0.05	0.10	0.11
Follow-up Question	0.04	0.06	0.05
Total Questions	0.13	0.13	0.14

Table A.3: P-values from permutation test using the coefficient of presenter gender in the ZINB positive model as the test statistic. Each response variable was regressed on the indicator variable of whether the presenter's first name has an even or odd number of letters.

A.2.5 ZINB versus randomization tests

The randomization tests posit that gender is an arbitrary label, which might as well have been assigned at random. They make no assumption about the distribution of the number and nature of questions; they do not even assume those things are random. Indeed, they "condition" on the number of questions of each type received by each presenter.

In contrast, the ZINB model assumes that questions are generated in the following way. First, toss a biased coin. If the coin lands heads, then the presenter receives no questions (in the pre-Q&A portion of the talk). If the coin lands tails, toss a different coin repeatedly, independently, until it lands heads some pre-specified number of times. The number of tosses

it takes to get the pre-specified number of heads is the number of questions asked in the pre-Q&A portion of the talk. The parameters in the models are the chance of heads for the first coin, chance of heads for the second coin, and the pre-specified number of heads for the second coin. These are the "natural" parameters for the negative binomial, but it is common to re-parametrize the distribution in terms of the mean and scaled standard deviation.

These parameters are in turn modeled as parametric functions of a pre-specified set of covariates, such as gender, proportion of female faculty, and years since the presenter earned a PhD.

The scientific research question is, "Are women interrupted more than men?" The ZINB analysis changes that question to "On the assumption that the number of questions was generated by a ZINB model with a specified parametric relationship to a given set of covariates, is the coefficient of "female" in the ZINB positive model zero?"

Focusing solely on the "positive model" (i.e., the distribution of pre-Q&A questions given that at least one question was asked) widens the gap between the scientific question and the statistical question. Suppose, for example, that there are 50 female and 50 male presenters. Every man receives 5 questions. One woman receives 50 questions and the others receive none. The only woman who contributes data to the "positive model" is the woman who received 50 questions, but all the men contribute data. The positive model would show that women receive more questions than men, even though on average they receive fewer.

A.3 Supplemental Results

Additional results are shown below. Table 7 gives the results for all presenters (nontenured and tenured) for only the pre-Q&A portion of the talk. Table 8 gives the results for only nontenured presenters for only the pre-Q&A portion of the talk. Table 9 gives the results for only nontenured presenters for the entire talk (pre and post Q&A). None of the omnibus tests are significant (P-value of 1).

	CEE	EECS	IEOR	ME	Physics
Time on Questions (in seconds)	0	14	-65	7	-4
	1	*0.25*	*0.65*	*0.29*	*0.84*
Acknowledged Question	0	0	1	1	0
	1	*0.65*	*0.51*	*0.43*	*0.96*
Unacknowledged Question	0	-1	-16	1	0
	1	*0.93*	*1*	*0.26*	*0.68*
Attempted Interruption	0	0	-4	0	0
	1	*1*	*0.95*	*1*	*1*
Follow-up Question	0	0	-11	0	0
	1	*0.56*	*0.87*	*0.54*	*0.96*
Scientific Comment	0	0	-2	0	0
	1	*1*	*0.40*	*1*	*1*
Non Scientific Comment	0	0	-1	0	0
	1	*1*	*0.69*	*1*	*1*
Positive Comment	0	0	0	0	0
	1	*1*	*1*	*1*	*1*
Clarifying Question	0	1	-9	-2	1
	1	*0.84*	*0.72*	*0.79*	*1*
Furthering Question	0	0	2	0	0
	1	*1*	*1*	*1*	*1*
Critical Element	0	0	-1	0	0
	1	*1*	*0.38*	*1*	*1*
Ad Hominem	0	0	0	0	0
	1	*1*	*1*	*1*	*1*
Self Referential	0	0	0	0	0
	1	*1*	*1*	*1*	*1*

Table A.4: For each department, difference in the median (female - male) for each category with P-value of permutation test in parentheses. Statistics based on **all** presenters and the **pre Q&A data only**.

	CEE	EECS	IEOR	ME	Physics
Time on Questions (in seconds)	0	8	-65	2	-4
	1	*0.41*	*0.65*	*0.41*	*0.84*
Acknowledged Question	0	-1	1	0	0
	1	*0.88*	*0.51*	*0.58*	*0.96*
Unacknowledged Question	0	-1	-16	1	0
	1	*0.92*	*1*	*0.47*	*0.68*
Attempted Interruption	0	0	-4	0	0
	1	*1*	*0.95*	*1*	*1*
Follow-up Question	0	0	-11	0	0
	1	*0.69*	*0.87*	*0.60*	*0.96*
Scientific Comment	0	0	-2	0	0
	1	*1*	*0.40*	*1*	*1*
Non Scientific Comment	0	0	-1	0	0
	1	*1*	*0.69*	*1*	*1*
Positive Comment	0	0	0	0	0
	1	*1*	*1*	*1*	*1*
Clarifying Question	0	0	-9	-2	1
	1	*1*	*0.72*	*0.65*	*1*
Furthering Question	0	0	2	0	0
	1	*1*	*1*	*1*	*1*
Critical Element	0	0	-1	0	0
	1	*1*	*0.38*	*1*	*1*
Ad Hominem	0	0	0	0	0
	1	*1*	*1*	*1*	*1*
Self Referential	0	0	0	0	0
	1	*1*	*1*	*1*	*1*

Table A.5: For each department, difference in medians (female - male) for each category with P-value of permutation test in parentheses. Statistics based on **non-tenured** presenters only and the **pre Q&A data only**.

	CEE	EECS	IEOR	ME	Physics
Time on Questions (in seconds)	-6	39	117	-13	0
	0.48	*0.11*	*0.52*	*0.55*	*0.64*
Acknowledged Question	1	1	7	-6	-2
	0.27	*0.34*	*0.37*	*0.97*	*1*
Unacknowledged Question	0	0	-7	2	1
	0.75	*0.66*	*0.65*	*0.22*	*0.35*
Attempted Interruption	0	0	-2	0	0
	0.94	*0.99*	*0.65*	*0.79*	*0.67*
Follow-up Question	1	1	-1	-2	-1
	0.41	*0.45*	*0.65*	*0.78*	*0.77*
Scientific Comment	-2	0	-3	2	0
	0.15	*1*	*0.57*	*0.14*	*1*
Non-scientific Comment	0	0	-2	1	0
	1	*1*	*0.59*	*0.99*	*1*
Positive Comment	0	0	0	0	0
	1	*1*	*1*	*1*	*1*
Clarifying Question	2	2	-5	-5	-3
	0.42	*0.42*	*0.72*	*0.42*	*0.32*
Furthering Question	1	-1	-5	-1	2
	0.56	*0.9*	*0.72*	*1*	*0.63*
Critical Element	0	0	-1	1	-1
	1	*1*	*1*	*0.97*	*0.65*
Ad Hominem	0	0	0	0	0
	1	*1*	*1*	*1*	*1*
Self Referential	0	0	0	1	0
	1	*1*	*1*	*0.73*	*1*

Table A.6: For each department, difference (female - male) in the median for each category with P-value of permutation test in parentheses. Statistics based on **non-tenured** presenters only and the **entire talk**.

Appendix B

Appendix for Chapter 8

B.1 Proof of Theorem

B.1.1 Assumptions

Our proof relies on the assumptions on sampling described in Section 8.2.1 and the Group-Match identification assumptions described in Section 8.2.2 of the main manuscript, with the exception of the exact matching assumption. In addition, we invoke three additional sets of assumptions, which we denote M, W, R following similar labeling in Otsu and Rai Otsu and Rai [2017], from whom we adapt our proof strategy.

Assumption M. (Conditions for $\hat{\Delta}_{adj}$)

1. Let the population distribution of $\mathbf{X}_i = (X_{i,1}, \ldots, X_{i,T})$ be continuous on $\mathbb{R}^{kT}$ with compact and convex support $\mathbb{X}^T$. In addition, let the density of $\mathbf{X}_i$ be bounded, and bounded away from zero on its support.

2. For some $r \geq 1$, $\frac{N_1^r}{N_0} \to \theta$ for $\theta \in (0, \infty)$.

3. For $z = 0, 1$, let $\mu_z^t(\mathbf{X})$ be Lipschitz in $\mathbb{X}^L$ for all $t = L+1, \ldots, T$.

4. For all $t \in \{L+1, \ldots, T\}$, and $d = 0, 1$, $E[Y_{i,t}^4(z)|D_i = d, \mathbf{X}_i]$ and $Cov(Y_{it}, Y_{it'}|D_i = d, \mathbf{X}_{i,t} = x, \mathbf{X}_{i,t'} = x')$ are bounded uniformly on $\mathbb{X}^T$, and $\text{Var}(Y_{i,t}(z)||D_i = d, \mathbf{X}_i)$ is Lipschitz in $\mathbb{X}^T$ and bounded away from zero.

Assumption R. (Conditions for $\mu_d(x)$)
For $d = 0, 1$ and λ satisfying $\sum_{l=1}^{kL} \lambda_l = kL$, the derivative $\partial^{kL}\mu_d^t(x)$ exists and satisfies $\sup_{x \in \mathbb{X}} |\partial^{kL}\mu_d^t(x)| \leq R$ for some $R > 0$ and for all $t \in \{L+1, \ldots T\}$. Furthermore, $\hat{\mu}_0(x)$ satisfies $|\hat{\mu}_0(\cdot) - \mu_0(\cdot)|_{kL-1} = o_p(N^{-1/2+1/(kL)})$.

Assumption W. (Conditions for W_i^*)

1. $(W_1^*, ..., W_N^*)$ is exchangeable and independent of $\mathbf{O} = (\mathbf{Y}, \mathbf{D}, \mathbf{X})$.

2. $\sum_{i=1}^N (W_i^* - \bar{W}^*)^2 \xrightarrow{p} 1$ where $\bar{W}^* = \frac{1}{N}\sum_{i=1}^N W_i^*$

3. $\max_{i=1,\ldots,N} |W_i^* - \bar{W}^*| \xrightarrow{p} 0$

4. $E[W_i^{*2}] = O(N^{-1})$ for all $i = 1, ..., N$

Assumption W is satisfied by construction for the nonparametric bootstrap, wild bootstrap, and Bayesian bootstrap approaches mentioned in the main text.

B.1.2 Lemmas

To bound the size of matching discrepancies, we compare those obtained by GroupMatch with instance replacement to nearest-neighbor matching at a fixed timepoint, in which only one instance from each control unit can potentially be used in a match. Nearest-neighbor matching matches each treated unit to the control instance that is most similar. While GroupMatch also uses nearest-neighbor matching, nearest-neighbor matching at a fixed timepoint considers a smaller pool of control instances since there is only one instance for each control unit that can potentially be used in a match. Comparing matching discrepancies between GroupMatch with instance replacement and nearest-neighbor matching is important to apply Lemma A.2, used to prove Theorem 2, in Abadie and Imbens Abadie and Imbens [2011]. For each treated unit i, let $j_m(i)$ and $j_m^{gm}(i)$ represent the mth instance used as a match for nearest-neighbors at a fixed timepoint and GroupMatch with instance replacement respectively. Let $U_{m,i} = \mathbf{X}_{j_m(i)} - \mathbf{X}_{i,T_i}$ and $U_{m,i}^{gm} = \mathbf{X}_{j_m^{gm}(i)} - \mathbf{X}_{i,T_i}$ be the matching discrepancies under nearest neighbors at a fixed timepoint and GroupMatch with instance replacement matching respectively.

Lemma 1. $||U_{m,i}^{gm}|| \leq ||U_{m,i}||$ for all m, i.

Proof. Because nearest neighbors (NN) matching at a fixed timepoint only considers one instance from each control trajectory whereas GroupMatch with instance replacement (GM) considers multiple, the set of control instances that can be used in a match in NN matching, $\mathcal{C}$, is a subset of the set of control instances that can be used in a match in GM, $\mathcal{C}^{gm}$: $\mathcal{C} \subseteq \mathcal{C}^{gm}$. Both NN and GM match treated units to the control instance that minimizes $U_{m,i}$ and $U_{m,i}^{gm}$, so we have that

$$||U_{m,i}^{gm}|| = min_{m,j \in \mathcal{C}^{gm}} |\mathbf{X}_j - \mathbf{X}_{i,T_i}| \leq min_{m,j \in \mathcal{C}} |\mathbf{X}_j - \mathbf{X}_{i,T_i}| = ||U_{m,i}||$$

where min_m denotes the mth minimum.

Lemma 2. $E[K_M(i,t)^q]$ is bounded uniformly in N.

Proof. This proof follows closely with the proof of Lemma 3 in Abadie and Imbens Abadie and Imbens [2006] with modifications described below. Let f^t be the density of $\mathbf{X}_{i,t}$ and define $\underline{f} = inf_{x,w,t} f_w^t(x)$ and $\overline{f} = sup_{x,w,t} f_w^t(x)$. We define the catchment area, $\mathbb{A}_M(i,t)$ as the subset of $\mathbb{X}$ such that control unit i at time t is matched to each observation j at time t' with $W_j = 1 - W_i$ and $\mathbf{X}_{j,t'} \in \mathbb{A}_M(i,t)$:

$$\mathbb{A}_M(i,t) = \{x | \sum_{l|W_l = W_i, l \neq i} 1\{(min_{t'} ||\mathbf{X}_{l,t'} - x||) \leq ||\mathbf{X}_{i,t} - x||\} 1\{min_{t'} ||\mathbf{X}_{i,t'} - x|| \geq ||\mathbf{X}_{i,t} - x||\} \leq M\}$$

Ultimately, we want to bound the volume of the catchment area. In order to do this, we need to bound the probability that the distance to a match exceeds some value. To derive this bound, we must account for our trajectory structure by showing the following inequality:

$$
\begin{aligned}
&Pr(||\mathbf{X}_{j,t} - \mathbf{X}_{i,t'}|| > uN_{1-W_i}^{-1/k} | W_1, ..., W_N, \mathbf{X}_{i,t'} = x, j \in \mathcal{J}_M(i)) \\
&\leq Pr(||\mathbf{X}_{j,t} - \mathbf{X}_{i,t'}|| > uN_{1-W_i}^{-1/k} | W_1, ..., W_N, \mathbf{X}_{i,t'} = x, (j,t) = j_M(i,t')) \\
&= \sum_{m=0}^{M-1} \binom{N_{1-W_i}}{m} Pr(min_t ||\mathbf{X}_{j,t} - \mathbf{X}_{i,t'}|| > uN_{1-W_i}^{-1/k} | W_1, ..., W_N, W_j = 1 - W_i, \mathbf{X}_{i,t'} = x)^{N_{1-W_i}-m} \times \\
&Pr(min_t ||\mathbf{X}_{j,t} - \mathbf{X}_{i,t'}|| \leq uN_{1-W_i}^{-1/k} | W_1, ..., W_N, W_j = 1 - W_i, \mathbf{X}_{i,t'} = x)^m \\
&\leq \sum_{m=0}^{M-1} \binom{N_{1-W_i}}{m} Pr(||\mathbf{X}_{j,t} - \mathbf{X}_{i,t'}|| > uN_{1-W_i}^{-1/k} | W_1, ..., W_N, W_j = 1 - W_i, \mathbf{X}_{i,t'} = x)^{N_{1-W_i}-m} \times \\
&Pr(||\mathbf{X}_{j,t} - \mathbf{X}_{i,t'}|| \leq uN_{1-W_i}^{-1/k} | W_1, ..., W_N, W_j = 1 - W_i, \mathbf{X}_{i,t'} = x)^m
\end{aligned}
$$

The second inequality follows from the fact that the probability that the minimal distance over a trajectory is less than or equal to any particular instance. The rest of the proof follows directly from the proof of Abadie and Imbens Abadie and Imbens [2006]'s Lemma 3 after substituting in our catchment area, this inequality, and indexing over time.

Lemma 3. Given that $E[K_M(i,t)^q]$ is uniformly bounded, $\sum_i \sum_t K_M(i,t) = N_1$, and $K_M(i,t) \geq 0$, we have that $\sum_i \sum_t \sum_{t'} \sum_{t''} \sum_{t'''} E[K_M(i,t)K_M(i,t')K_M(i,t'')K_M(i,t''')] \leq cN_1^{1+o(1)}$ for some constant c.

Proof. For all $\epsilon > 0$, we have that

$$
\begin{aligned}
&\sum_i \sum_t \sum_{t'} \sum_{t''} \sum_{t'''} E[K_M(i,t)K_M(i,t')K_M(i,t'')K_M(i,t''')] \\
&= \sum_i \sum_t \sum_{t'} \sum_{t''} \sum_{t'''} E[K_M(i,t)K_M(i,t')K_M(i,t'')K_M(i,t'''); K_M(i,t), K_M(i,t'), K_M(i,t''), K_M(i,t''') \leq N_1^\epsilon] + \\
&\sum_i \sum_t \sum_{t'} \sum_{t''} \sum_{t'''} E[K_M(i,t)K_M(i,t')K_M(i,t'')K_M(i,t'''); max_{t_1=t,t',t'',t'''} K_M(i,t_1) > N_1^\epsilon] \\
&\leq \sum_i \sum_t \sum_{t'} \sum_{t''} \sum_{t'''} E[K_M(i,t)K_M(i,t')K_M(i,t'')K_M(i,t'''); K_M(i,t), K_M(i,t'), K_M(i,t''), K_M(i,t''') \leq N_1^\epsilon] + \\
&4 \sum_i \sum_t \sum_{t'} \sum_{t''} \sum_{t'''} E[K_M(i,t)K_M(i,t')K_M(i,t'')K_M(i,t'''); K_M(i,t_1) > N_1^\epsilon]
\end{aligned}
$$

This upper bound is derived from the fact that

$$
\begin{aligned}
&E[K_M(i,t)K_M(i,t')K_M(i,t'')K_M(i,t'''); max_{t_1=t,t',t'',t'''} K_M(i,t_1) > N_1^\epsilon] \\
&\leq \sum_{t_1 \in \{t,t',t'',t'''\}} E[K_M(i,t)K_M(i,t')K_M(i,t'')K_M(i,t'''); K_M(i,t_1) > N_1^\epsilon]
\end{aligned}
$$

and since we are summing over all t, t', t'', t''' we are able to replace the summation above with multiplying by four. We bound each of the terms separately. First,

$$\sum_i \sum_t \sum_{t'} \sum_{t''} \sum_{t'''} E[K_M(i,t)K_M(i,t')K_M(i,t'')K_M(i,t'''); K_M(i,t), K_M(i,t'), K_M(i,t''), K_M(i,t''') \leq N_1^\epsilon]$$
$$\leq N_1^{3\epsilon} \sum_i \sum_t \sum_{t'} \sum_{t''} \sum_{t'''} E[K_M(i,t)]$$
$$= T^3 N_1^{3\epsilon+1}$$

For the next term we use proof by contradiction to show

$$\sum_i \sum_t \sum_{t'} \sum_{t''} \sum_{t'''} E[K_M(i,t)K_M(i,t')K_M(i,t'')K_M(i,t'''); K_M(i,t) > N_1^\epsilon] \leq N_1.$$

Suppose $\sum_i \sum_t \sum_{t'} \sum_{t''} \sum_{t'''} E[K_M(i,t)K_M(i,t')K_M(i,t'')K_M(i,t'''); K_M(i,t) > N_1^\epsilon] > N_1$, then:

$$\sum_i \sum_t \sum_{t'} \sum_{t''} \sum_{t'''} E[K_M(i,t)^{\frac{r}{\epsilon}+1} K_M(i,t')K_M(i,t'')K_M(i,t'''); K_M(i,t) > N_1^\epsilon]$$
$$\geq (N_1^\epsilon)^{\frac{r}{\epsilon}} \sum_i \sum_t \sum_{t'} \sum_{t''} \sum_{t'''} E[K_M(i,t)K_M(i,t')K_M(i,t'')K_M(i,t'''); K_M(i,t) > N_1^\epsilon]$$
$$= N_1^r \sum_i \sum_t \sum_{t'} \sum_{t''} \sum_{t'''} E[K_M(i,t)K_M(i,t')K_M(i,t'')K_M(i,t'''); K_M(i,t) > N_1^\epsilon]$$
$$> N_1^r N_1 = N_1^{r+1}$$
$$\implies NT^4 sup_{i,t,t',t'',t'''} E[K_M(i,t)^{\frac{r}{\epsilon}+1} K_M(i,t')K_M(i,t'')K_M(i,t'''); K_M(i,t) > N_1^\epsilon] > N_1^{r+1}$$
$$\implies sup_{i,t,t',t'',t'''} E[K_M(i,t)^{\frac{r}{\epsilon}+1} K_M(i,t')K_M(i,t'')K_M(i,t'''); K_M(i,t) > N_1^\epsilon] > \frac{N_1^{r+1}}{NT^4} = cN_1$$

for some constant c. This follows from Assumption M.
So, then we have that:

$$sup_{i,t} E[K_M(i,t)^{\frac{r}{\epsilon}+4}]$$
$$\geq sup_{i,t,t',t'',t'''} E[K_M(i,t)^{\frac{r}{\epsilon}+1} K_M(i,t')K_M(i,t'')K_M(i,t''')]$$
$$> sup_{i,t,t',t'',t'''} E[K_M(i,t)^{\frac{r}{\epsilon}+1} K_M(i,t')K_M(i,t'')K_M(i,t'''); K_M(i,t) > N_1^\epsilon]$$
$$> cN_1$$

The first inequality holds by applying Cauchy-Schwarz twice. But then $E[K_M(i,t)^q]$ is not uniformly bounded for all q, so by contradiction

$$\sum_i \sum_t \sum_{t'} \sum_{t''} \sum_{t'''} E[K_M(i,t)K_M(i,t')K_M(i,t'')K_M(i,t'''); K_M(i,t) > N_1^\epsilon] \leq N_1.$$

So we have that:

$$\sum_i \sum_t \sum_{t'} \sum_{t''} \sum_{t'''} E[K_M(i,t)K_M(i,t')K_M(i,t'')K_M(i,t''')]$$
$$= \sum_i \sum_t \sum_{t'} \sum_{t''} \sum_{t'''} E[K_M(i,t)K_M(i,t')K_M(i,t'')K_M(i,t'''); K_M(i,t), K_M(i,t'), K_M(i,t''), K_M(i,t''') \leq N_1^\epsilon] +$$
$$\sum_i \sum_t \sum_{t'} \sum_{t''} \sum_{t'''} E[K_M(i,t)K_M(i,t')K_M(i,t'')K_M(i,t'''); max_{t_1=t,t',t'',t'''} K_M(i,t_1) > N_1^\epsilon]$$
$$\leq \sum_i \sum_t \sum_{t'} \sum_{t''} \sum_{t'''} E[K_M(i,t)K_M(i,t')K_M(i,t'')K_M(i,t'''); K_M(i,t), K_M(i,t'), K_M(i,t''), K_M(i,t''') \leq N_1^\epsilon] +$$
$$4 \sum_i \sum_t \sum_{t'} \sum_{t''} \sum_{t'''} E[K_M(i,t)K_M(i,t')K_M(i,t'')K_M(i,t'''); K_M(i,t) > N_1^\epsilon]$$
$$\leq T^3 N_1^{3\epsilon+1} + 4c'N_1 \leq cN_1^{1+o(1)}$$

for some constant c.

B.1.3 Proof

We follow the proof of Theorem 1 in Otsu and Rai Otsu and Rai [2017] closely, adapting where necessary to address the possible presence of multiple correlated potential outcomes from the same trajectory multiple control instances in the matched design. In the case where every control unit has only one instance, our argument reduces to exactly that presented in Otsu and Rai Otsu and Rai [2017].

We decompose $\sqrt{N_1}U^*$ as follows:

$$\begin{aligned}
\sqrt{N_1}U^* &= \sum_{i=1}^{N} W_i^*(\hat{\Delta}_i - D_i\hat{\Delta}_{adj}) \\
&= \sum_{i=1}^{N} (W_i^* - \bar{W}^*)(\hat{\Delta}_i - D_i\hat{\Delta}_{adj}) \\
&= \sum_{i=1}^{N} (W_i^* - \bar{W}^*)(D_i(\hat{\Delta}_i - \hat{\Delta}_{adj}) + (1-D_i)\hat{\Delta}_i) \\
&= \sum_{i=1}^{N} (W_i^* - \bar{W}^*)[D_i(Y_{i,T_i} - \hat{\mu}_0(\mathbf{X}_{i,T_i}) - \hat{\Delta}_{adj}) + (1-D_i)\sum_{t=1}^{T} \frac{K_M(i,t)}{C}(Y_{i,t} - \hat{\mu}_0(\mathbf{X}_{i,t}))] \\
&= \sqrt{N_1}(T^* + R_{1N_1}^* + R_{2N_1}^*)
\end{aligned}$$

We define the following:

$$\begin{aligned}
e_{i,t} &= Y_{i,t} - \mu_{D_i}(\mathbf{X}_{i,t}) \\
\xi_{i,t} &= (2D_i - 1)(\mu_{D_i}(\mathbf{X}_{i,t}) - \mu_{1-D_i}(\mathbf{X}_{i,t})) - \Delta
\end{aligned}$$

We can now rewrite the three components as follows.

$$\begin{aligned}\sqrt{N_1}T^* &= \sum_{i=1}^{N}(W_i^* - \bar{W}^*)(D_i(e_{i,t=T_i} + \xi_{i,t=T_i}) - (1-D_i)\sum_{t=1}^{T}\frac{K_M(i,t)}{C}e_{i,t}) \\ &= \sum_{i=1}^{N}(W_i^* - \bar{W}^*)[D_i((Y_{i,t=T_i} - \mu_1^{T_i}(\mathbf{X}_{i,t})) + (\mu_1^{T_i}(\mathbf{X}_{i,t}) - \mu_0(\mathbf{X}_{i,t}) - \Delta)) \\ &- (1-D_i)\sum_{t=1}^{T}\frac{K_M(i,t)}{C}(Y_{i,t} - \mu_0(\mathbf{X}_{i,t}))] \\ &= \sum_{i=1}^{N}(W_i^* - \bar{W}^*)[D_i(Y_{i,t=T_i} - \mu_0(\mathbf{X}_{i,t}) - \Delta) - (1-D_i)\sum_{t=1}^{T}\frac{K_M(i,t)}{C}(Y_{i,t} - \mu_0(\mathbf{X}_{i,t}))]\end{aligned}$$

$$\sqrt{N_1}R_{1N_1}^* = \sum_{i=1}^{N}(W_i^* - \bar{W}^*)(D_i(\mu_0(\mathbf{X}_{i,t}) - \hat{\mu}_0(\mathbf{X}_{i,t})) - (1-D_i)\frac{K_M(i,t)}{C}\sum_{t=1}^{T}(\mu_0(\mathbf{X}_{i,t}) - \hat{\mu}_0(\mathbf{X}_{i,t})))$$

$$\sqrt{N_1}R_{2N_1}^* = \sum_{i=1}^{N}(W_i^* - \bar{W}^*)D_i(\Delta - \hat{\Delta}_{adj})$$

We have that $Pr\{\sqrt{N_1}R_{1N_1}^* > \epsilon|\mathbf{O}\} \xrightarrow{p} 0$ and $Pr\{\sqrt{N_1}R_{2N_1}^* > \epsilon|\mathbf{O}\} \xrightarrow{p} 0$ for any $\epsilon > 0$, by the same argument as Otsu and Rai Otsu and Rai [2017] which utilizes our Assumptions W, R, Lemma 2 and the Markov Inequality. This part of the argument also relies on Lemma 1, although the reliance is not explicit in Otsu and Rai Otsu and Rai [2017]; see Abadie and Imbens Abadie and Imbens [2011] for a similar derivation where the role of Lemma 1 is more clear.

Next, to show that that $\sup_r|Pr\{\sqrt{N_1}T^* \leq r|\mathbf{O}\} - Pr\{\sqrt{N_1}(\hat{\Delta}_{adj} - \Delta) \leq r\}| \xrightarrow{p} 0$, we define:

$$\begin{aligned}\eta_i &= [D_i(Y_{i,t=T_i} - \mu_0(\mathbf{X}_{i,t}) - \Delta) - (1-D_i)\sum_{t=1}^{T}\frac{K_M(i,t)}{C}(Y_{i,t} - \mu_0(\mathbf{X}_{i,t}))]/\sqrt{N_1} \\ &= D_i(e_{i,t=T_i} + \xi_{i,t=T_i}) - (1-D_i)\sum_{t=1}^{T}\frac{K_M(i,t)}{C}e_{i,t}\end{aligned}$$

We have the following:

$$\sigma_N^2 = \sigma_{1N}^2 + \sigma_2^2$$

$$\sigma_{1N}^2 = \frac{1}{N_1}\sum_{i=1}^{N} Var(\sum_{t=1}^{T}(D_i 1\{t = T_i\} + \frac{K_M(i,t)}{C}(1 - D_i))Y_{it}|\mathbf{D}, \mathbf{X})$$

$$= \frac{1}{N_1}\sum_{i=1}^{N} Cov\Big\{\sum_{t=1}^{T}(D_i 1\{t = T_i\} + \frac{K_M(i,t)}{C}(1 - D_i))Y_{it},$$

$$\sum_{t=1}^{T}(D_i 1\{t = T_i\} + \frac{K_M(i,t)}{C}(1 - D_i))Y_{it}|\mathbf{D}, \mathbf{X}\Big\}$$

$$= \frac{1}{N_1}\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{t'=1}^{T}(D_i 1\{t = T_i\} + \frac{K_M(i,t)}{C}(1 - D_i))(D_i 1\{t' = T_i\} + \frac{K_M(i,t')}{C}(1 - D_i))Cov(Y_{it}, Y_{it'})$$

$$\sigma_2^2 = E[((\mu_1^{T_i}(\mathbf{X}_{i,t}) - \mu_0(\mathbf{X}_{i,t})) - \Delta)^2 | D_i = 1]$$

Within σ_{1N}^2, note that $Var(\sum_{t=1}^{T}(D_i 1\{t = T_i\} + \frac{K_M(i,t)}{C}(1 - D_i))Y_{it}|\mathbf{D}, \mathbf{X})$ reduces to $Var(Y_{it})$ for treated units. However, for control units, we have this variance term plus extra covariance terms between that control unit instance and the other instances in its trajectory. From here, we are able to follow the same proof strategy as in Otsu and Rai Otsu and Rai [2017], with our modified assumptions, and some modifications (detailed below) to Lemmas (i)-(iii) in Otsu and Rai Otsu and Rai [2017] to account for the extra covariance terms resulting from the trajectory structure of the control units. Lemmas (i)-(iii) in Otsu and Rai Otsu and Rai [2017] show that the sampling variance of the η_i's converge to the population variance, σ_N^2, and that other random variables converge in probability to 0. They show this by leveraging the boundedness of higher order moments of η_i and use of the Markov inequality. We are able to utilize the same arguments, with changes laid out more explicitly for lemma (i), below, and generalized for lemmas (ii) and (iii).

To apply Lemmas (i)-(iii) in Otsu and Rai Otsu and Rai [2017] we define:

$$\hat{\sigma}_{1N}^2 = \frac{1}{N_1}\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{t'=1}^{T}(D_i 1\{t = T_i\} + \frac{K_M(i,t)}{C}(1 - D_i))(D_i 1\{t' = T_i\} + \frac{K_M(i,t')}{C}(1 - D_i))e_{i,t}e_{i,t'}$$

Then note that:

$$
\begin{aligned}
&E[(\hat{\sigma}^2_{1N} - \sigma^2_{1N})^2] \\
&= E[\{(\frac{1}{N_1}\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{t'=1}^{T}(D_i 1\{t = T_i\} + \frac{K_M(i,t)}{C}(1-D_i))(D_i 1\{t' = T_i\} + \frac{K_M(i,t')}{C}(1-D_i))e_{i,t}e_{i,t'}) - \\
&(\frac{1}{N_1}\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{t'=1}^{T}(D_i 1\{t = T_i\} + \frac{K_M(i,t)}{C}(1-D_i))(D_i 1\{t' = T_i\} + \frac{K_M(i,t')}{C}(1-D_i))Cov(Y_{it}, Y_{it'}))\}^2] \\
&= \frac{1}{N_1^2}E[\{\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{t'=1}^{T}(D_i 1\{t = T_i\} + \frac{K_M(i,t)}{C}(1-D_i))(D_i 1\{t' = T_i\} + \frac{K_M(i,t')}{C}(1-D_i)) \\
&\times (e_{i,t}e_{i,t'} - Cov(Y_{it}, Y_{it'}))\}^2] \\
&= \frac{1}{N_1^2}E[\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{t'=1}^{T}\sum_{j=1}^{N}\sum_{t_1=1}^{T}\sum_{t_1'=1}^{T}(D_i 1\{t = T_i\} + \frac{K_M(i,t)}{C}(1-D_i))(D_i 1\{t' = T_i\} + \frac{K_M(i,t')}{C}(1-D_i)) \\
&\times (D_j 1\{t_1 = T_j\} + \frac{K_M(j,t_1)}{C}(1-D_j))(D_j 1\{t_1' = T_j\} + \frac{K_M(j,t_1')}{C}(1-D_j)) \\
&\times (e_{i,t}e_{i,t'} - Cov(Y_{it}, Y_{it'}))(e_{j,t_1}e_{j,t_1'} - Cov(Y_{jt_1}, Y_{jt_1'}))] \\
&= \frac{1}{N_1^2}\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{t'=1}^{T}\sum_{t_1=1}^{T}\sum_{t_1'=1}^{T}E[(D_i 1\{t = T_i\} + \frac{K_M(i,t)}{C}(1-D_i))(D_i 1\{t' = T_i\} + \frac{K_M(i,t')}{C}(1-D_i)) \\
&\times (D_i 1\{t_1 = T_i\} + \frac{K_M(i,t_1)}{C}(1-D_i))(D_i 1\{t_1' = T_i\} + \frac{K_M(i,t_1')}{C}(1-D_i)) \\
&\times E[(e_{i,t}e_{i,t'} - Cov(Y_{it}, Y_{it'}))(e_{i,t_1}e_{i,t_1'} - Cov(Y_{it_1}, Y_{it_1'}))|D_i, \mathbf{X}_{i,t}, \mathbf{X}_{i,t'}, \mathbf{X}_{i,t_1}, \mathbf{X}_{i,t_1'}]] \\
&\leq \frac{1}{N_1^2}(N_1 + cN_1^{1+o(1)})\times \\
&3 \times max(sup_{d,x}E[e_{it}^4|D_i = d, \mathbf{X}_{i,t} = x], \\
&- 2sup_{d,x,x'}E[e_{it}^2|D_i = d, \mathbf{X}_{i,t} = x]sup_{d,x,x'}Cov(Y_{it}, Y_{it'}|D_i = d, \mathbf{X}_{i,t} = x, \mathbf{X}_{i,t'} = x'), \\
&sup_{d,x,x'}Cov(Y_{it}, Y_{it'}|D_i = d, \mathbf{X}_{i,t} = x, \mathbf{X}_{i,t'} = x')^2) \to 0
\end{aligned}
$$

The inequality follows from Lemma 3. The convergence follows from Assumption M(4) and Lemma 2.

Difference-in-differences Estimator

This proof extends easily to the difference-in-differences ATT estimator described in Section 8.4.3 with the modifications described below.

We replace $\hat{\Delta}_{adj}$ with $\hat{\Delta}_{DiD}$ and the outcome $Y_{i,t}$ with the difference-in-differences outcome $Y_{i,t} - Y_{i,t-1}$

We modify

$$e_{i,t} = Y_{i,t} - Y_{i,t-1} - (\mu_{D_i}(\mathbf{X}_{i,t}) - \mu_{D_i}(\mathbf{X}_{i,t-1}))$$

and

$$\xi_{i,t} = (2D_i - 1)((\mu_{D_i}(\mathbf{X}_{i,t}) - \mu_{1-D_i}(\mathbf{X}_{i,t})) - (\mu_{D_i}(\mathbf{X}_{i,t-1}) - \mu_{1-D_i}(\mathbf{X}_{i,t-1})))$$

.

We also update our variance formulas to reflect the additional covariance terms introduced by the difference-in-differences outcome. In particular, our formula for σ^2_{1N} includes additional terms:

$$\begin{aligned}
\sigma^2_{1N} &= \frac{1}{N_1}\sum_{i=1}^{N} Var(\sum_{t=1}^{T}(D_i 1\{t = T_i\} + \frac{K_M(i,t)}{C}(1 - D_i))(Y_{it} - Y_{i,t-1})|\mathbf{D}, \mathbf{X}) \\
&= \frac{1}{N_1}\sum_{i=1}^{N} Cov\Big\{\sum_{t=1}^{T}(D_i 1\{t = T_i\} + \frac{K_M(i,t)}{C}(1 - D_i))(Y_{it} - Y_{i,t-1}), \\
&\sum_{t=1}^{T}(D_i 1\{t = T_i\} + \frac{K_M(i,t)}{C}(1 - D_i))(Y_{it} - Y_{i,t-1})|\mathbf{D}, \mathbf{X}\Big\} \\
&= \frac{1}{N_1}\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{t'=1}^{T}(D_i 1\{t = T_i\} + \frac{K_M(i,t)}{C}(1 - D_i))(D_i 1\{t' = T_i\} + \frac{K_M(i,t')}{C}(1 - D_i))\times \\
&(Cov(Y_{it}, Y_{it'}) - Cov(Y_{it}, Y_{i,t'-1}) - Cov(Y_{i,t-1}, Y_{it'}) + Cov(Y_{i,t-1}, Y_{i,t'-1}))
\end{aligned}$$

To show that $E[(\hat{\sigma}^2_{1N} - \sigma^2_{1N})^2] \to 0$ we use a similar argument to before, where we are able to bound the difference between the $e_{i,t}e_{i,t'}$ and $Cov(Y_{it}, Y_{it'})$ by a fourth order polynomial in e_{it} terms and use the assumption that the expectation of these fourth moments is bounded.

B.2 Weighted Least Squares

Weighted least squares (WLS) is commonly used with matching weights to calculate the ATT and its corresponding confidence interval Ho et al. [2007], Stuart et al. [2011]. For WLS, we have the following estimator for β:

$$\hat{\beta} = (X^T W X)^{-1} X^T W Y$$

Where W is a diagonal matrix with entries corresponding to the matching weights. Note that $W = W^T$. We can compute the variance as follows:

$$Var(\hat{\beta}) = (X^T W X)^{-1} X^T W [Var(Y)]((X^T W X)^{-1} X^T W)^T$$

Software to compute WLS estimates, such as lm and Zelig (which calls on lm) in R, often assumes that $Var(Y) = \sigma^2 W^{-1}$. If this is true we get a cancellation that yields a nicer formula for variance:

$$
\begin{aligned}
Var(\hat{\beta}) &= (X^TWX)^{-1}X^TW\sigma^2W^{-1}((X^TWX)^{-1}X^TW)^T \\
&= (X^TWX)^{-1}X^TWW^{-1}W^TX\sigma^2(W^TWX)^{-1} \\
&= (X^TWX)^{-1}(X^TWX)\sigma^2(W^TWX)^{-1} \\
&= \sigma^2(W^TWX)^{-1}
\end{aligned}
$$

In R, lm (and Zelig) use this formula[1] in order to compute standard errors and confidence intervals for WLS regression. However, when this assumption is not true, as is the case in our simulations (and often in practice), this formula is incorrect.

Assume $Var(Y) = \sigma^2 I$, as in some of our simulations, then:

$$
\begin{aligned}
Var(\hat{\beta}) &= (X^TWX)^{-1}X^TW\sigma^2I((X^TWX)^{-1}X^TW)^T \\
&= \sigma^2(X^TWX)^{-1}X^TW^2X(X^TWX)^{-1}
\end{aligned}
$$

To get an unbiased estimator of σ^2 we note that:

$$
E[e^TIe] = tr(I\Sigma_{ee})
$$

Where $e = Y - \hat{Y} = Y - X\hat{\beta} = Y - X(X^TWX)^{-1}X^TWY$. Now,

$$
\begin{aligned}
\Sigma_{ee} &= Cov(Y - X(X^TWX)^{-1}X^TWY, Y - X(X^TWX)^{-1}X^TWY) \\
&= Var(Y) - 2Cov(Y, X(X^TWX)^{-1}X^TWY) + Var(X(X^TWX)^{-1}X^TWY) \\
&= \sigma^2I - 2X(X^TWX)^{-1}X^TWVar(Y) + (X(X^TWX)^{-1}X^TW)^T(X(X^TWX)^{-1}X^TW)Var(Y) \\
&= \sigma^2(I - 2X(X^TWX)^{-1}X^TW + WX(X^TWX)^{-1}X^TX(X^TWX)^{-1}X^TW)
\end{aligned}
$$

Thus, to get an unbiased estimator of σ^2 we must divide e^Te by the trace of $I - 2X(X^TWX)^{-1}X^TW + WX(X^TWX)^{-1}X^TX(X^TWX)^{-1}X^TW$. When we use this formula to create confidence intervals for WLS, instead of using lm, our simulations cover in the linear DGP, uncorrelated errors case. However, we do not expect that this correction addresses all issues with model-based parametric standard errors after matching under rolling enrollment — in particular, many of the problems identified by Abadie and Spiess Abadie and Spiess [2021] likely persist in some form — and we recommend use of the block bootstrap procedure discussed in the main manuscript instead.

B.3 Additional Simulations

In this section, we provide additional simulations with the same set-up as Section 8.5 but worse overlap between treatment and control units.

[1] Last verified 04-11-2022.

In particular, we modify the distribution of X_2 and X_6 by increasing the mean for treated units by a multiple of 4. Now, for treated units:

$$X_2 \sim N(1,1) \text{ and } X_6 \sim N(2,1)$$

The rest of the simulation set-up remains the same. Table B.1 summarizes the average standardized difference in X_2 and X_6 between matched treatment and control groups. We do not include the other covariates in this table, because they are well matched with average differences close to 0. As evidenced by the table, this simulation set up has poor overlap leading to bad balance after matching.

	X_2	X_6
Average Difference	0.52	0.93

Table B.1: Average standardized difference between matched treatment and control groups for X_2 and X_6 for our simulation set-up under the bad overlap specification.

Tables B.2 and B.3 summarize the coverage and average confidence interval length results for these simulations. Unsurprisingly, WLS Cluster continues to perform well in the linear DGP settings, even with bad overlap, because it has the correct model form. However, WLS and WLS Cluster's coverage decreases substantially with increased misspecification. On the other hand, our bootstrap method does not perform as well because of the low quality of the matches. In general, we'd advise against trusting matching results when there is high imbalance in the matched data.

Coverage	WLS	WLS Cluster	Bootstrap Bias Corrected
Linear DGP	84.2%	94.1%	87.2%
Linear DGP, Correlated Errors	77.3%	87.3%	73.7%
Nonlinear DGP, Correlated Errors	71.1%	83.1%	73.7%

Table B.2: Coverage of the WLS, WLS cluster and bootstrap bias corrected methods of inference for our three simulation set-ups under our bad overlap specification.

Average CI Length	WLS	WLS Cluster	Bootstrap Bias Corrected
Linear DGP	0.30	0.40	0.35
Linear DGP, Correlated Errors	0.31	0.39	0.32
Nonlinear DGP, Correlated Errors	0.32	0.41	0.44

Table B.3: Average 95% confidence interval length for the WLS, WLS cluster and bootstrap bias corrected methods of inference for our three simulation set-ups under our bad overlap specification.

B.4 Falsification Test Simulations

In this section, we illustrate the timepoint agnosticism falsification test presented in Section 8.6 via simulation. We generate a dataset of 1000 control units each with 4 covariates

and 2 instances occuring at times t_0 and t_1. Two of the covariates vary with time, and two are uniform across time:

$$X_{1,i,t}, X_{2,i,t}, X_{3,i,t_0}, X_{4,i,t_0} \sim N(0,1) \text{ for } t = t_0, t_1$$
$$X_{j,i,t_1} = X_{j,i,t_0} + \epsilon_{j,i} \text{ for } j = 3, 4$$
$$\epsilon_{j,i} \sim N(0, 0.5^2) \text{ for } j = 3, 4$$

The outcome variable is a linear combination of the four covariates, a time trend controlled by parameter γ, and an error term ($\epsilon_{i,t} \sim N(0,1)$):

$$Y_{i,t} = log(4)(X_{1,i,t} + X_{4,i,t}) + log(10)(X_{3,i,t} + X_{4,i,t}) + 1\{t = t_1\}\gamma + \epsilon_{i,t}$$

We generate data for the setting $\gamma = 0$, which does not include a time varying component, 1000 times. On each dataset, we perform the test for timepoint agnosticism outlined in this section, with 1-1 matching and bias correction. Our simulated data only contains two timepoints and the sample size is balanced so we choose the first timepoint as t_0, our new control group, and the second timepoint as t_1, our new treated group. In 0.05 of the simulations the P-value is less than 0.05, which shows that type I error is controlled. Now, we add in a time trend where there is an additional term, γ, added to the second timepoint. For $\gamma = 0.1$ for the second timepoint (resulting in a time trend of 0.1), our simulations result in a P-value less than 0.05 in 0.60 of the 1000 simulations. For $\gamma = 0.25$ for the second timepoint (resulting in a time trend of 0.25), our simulations result in a P-value less than 0.05 in all of the 1000 simulations. Table B.4 summarizes these results.

Figure B.1 shows two simulated datasets with $\gamma = 0.1$. The test for timepoint agnosticism detects the trend in one of the two datasets. Overall, the simulations show that the test is not a panacea for issues with timepoint agnosticism, sometimes failing to detect small violations. However, it still adds substantial value to the analysis pipeline, detecting moderate violations of timepoint agnosticism not especially obvious to the eye in visualization plots with a very high rate of success.

Time Trend, γ	0	0.1	0.25
Proportion P-values < 0.05	0.05	0.60	1

Table B.4: Summary of timepoint agnosticism simulation results. Proportion of 1000 simulations where the P-value from the test is less than 0.05, for time trends, $\gamma = 0, 0.1, 0.25$.

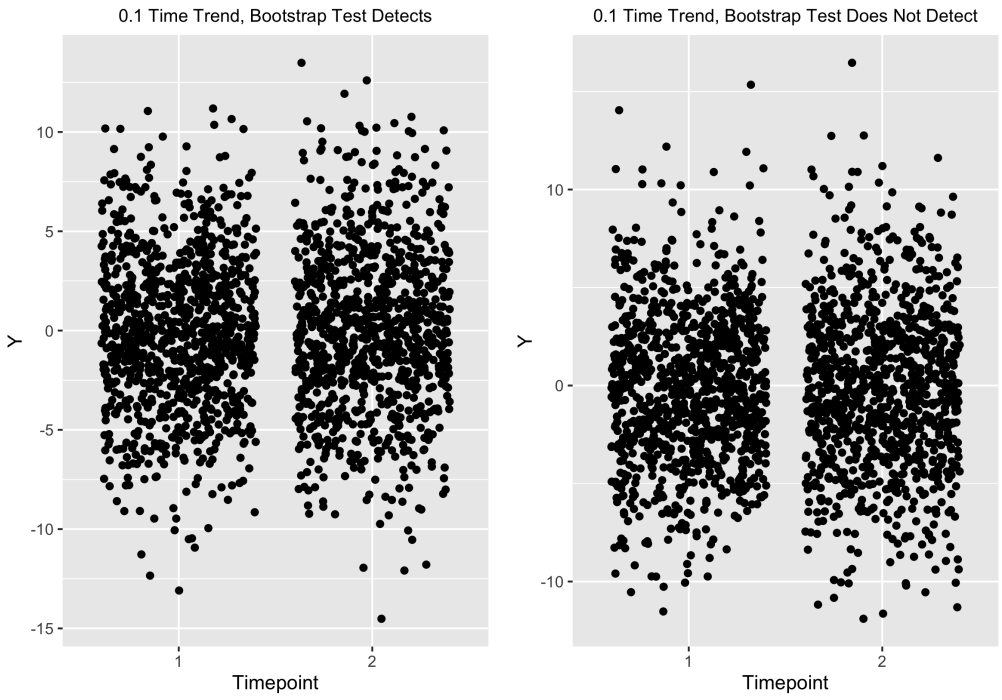

Figure B.1: Simulated datasets with time trend, $\gamma = 0.1$. The figures show outcome data for a dataset where the timepoint agnosticism test detects the time trend and does not detect the time trend respectively.