8-26-2022

Public Administrator Aversion to Randomized Controlled Trials

Emily Bryn Cardon
Syracuse University, emily.cardon@gmail.com

Abstract

This dissertation examines whether public administrators are averse to randomized controlled trials, and why. More specifically, it explores whether public administrators are reluctant or opposed to conducting RCTs for the evaluation of social welfare programs in situations where they would help to answer an important and useful policy question and are feasible to conduct. While RCTs are widely used to evaluate clinical interventions and practices, they are used far less often in social policy.

I first propose a conceptual framework for understanding the public administrator's decision-making process as they contemplate whether to conduct an RCT, outlining nine factors that likely contribute to RCT aversion. Then, using a survey experiment with a nationally representative sample, I investigate three research questions. First, do people, on average, prefer a quasi-experiment to the RCT? Second, do features of the policy environment that create a greater perceived difference in treatment between groups contribute to more RCT aversion? And third, do preferences for the RCT differ for public administrators compared to their non-public administrator peers?

I find that the majority of people demonstrate a strong preference for the quasi-experiment to the RCT. Public administrators are RCT averse on average, but less so compared to their general public peers. Additionally, I find that public administrators are likely to be more RCT averse when the intervention is perceived to be very different, and potentially superior, to the status-quo option available to members of the control group. People who are not public administrators are not sensitive to features of the policy environment. I conclude by outlining several avenues for future exploration of public administrator RCT aversion, and implications for

social policy researchers, evidence-based policy advocates, and public policy and administration educators.

PUBLIC ADMINISTRATOR AVERSION TO RANDOMIZED CONTROLLED TRIALS

by

Emily Bryn Cardon

B.A., Boston College, 2008
MPA, Syracuse University, 2013

Dissertation
Submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Public Administration.

Syracuse University
August 2022

Acknowledgments

> “Things take the time they take. don’t worry. How many roads did St. Augustine follow before he became St. Augustine?” - Mary Oliver

Had I come across this quote from my favorite poet when I started work on my dissertation, I may have felt very different about the process. While this project has taken far longer than I expected, it is a product of all of the learning I have had along the way. And for that, I am indebted to so many people. I have not walked these roads alone.

My deepest thanks to all of my committee members, Robert Bifulco, Alfonso Flores-Lagunes, Sarah Hamersma, and Katherine Michelmore. They have graciously shared their expertise with me over the years, in many classrooms and through several drafts. I am a better researcher and scholar because of them.

I could not be more grateful to Leonard Lopoo for his never-failing patience and encouragement over just shy of a decade. He has been endlessly generous with his time, challenged me to think deeper and broader, and gave me direction when I felt stuck. Not every student can say they their advisor cares about them personally. I am lucky to have known this from the start.

I am so fortunate to be a member of the Behavioral Insights Team, where I am surrounded by the brightest and kindest colleagues I have ever known. Joining the team gave me a sense of purpose and confidence when I felt directionless. Thank you to everyone, especially Michael Hallsworth, Cecily Wallman-Stokes, and Kelsey Irvine, who believed in me and spent many hours talking with me as I worked through the outlines of what is now this finished dissertation.

I was blessed to meet Gayle Scroggs when I most needed a change in approach. Gayle's perspective and reassurance helped me carve a path to finishing that felt true to my strengths. My only regret has been not finding and working with Gayle sooner.

To my friends and my family: this would not have been possible without you. Knowing that you would see me exactly the same, whether or not I had letters after my name, gave me the freedom to keep going. Thank you for everything.

This program brought two very important people into my life. Laura, thank you for doing this with me. Being your friend has been an amazing gift. And Joe, I'm so lucky you decided to come back to Syracuse. I like you the most.

This dissertation is dedicated to my Ronan, who was with me nearly every step of the way.

Table of Contents

List of Illustrative Materials

List of Figures

List of Tables

Chapter 1. Introduction

In 2019, the US government passed the bi-partisan Foundations for Evidence-Based Policymaking Act, referred to as the Evidence Act for short. The goal of the law is straightforward: to help federal agencies advance evidence-building, specifically by making it easier to access data and expanding the capacity to conduct evaluations (Office of the Assistant Secretary for Planning and Evaluation, n.d.). The Evidence Act was the culmination of recommendations from the Commission on Evidence-Based Policymaking, which was convened to study what would be needed to advance evidence-based policy in the US federal government (Abraham & Haskins, 2017).

Both the creation of the Commission and the passage of the Evidence Act illustrate the advancement of the evidence-based policy movement in the US. The central principle behind this movement is that public policy decision-making should be based on rigorous evidence and the best available data about an issue throughout the policy-making process such as creating new policies or programs, or adapting current programs to achieve better results (Head, 2015; Results for America, 2019).

Advancing evidence-based policy requires two things: an ability to generate sufficient evidence to be useful to public policy decision-making and the will or incentive to apply the evidence. Haskins (2018) argues that an increase in both of these areas has contributed to the rise of the evidence-based movement. First, the field of social science has developed a sophisticated set of methods to be able to determine whether social programs are effective in a real-world context. Second, a sufficient amount of evidence has been generated to be useful to policy-makers. Haskins (2018) cited the growing number of evidence clearinghouses, which centralize and grade research findings. Third, policy-makers increasingly cite evidence as a way to boost political support for programs and policies; a phenomenon also documented by Shulock (1999). However, while evaluation and the evidence it produces may be increasing in popularity, social policy experiments are still not the typical or default approach in the US. For the vast majority of public programs, there is little evidence suggesting that the program works, by how much, and for whom. Cook and Ludwig (2006) claimed that "for many social policy applications we either must give up on the goal of evidence-based policy, or develop a broader conception of what counts as evidence" (p. 694). This debate around what should count as rigorous evidence is active with many researchers and most evidence clearinghouses favoring the randomized controlled trial as a gold standard (Baldassarri & Abascal, 2017; Heinrich, 2007; Orr et al., 2019) and others arguing that randomized experiments should not be afforded special status (Cook & Ludwig, 2006; Deaton & Cartwright, 2018; Heckman & Smith, 1995).

The relatively low use of policy experiments to evaluate social welfare programs, and the debate around whether they are particularly useful or even appropriate, is especially interesting when contrasted with the health and medical field. In the clinical context, the use of the randomized controlled trial has been default federal policy since the 1960s. The RCT is viewed as an integral part of medical quality assurance, validating the safety and efficacy of new treatments and interventions. In response to the thalidomide drug crisis, the FDA approval process raised the bar by requiring large and well-controlled randomized trials for approval of new drugs and clinical practices (White Junod, n.d.). However, a similar mandate for applying the same methodology to social programs and public administration does not exist. What might account for this difference?

Randomized controlled trials are not a practical methodology for answering every policy question. For example, if we want to understand the impact of neighborhood choice on later life economic outcomes, we cannot feasibly create and randomize residents to a control neighborhood where there is limited schooling and economic opportunity. For these types of questions, different methodologies are needed to simulate a counterfactual. However, there are many questions in public policy and administration that are important and could be feasibly answered using a "randomized controlled trial" (RCT). Yet, RCTs may be avoided for one reason or another. I call this RCT aversion.

I define RCT aversion as a systematic preference against choosing a randomized controlled trial, in the cases where an RCT would be both practical and useful to answer a policy question of interest. This definition has two components worth some elaboration. First, it needs to be a conscious preference, in that the individual makes an intentional decision to choose other evaluation options over the RCT. A person cannot be RCT averse if they simply are not aware of the RCT as an option, or are not aware that they are making an evaluation decision in the first place. Secondly, RCT aversion is something systematic and could be observed on average in a population and across different contexts. RCT aversion is not the result of random and unique individual preferences, in the way individuals may prefer some colors or flavors to others. In other words, RCT aversion is not the random noise we might observe when individuals make choices about evaluation methods.

This dissertation investigates whether RCT aversion exists, and if so - why. It is important to note that aversion in this context does not exclusively mean a bias or error in judgment, although there may be biases and cognitive errors that do contribute to RCT aversion. But rather, RCT aversion may be the result of an entirely rational calculation on the part of individuals. People may be correctly identifying that the RCT is not the optimal choice in a given context, for a variety of reasons.

While there have been several discussions of why randomized experiments are not used more widely (Cook & Payne, 2002; Gueron, 2002; Maynard, 2006), there have been few studies to explicitly focus on RCT or experimental aversion. Meyer and colleagues (Heck et al., 2020; Meyer et al., 2019) find evidence across 19 survey experiments that people may approve implementing or scaling an untested policy or treatment, and disapprove of using a randomized controlled trial to determine which policy option is the most efficacious. They find this effect to persist across several survey designs and policy domains. In contrast to these results, Mislavsky et al. (2019) do not find evidence of general experiment aversion. In six experiments, they found that people's approval of an experiment was related to the acceptability of the policies it was testing. They, therefore, concluded that objections to experiments have less to do with the methodology and more to do with the policy interventions.

In contrast to these previous studies on experimental aversion, this study focuses on a particular group of people: public administrators. The evidence-based policy movement is interested in advancing both the generation and the application of evidence to policy. Randomized controlled trials are often a strong methodological choice for understanding the effectiveness of social welfare interventions. While there are many actors involved in implementing an RCT, public administrators are key stakeholders in this process. They will be gatekeepers in deciding whether the RCT is able to move forward: if they are RCT-believers, they have significant influence in pushing it along, and if they are RCT-skeptics, they can halt progress. Therefore, understanding whether and why public administrators may be RCT averse is particularly important for increasing the use of policy experiments.

While my focus is on public administrators, I do employ a wide definition of who may be included in that group. Specifically, I use public administrators to refer to someone who has significant or final decision-making authority over the management or implementation of a public policy, program, or service. Excluded from this definition are people who have lower levels of decision-making power, such as entry-level staff, or who work in an organization involved in the delivery of public goods or services but not in a role that would oversee or participate in evaluation decisions, such as the accounting department. However, the public administrator I have in mind could be at various levels of management (such as a mid-level manager or the top bureaucrat), in a variety of roles (such as overseeing a single program or sitting in a central performance office), in various levels of government (local or federal), and even across organization types (a government, a non-profit, or a private entity contracted by the government.)

I am particularly interested in whether the decision-making of public administrators differs from the decision-making of other groups. Public service motivation theory suggests there is reason to expect that public administrators will differ from others in notable ways. For example, when compared to private sector employees, public administrators are found to value public service at higher levels and to be motivated by intrinsic factors more than extrinsic rewards (Crewson, 1997; Perry & Hondeghem, 2008; Vandenabeele et al., 2018). How public service motivation may relate to RCT aversion is one of the questions this dissertation seeks to explore. Of course, there may be other professional factors that make public administrators more or less RCT averse than others, such as greater knowledge about costs or organizational dynamics. Public administrators may also differ systematically along some correlated demographic dimensions, such as greater educational attainment or exposure to program evaluation. In other words, while there is a strong reason to expect that public administrators may make evaluation choices differently than a group of non-public administrator peers, it is not immediately clear whether they will be more or less RCT averse.

The purpose of this dissertation is to understand whether public administrators are RCT averse, and if so, why. Specifically, are public administrators hesitant or opposed to using RCTs to evaluate social welfare programs? It is organized into two main parts. Chapter 2 outlines my conceptual framework and explores potential reasons why public administrators may be RCT averse. I first define a simple model of the evaluation decisions a public administrator must make to arrive at the RCT design. The benefit of this decision tree is that it helps to clarify three different components of RCT aversion: whether someone is averse to evaluations in general, averse to impact evaluations specifically, or averse to randomized controlled trials in particular. Next, I describe the public administrator's utility model, and how their broader decision-making may be shaped by the limitations in how people optimize choices, balance between individual and collective interest, and the influence of risk. Finally, I outline nine factors that are likely to influence the public administrator's utility calculation, and either leads to greater or less RCT aversion.

Overall, the main conclusion from the discussion of the nine factors is that there are very few clear incentives for public administrators to conduct evaluations in general, let alone impact evaluations and RCTs. At the same time, there are several clear disincentives. Risk aversion makes this dynamic worse, as the disincentives are likely to be weighed more than equivalent incentives. In other words, to favor the RCT, the public administrator must believe that the benefits far outweigh the potential downsides. The main contribution of this chapter is to outline a potential research agenda to better understand the public administrator's evaluation decision-making process.

Chapter 3 turns to an empirical test of one of the nine factors: that one source of RCT aversion lies in the challenge it presents to distributive justice norms. The randomized controlled trial requires that the public administrator explicitly deny a social welfare intervention to an equivalent group of people on the basis of an arbitrary assignment mechanism. Other impact evaluation methods do not present the same tradeoffs, at least not explicitly, and therefore public administrators are not likely to view them as a challenge to ethical norms.

Using a survey experiment with a nationally-representative sample, I investigate the factors that may affect evaluation choices. Specifically, in the context of a city-run lead abatement program, I present participants with an evaluation scenario and ask them to decide between an RCT or a quasi-experimental (QE) approach. I use a vignette factorial survey methodology to answer three research questions. First, do people, on average, prefer a quasi-experiment to the RCT? Second, do features of the policy environment that create a greater perceived difference in treatment between groups contribute to more RCT aversion? And third, do preferences for the RCT differ by certain characteristics? Specifically, do public administrators have different evaluation preferences compared to their non-public administrator peers?

I find that the majority of people demonstrate a strong preference for the quasi-experiment to the RCT. Public administrators are RCT averse on average, but less so compared to their general public peers. Specifically, they are about 13 percent more likely to prefer an RCT to a quasi-experiment. This higher likelihood to select the RCT evaluation option can be partially explained by greater educational attainment, greater prior experience with program evaluation, and greater decision-making authority. Finally, public administrators are likely to be more RCT averse when the intervention is perceived to be very different, and potentially superior, to the status-quo option available to members of the control group. People who are not public administrators are not sensitive to features of the policy environment. Chapter 4 summarizes the dissertation and provides concluding thoughts. In particular, I provide a set of implications for researchers, evidence-based policy advocates, and policy and evaluation educators. The study concludes with a bibliography and appendices.

Chapter 2. Conceptual Model

Introduction

This central research question of this dissertation is whether public administrators are averse to using randomized controlled trials (RCTs), and if so, why. Specifically, are public administrators hesitant or opposed to using RCTs to evaluate social welfare programs?

The previous chapter outlined the role of RCTs in the evidence-based policy movement in the US and the limited literature on experiment aversion. RCTs have played an important role in providing evidence about what works in the health and medical field and have become the default scientific method of choice for establishing the safety and efficacy of new innovations. However, the RCT method has been used less widely in areas of public policy to evaluate social welfare interventions.

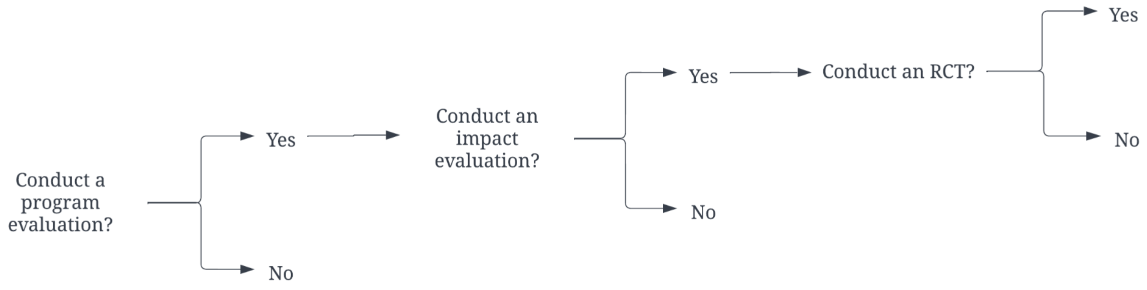
This chapter explores potential reasons why public administrators may be RCT averse. First, I define a simple model of the evaluation decisions a public administrator must make to arrive at the RCT design. Second, I describe the public administrator's utility model, and how their broader decision-making may be shaped by the limitations in how people optimize choices, the balance between individual and collective interest, and the influence of risk. Finally, I outline nine factors that are likely to influence the public administrator's utility calculation, and either contribute to greater RCT aversion or less.

These nine factors are potential starting points for testable hypotheses, and therefore may provide a roadmap to a broader research agenda around the public administrator's role in evidence generation. However, the net effect of these factors on RCT aversion will be highly context-specific. It is beyond the scope of this dissertation to provide a test for all of these factors and their inter-relationship. The next chapter will focus on the last of these factors that I argue uniquely impacts aversion to RCTs: the perception that the RCT presents a unique conflict with distributive justice norms.

The policy RCT decision tree

To start to unpack RCT aversion, it's helpful to model the choice set facing the public administrator. As the public administrator considers whether and how to evaluate a social welfare intervention, they face a range of options. Below, I present a simplified model of how a public administrator may arrive at the decision to conduct an RCT (Figure 1). The decision tree has three significant decision points: whether to evaluate or not, whether to conduct an impact evaluation or not, and whether to conduct an RCT. Each decision is nested within the decision that came before. For this discussion, I assume that each choice is made separately and that evaluation decision-making events are independent. In reality, this is likely to be more complicated, as individuals are likely to make joint decisions and period one choices are likely to influence choices in period two and beyond.

Figure 1: RCT Decision-Tree



Imagine there is a public administrator named Vicky. Vicky is a senior leader in the Mayor's Office for Cityville and is responsible for performance and innovation. Vicky helps to ensure that delivery and management of Cityville's programs and services are aligned with the Mayor's goals and strategy. She has been tasked with finding ways that Cityville government can advance economic growth and mobility for Cityville residents, specifically through targeting health, education, and financial stability.

Relative to cities of similar size, Cityville struggles with housing stability. The housing stock is old, and many residents are tenants. Past studies have shown that code compliance is low, with many residents staying in units of substandard quality. Vicky knows that safe and healthy housing contributes to positive health outcomes, with the biggest impacts on vulnerable individuals such as the young and elderly. The Mayor has signed off on an initiative to improving housing quality across Cityville, with the aim of improving environmental health and resident well-being.

In facing the first choice in the policy RCT decision tree, Vicky must decide whether to pursue an evaluation of this housing initiative. As the initiative represents a change in practice, there is an opportunity to plan a systematic investigation around what is implemented, whether standards of implementation have been met, what outcomes the initiative achieves, and whether the initiative has had the intended impact. Alternatively, Vicky may choose not to evaluate at all or simply report on programmatic spend. While both evaluation and performance management share an intention to ensure the initiative achieves its goals, and may even use the same data sources, evaluation uses scientific methods to answer specific research questions. In contrast, performance management is often concerned with monitoring the activity of intermediate outcomes to a prescribed benchmark or threshold (Heinrich, 2007). Therefore, pursuing an evaluation is likely to add additional activity and cost beyond what performance management may demand.

If Vicky decides to pursue a program evaluation around the housing initiative, the next choice she faces in the RCT decision tree is around the type of evaluation. RCTs are generally considered to be a type of impact evaluation, where the focus of study is around the program's or policy's efficacy in shifting key outcomes of interest. In other words, impact evaluation aims to assess whether an intervention has caused changes in observed outcomes. Impact evaluations answer questions around a program or policy's value add and whether it has made a difference. An alternative evaluation approach would be to engage in either a formative or process evaluation, which helps to answer different research questions. For example, a formative evaluation may help Vicky understand operational readiness for a new housing initiative, and which types of interventions are likely to be most effective in Cityville's context. A process evaluation may help Vicky understand how the initiative is implemented compared to its design, who participates in the initiative, and what happens over time.

While most evaluation methods are not mutually exclusive and can be used either in succession or in parallel, they can be conducted in isolation. It is possible, and common, to conduct an RCT without conducting a full process evaluation as well. There is also a range of options within each evaluation type, with some involving more resource than others and some achieving greater internal and external validity than others. Finally, each method will require different inputs and conditions. Not every process or impact evaluation will be feasible. Vicky will need to weigh tradeoffs in which types of questions are most valuable to answer, and how much she is willing to pay for those evaluation outputs. Exploring how Vicky may engage with those tradeoffs is the purpose of the sections that follow.

Should Vicky decide to pursue an impact evaluation, she faces the final choice in the RCT decision tree: whether to pursue an RCT or use a different impact evaluation approach. Key to the RCT method is the use of randomized assignment and a control group that does not receive the intervention. The RCT is also experimental rather than observational: RCTs force public administrators to manipulate implementation of the new program or policy, and the study must be initiated before the intervention has been rolled out at scale. Alternatively, there is a range of impact evaluations that relax these conditions, whether it's using a different assignment mechanism, comparison groups that are not "pure controls," or are observational rather than experimental. For example, natural experiments are an observational method that exploits seemingly random assignment to approximate the RCT design. Whether Vicky chooses to pursue the RCT in the evaluation of the housing initiative is dependent on the RCT being feasible to implement and her preference for the RCT approach. Factors that are likely to influence Vicky's preferences are discussed in the following sections.

The benefit to viewing the RCT decision tree in this simplified model is that it helps tease out three different components of potential public administrator RCT aversion. First, public administrators may be averse to evaluations in general, regardless of type. Factors contributing to this evaluation aversion, therefore, make it less likely that an RCT is chosen, but they are also likely to lead to fewer evaluations overall. Second, public administrators may be averse to impact evaluations specifically and prefer formative or process evaluation approaches. These factors are also likely to present as RCT aversion and lead to fewer RCTs, but they also contribute to aversion to quasi-experimental or natural experimental evaluation approaches as well. Finally, public administrators may be averse to the RCT approach and prefer alternative impact evaluation methods instead.

I hypothesize that at each decision point along the path to selecting the RCT as the preferred evaluation approach, the public administrator is likely to be influenced by factors unrelated to just the practicalities of the evaluation design. Instead, what evaluations are pursued is determined by a complex aggregation of individual preferences. The question I turn to next is, what factors are likely to influence public administrator evaluation decision-making? In other words, what might systematically lead public administrators away from using the RCT approach, even in cases where an RCT is feasible and practical to run?

The general utility model

Rational Choice Theory

For the purposes of this model, I start with a rational choice theory perspective (Shafir & LeBoeuf, 2002). I assume that public administrators are utility maximizers, subject to the constraints they face and reflecting their individual preferences. Each evaluation option provides a given level of utility to the public administrator. When deciding among evaluation options, public administrators will compare the options and choose the strategy that provides them with the most satisfaction. The public administrator also compares the utility from different evaluation strategies to alternative uses of those resources and choose the option that provides the most satisfaction. This means that the "best option" for the public administrator may be no evaluation at all. In sum, the public administrator will always choose the evaluation strategy (including no evaluation) that provides the maximum utility relative to their other choices.

How much utility the public administrator derives from evaluation, and for particular types of evaluations, will be determined primarily by their preferences. Context, personal development, experience, and personality are likely to play a significant role in determining the individual public administrator's tastes and interests. As is standard, I assume that preferences are complete (the public administrator will always be able to say whether they prefer the RCT to the quasi-experiment, the quasi-experiment to the RCT, or they are indifferent between the two) and transitive (if they prefer the RCT to the quasi-experiment, and the quasi-experiment to the pre-post, then they will prefer the RCT to the pre-post) (Buskens, 2015).

Bounded Rationality

Classic starting assumptions in rational choice theory are that actors are rational and self-interested, engaging in one-time interactions within a transparent context and with complete information (Buskens, 2015; Zey, 2015). Relaxing these assumptions broadens the model. For example, theories have loosened the contextual assumptions of the classical rational choice theory model, recognizing that actors often interact repeatedly within a social environment with given norms (Coleman, 1988, 1990), are embedded within networks (Granovetter, 2002), or are influenced by others (von Neumann & Morgenstern, 2007).

There are two implications of this broader definition for the model of public administrator evaluation decision-making. First, public administrators do not operate in a vacuum. Their social environment is likely to be extremely influential, as organizational culture and political context affect their utility function. Similarly, public administrators will need to make evaluation decisions multiple times. A decision to evaluate that leads to a poor or sub-optimal outcome is likely to influence the decision to evaluate in the next round.

There have also been extensions to the rational and selfish assumptions of the neo-classical rational choice model. Simon's (1955, 1990) theory of bounded rationality recognizes that the ability of individuals to engage in complex decision-making is limited, especially given the role of uncertainty and the limits of time and information (Gigerenzer, 2008, 2010; Simon, 1990). Therefore, people may satisfice, or rely on imitation and heuristics leading to satisfactory, but not necessarily fully optimal, outcomes (Buskens, 2015; Gigerenzer, 2010; Gigerenzer & Selten, 2001).

A bounded rationality and satisficing perspective means that we do not expect public administrator decision-making to be perfect. Instead, their choices may stray from optimal because of the public administrator's biases and perceptions, lack of information, or cognitive processes. In other words, public administrators may end up choosing a second or third-best option and not fully maximizing their total utility. For this model, it does not matter whether these issues are viewed as simply optimization under constraints (Arrow, 2004), or systematic errors in judgment (Kahneman, 2003). In any given situation, the public administrator is "doing their best" to maximize utility, but may have made a different choice given more time or information.

The theory of rationality holds that individuals aim to optimize their wellbeing or utility, given their preferences. However, while these preferences will be largely self-interested, they do not need to be entirely selfish. Preferences can also be altruistic. The inequality aversion model by Fehr and Schmidt (1999) proposes that an actor experiences utility based on what they receive, but they also will experience disutility from differences between what they receive and what others receive. There has also been a significant literature investigating the "unselfish" components of motivation and behavior (Simon, 1991).

Public Service Motivation

Public service motivation (PSM) theory fits this tradition of investigated altruistically motivated preferences. In attempting to explain the behavior of public service employees and bureaucrats, PSM theory argues that the classical rational choice perspective comes up short. In public organizations, goals are often less specified, and performance is difficult to link to external and conditional rewards. Instead, PSM emphasizes the role of intrinsic, or self-determined, motivation, in explaining the efforts and choices of public servants (Wang et al., 2020). PSM theory has since expanded to more broadly mean a willingness or inclination to contribute to society (Perry & Hondeghem, 2008; Vandenabeele et al., 2018). Vandenabelle

(2007) defines PSM as “the belief, values, and attitudes that go beyond self-interest and organizational interest, that concern the interest of a larger political entity and that motivate individuals to act accordingly whenever appropriate” (p. 547). In other words, individuals motivated by PSM will demonstrate a greater orientation towards collective welfare, rather than just individual self-interest.

In addition to seeking to explain pro-social motivation, PSM theory also predicts that a public service orientation influences selection and sorting into public service work (B. E. Wright & Christensen, 2010). People with high PSM are predicted to cluster in government and nonprofit organizations due to better value alignment between employee and organization (Holt, 2018). Holt (2018) studies this relationship using longitudinal data, and finds that PSM-related values measured before entry into the labor market can predict employment sector.

Therefore, due to their PSM orientation, public administrators may have a utility function that is more heavily weighted towards altruism and public well-being. When applied to decisions around program evaluation, this suggests that public administrators are likely to choose the approach likely to maximize the positive impact on public welfare, all else being equal. They may also be willing to choose an approach that may have neutral, unclear, or potentially negative consequences for themselves, if they feel the evaluation approach is beneficial for the public good and aligned with their PSM values. How public administrators trade-off between social welfare and their own personal interests with regard to evaluation decisions is unclear, and there is likely to be a limit on how much personal loss a public administrator is willing to bear. The key point is that public administrators are likely to weigh the benefit to the public in their own utility calculation.

Prospect Theory

Another extension to rational choice theory is prospect theory, which examines how people make decisions under uncertainty. The standard rational choice model assumes that outcomes are known to the decision-maker, but often there is a degree of chance at play. Decisions can be “risky” if the probability or likelihood of a given outcomes is known, or decisions can be “ambiguous” if even the probabilities are unknown. This dimension of probability is then another factor in a person’s rational choice calculation, as they now need to decide both how to weight the decision’s utility and the likelihood of it occurring (LeBoeuf & Shafir, 2005; Shafir & LeBoeuf, 2002; Thaler, 1980).

Prospect theory, as described by Kahneman and Tversky (Kahneman & Tversky, 1979; Tversky & Kahneman, 1991), predicts that people prefer certain or fixed outcomes to outcomes that are risky or uncertain, even if both outcomes have equal expected value. This preference toward certainty over risk is known as risk aversion. For the public administrator, risk aversion is likely to bias them away from evaluation as an activity in general. All evaluations involve some level of uncertainty in the outcome, as evaluations where the results are known won’t be worth running. If the public administrator needs to conduct an evaluation, risk aversion is likely to lead them to prefer methodologies with more certainty and less uncertainty. This might look like running evaluations they have previous experience with, as there is more familiarity with the process and less that is unknown. It may also look like a preference towards evaluations where they have more control over the outcome, such as evaluation that studies the output of a process rather than a more distal impact measure.

Prospect theory also suggests that people will tend to view outcomes in relation to reference points, rather than by some fixed objective criteria. In other words, people tend to perceive outcomes as either a loss or gain relative to their current status, rather than a summation of the outcome’s total value. With regard to the public administrator, this means that they will be viewing how an evaluation will help or hurt them, relative to the status quo. Research has shown that losses tend to loom larger than gains, potentially as much as twice the impact (Akalis, 2008; Vis, 2011). Known specifically as loss aversion, this distaste for losses leads to a general preference for the status quo because the disadvantages from change outweigh the advantages (Akalis, 2008; Kahneman et al., 1991; Quattrone & Tversky, 1988).

Loss aversion and status quo bias means that public administrators are likely to be more sensitive to perceived losses than gains of equivalent value. When conducting a mental cost-benefit analysis about whether to run an evaluation, impact evaluation or RCT, public administrators are likely to not be equally comparing the pros and the cons. The RCT will need to have many more benefits, or gains, to overcome any potential costs or losses. This leads to a general conservatism in approach, where the public administrator is biased away from change and innovation. Most programs have not been evaluated or have supporting evidence, and RCTs are even more infrequent (P. J. Cook & Ludwig, 2006; Heinrich, 2007; Sanderson, 2002). Given that “running an RCT” is seldom the status quo position, prospect theory predicts that the public administrator’s utility calculation is likely to be biased away from the RCT. In other words, public administrators are likely to be evaluation, impact evaluation, and RCT averse simply because all three choices are inherently risky.

Risk aversion, loss aversion, and status quo bias are phenomena that affect everyone. However, evidence suggests that people range in their risk tolerance, with some people are more risk averse than others. For example, through an experiment with a car manufacturer’s customers, Gächter et al. (2022) find evidence that loss aversion increases with age, income, and wealth, and decreases with educational attainment. Therefore, if public administrators are older and more economically secure than the general public, we may expect public administrators to potentially be more risk averse on average. Conversely, if public administrators are likely to have more educational attainment, we may expect their risk aversion to be tempered.

There is also some suggestive evidence that people with low-risk preferences selectively sort into the public sector. For example, Bellante and Link (1981) found that public employees were more risk averse than private employees across a range of behaviors, including cigarette and alcohol use, seat belt use, and automotive insurance. Dong (2017) used data from the National Longitudinal Survey to find that higher levels of risk aversion, as measured by income gamble questions, was associated with a higher likelihood to choose public sector work later in life. Risk preferences have also been linked to innovation in public service delivery (Torugsa & Arundel, 2017). Whether it is related to the nature of public service, the culture within public organizations, or characteristics of the individual public administrators, it is likely that public administrators will tend to be on the more risk averse end of the spectrum.

Influencing factors

Factor 1: Costs

All else equal, the higher the cost of an evaluation, the less likely the public administrator is to use the RCT

Feasibility is likely to be a main factor in driving public administrator decision-making about whether and how to evaluate. While resource budgets can shift over time, in any given moment, the public administrator will have a budget constraint that limits their evaluation choice set. Public administrators are going to be cost minimizers. In other words, for evaluations with equivalently perceived benefits, they are likely to choose the evaluation approach that has the lowest possible cost. As the cost of a given evaluation approach increases, public administrators face an incentive to substitute away from that approach. This may mean shifting to lower cost evaluation designs or reducing the quantity of evaluation overall.

While this dynamic may lead to fewer RCTs overall, public administrator RCT aversion is not the cause. Even if the public administrator may prefer the RCT in the abstract, resource constraints limit the evaluation choices available to them. At the same time, it is important to note that the value of the output, or the evaluation in this case, is subjective and will be partially informed by the public administrator's preferences and knowledge. In other words, whether the public administrator sees the outputs of an RCT or a simple comparison study as having different or similar levels of quality is key to understanding cost tradeoffs.

For each specific evaluation opportunity, there will be a range of evaluation approaches. Within each approach, there are more or less expensive versions for each evaluation design. RCTs will not always be more expensive than other types of impact evaluations or other types of evaluation. For example, if the RCT uses administrative data that is already routinely collected, and the technology to randomize is built-in to the system, then conducting the RCT may be marginally more expensive than no evaluation, with the costs mostly driven by additional staff time to analyze the data. Quasi-experiments may be more expensive in this case, as additional data may have to be gathered and partners with specific statistical expertise brought in to conduct analysis. Then again, RCTs may be more expensive if they require new data collection or processes to complete random assignment. Costs are likely to be highly specific to the context of the individual evaluation opportunity and rely on what evaluation inputs are required to answer a set research question and what current inputs are both available and how much they cost.

However, it is difficult to think of an example where conducting an evaluation of any type is going to be less expensive than conducting no evaluation at all. Because public administrators will have many demands for their limited resources, and many ways to increase utility through spending their budget, their budget constraint is likely to lead toward fewer evaluations than more, all else equal. As the associated costs of evaluation decrease, I predict that there are likely to be more evaluations of every type. These associated costs include staff time and skills, data, and technological infrastructure.

Staff time is an opportunity cost, as fewer competing priorities means that more effort and focus can be devoted to evaluation. The available evaluation skill set is also a resource. The broader or deeper this skillset within the organization, the cheaper the evaluation is to run. In other words, evaluation opportunities do not need to compete among each other for limited staff time, and employees can be more flexibly staffed to accommodate evaluation needs. Public organizations can also acquire skillsets externally, through partnerships and consultations, but procuring and managing those relationships also incurs a cost.

Data is a necessary input for all evaluations. Some data are more resource-intensive to collect and process than others. I predict that cost-minimization means that public administrators are likely to prefer evaluations making use of readily available, and thus cheaper, data. For example, simple comparisons that group individuals based on a few characteristics are likely to be cheaper than matching designs that make use of an extensive set of socio-demographic factors. Evaluation designs that make use of smaller time periods are likely to be preferred to longer time periods, as it minimizes the time a data gathering instrument needs to be active. Access to administrative data has been widely acknowledged as a factor in enabling low-cost randomized controlled trials (Shankar, 2014). Increasing access to government data was a major recommendation of the Commission on Evidence-based Policymaking and was reflected in the passage of the Foundations for Evidence-based Policymaking Act of 2018 (Hart & Davis, 2017).

Better access to data will also lower the cost of other types of evaluation, and therefore, while reducing barriers to the RCT, may not change the RCTs relative cost.

In addition to the data itself, evaluations often make use of specific technology and other fixed capital resources. Cost minimization means that public administrators are likely to prefer the evaluation methodology that requires the least amount of capital investment and outlays. Therefore, evaluation designs that make use of free and widely available technology are likely to be preferred to evaluation designs that require expensive and restricted technology. For example, an evaluation that can be completed using Excel is likely to be preferred to one that requires proprietary, paid software such as Stata or SAS; free survey software is likely to be preferred to Qualtrics or SurveyCTO. Cost minimization may lead to a preference for simpler, less sophisticated designs that can be completed with basic or low-cost technology, than more sophisticated and potentially robust designs that require more expensive software or resources. While this doesn’t necessarily lead to greater RCT aversion specifically, it does incentivize the public administrator to prefer low-cost trials to more complicated ones.

It also matters who bears the costs, and when. Consider the example of two evaluation approaches, an RCT and a quasi-experiment. Assume that both approaches cost the same, requiring the same amount of overall time, data, and resource. However, the RCT requires more time from the public administrator and their staff, while the quasi-experiment relies heavily on the time of outside researchers whom the public administrator does not need to pay directly. In this case, cost-minimization incentivizes the public administrator to choose the quasi-experiment, because the overall cost to them is less, even while the total cost is the same. Because experimental approaches are likely to rely on some amount of additional time from public administrators or their staff, cost minimization is likely to incentivize the public administrator to prefer observational evaluation designs.

The timing of costs may also matter. Overall, public administrators are likely to prefer to delay costs as much as possible and bring forward benefits. Evidence has shown that people to engage in temporal discounting, where future rewards are weighted less than current ones (Frederick et al., 2002; O'Donoghue & Rabin, 1999). Experimental approaches generally require lots of resource investment up front, as time is spent designing and implementing the experiment. Observational approaches on the other hand, potentially have a smaller gap in time between when resources are spent, and the results are finalized. Timing around the financial year may also incentivize the public administrator towards some evaluation approaches and away from others. If there is a need to spend budget within a particular window, the public administrator may prefer evaluation designs with shorter observation periods. If there are budget caps, this may incentivize designs where costs are spread out over time. Again, the role of timing in contributing to RCT aversion is likely to be context-specific, but I hypothesize overall it is likely to incentivize observational designs over experiments, and it may impact the type of RCTs the public administrator is willing to consider.

In summary, the impact of cost-minimization on preferences for or against the RCT is likely to vary considerably based on the specific context. Some RCT designs are more costly than others. A RCT of two different versions of a mailed notice, using administrative data and a simple randomization approach, may require little additional investment beyond the staff time required to conduct the randomization and analyze the results. Alternatively, an RCT of a clinic-based intervention, using observational or survey data or new measures and a complex randomization and implementation procedure, may require substantial material and time investment. Overall, however, I hypothesize that cost minimization likely increases potential RCT aversion. Cost minimization likely leads the public administrator to have preferences for less evaluation overall, for observational designs over experimental, and for low-cost RCT designs compared to higher-cost designs, all else being equal.

Factor 2: Investments and default practices

All else equal, the more current resources are in place for implementing an RCT, the more likely the public administrator is to use the RCT

Prior investments in evaluation capacity are likely to impact a public administrator's evaluation preferences by making certain approaches cheaper than others. Some evaluation inputs can be viewed as fixed costs, in that once the investment has been made, future uses of that resource are less expensive (X. H. Wang & Yang, 2001). For example, once a software license has been purchased, that software can then be used in an unlimited number of future uses, or until the license expires or software becomes obsolete.

Fixed costs may also make certain evaluation approaches riskier, if it is not clear that they will be used widely in the future. Hiring and certain technologies are expensive investments. Investing in a new data system for a one-off evaluation is likely cost prohibitive if it is not clear how, or how much, it could be leveraged going forward. If investments are sunk costs, a variant on fixed costs in which the costs cannot be recovered, then the risks are even higher (X. H. Wang & Yang, 2001). This means that public administrators are incentivized to choose evaluation approaches that make best use of current resources and that require minimal additional fixed investments.

Current resources can also be seen as practices and procedures. Defaults have been shown to be extremely powerful tool to influence behavior, ranging across domains as varied as organ donation, retirement savings, and health insurance selection (Johnson et al., 2013; Johnson

& Goldstein, 2003; Madrian & Shea, 2001). A meta-analysis by Jachimowitz et al. argues that defaults have been shown to be more effective when they are perceived as providing an endorsement of a choice or reflecting the status quo (Jachimowicz et al., 2019).

Prior evaluation activity can create default approaches. People become used to working together in a particular way. They tend to be most comfortable using the methods and tools they have used previously. As discussed above, risk aversion is likely to contribute to a bias towards the status quo, as the potential losses from a change are likely to loom larger than potential gains. If a government department has typically conducted its own process evaluations, collaborating with the IT department to run an RCT is a departure from the default approach. The RCT therefore requires additional energy and attention that the process evaluation does not. All else equal, the public administrator is therefore likely to prefer the process evaluation in this situation.

Transaction and friction costs also disincentivize departures from normal operations. Williamson's transaction cost theory proposes that the optimum or most efficient organizational structure will minimize the costs of exchange, caused largely through coordination (1979, 1981). This suggests that evaluation approaches that rely on greater numbers of stakeholders are more expensive, as each additional actor adds complexity to decision-making or implementation. For example, when choosing between two evaluation methods, one that can be run "in house" and another that requires a new partnership, the public administrator is likely to prefer the one that can be run with fewer transactions. Similarly, if an RCT requires legal review and approval, but a quasi-experiment does not, the public administrator is likely to prefer the quasi-experimental approach that does not involve added frictions.

In summary, prior activity is likely to influence the public administrator's choice of evaluation approach. Fixed costs may make certain methods cheaper or more expensive than others, while defaults make certain approaches feel more comfortable. Transaction and friction costs make it more difficult to depart from the status quo. This combination of factors suggests that a particular evaluation approach is likely to face the most resistance the first time it is used. The more it is used, the more these costs decrease.

Therefore, in making the decision to run an RCT, the public administrator is more likely to choose the RCT approach if the organization has conducted an RCT before. Organizations that have run RCTs in the past are more likely to run RCTs in the future. Public organizations that have not run many evaluations or impact evaluations in the past are not likely to favor the RCT approach. Additionally, the more similar a previous and currently proposed RCT are, using the same resources, people, and practices, the more likely the public administrator is to choose the RCT as the preferred evaluation approach.

Factor 3: Evaluation skills and capabilities

All else equal, as program evaluation capabilities increase, the more likely the public administrator is to use the RCT

Evaluation skills and capabilities are a type of organizational resource. As discussed above, as these skills become more widely available, and thus cheaper and easier to access, public administrators face lower evaluation costs. However, improved evaluation skills may also have a second type of direct effect through helping public administrators understand the value of evidence, recognize evaluation opportunities, and understand the tradeoffs between different methodologies. This is improvement in ability or capacity then helps improve evaluation decision-making. In other words, greater evaluation skillsets help public administrators recognize when they should be using the evaluation decision-tree and how to navigate between the options.

Additional methodological training is likely to increase a public administrator's demand for evaluation by improving the perceived value of causal evidence. In a field experiment, Mehmood et al. (2021) find evidence that training deputy ministers in econometrics and causal thinking shifted attitudes on the importance of causal inference, increased willingness-to-pay for evidence from RCTs, and increased the likelihood of choosing a policy for which there is experimental evidence. Their study suggests that the poor appreciation of the benefits of impact evaluations and causal evidence may be one factor limiting the wider use of RCTs by public administrators.

Greater evaluation training and skills may also help make opportunities for evaluations more salient. In consumer choice settings, Bordalo et al. theorize that consumer's attention is drawn to notable attributes, and that these attributes are disproportionately weighted in making their choice (Bordalo et al., 2013). Salience is also important in opportunity recognition, as individuals must identify meaningful patterns and notice connections (Baron & Ensley, 2006). A similar dynamic may exist in choosing evaluation strategies. RCTs require certain conditions to be met, including mechanisms for random assignment, differentiation of treatment, and sample sizes sufficient to detect changes or differences in outcomes. Further, public administrators need to recognize the opportunity to conduct an RCT before an intervention has been implemented. If public administrators don't have sufficient knowledge of when an RCT is likely to be applicable or helpful, an observed avoidance of the RCT isn't truly an active choice on the part of the public administrator. To be considered RCT aversion, the public administrator needs to be making a conscious choice to avoid the RCT. If they don't recognize there is an opportunity for such a choice to be made, it is not RCT aversion.

More advanced evaluation skillsets can also help public administrators assess the evaluation options available to make an informed decision. For any given evaluation opportunity, there are generally multiple different methodological approaches that could be used, and then many variations within each approach. To decide between options, public administrators need to know what research questions each approach will answer, which methodology is likely to have the highest validity, and what implementation risks are most likely to occur. The answers to these questions rely highly on the specific context and details of the evaluation opportunity, making this a complex decision-task that changes with each new situation. Heath and Tversky (1991) found support for a "competence hypothesis," which holds that ambiguity aversion, a specific type of risk aversion related to outcomes where the probability is unknown, will decrease as people feel increasingly competent in the decision area. In other words, as people feel increasingly knowledgeable, they are more willing to take a chance on their own judgment.

In summary, I hypothesize that as the public administrator's evaluation skills and capabilities increase, so will their willingness to use an RCT for the evaluation of social welfare programs. The more targeted the skills gains are in the areas of causal inference and experimental design, the greater the decreases in RCT aversion. This is likely to operate through three channels. First, improvements in methodological capabilities are likely to increase the perceived value of causal evidence. Second, additional evaluation skills can make evaluation opportunities more salient. Finally, more advanced evaluation skillsets can help public administrators make more informed decisions about which evaluation designs to select. In other words, it should lead to more RCTs in the situations in which the RCT is the most appropriate evaluation design.

Factor 4: Demand

All else equal, as policy process increases demand for evidence, the more likely the public administrator is to use the RCT

Evaluation is largely an instrumental, not aesthetic, activity. Public administrators engage with and conduct evaluation to contribute to generation of useful knowledge about a program or policy. This can be for decision-making or accountability purposes. A key determining factor in the quantity and quality of policy evaluations is therefore how much demand there is for the evaluation outputs. Senior administrators, policy-makers, and the public need to be asking for the information an evaluation produces. We can think about this in terms of classic micro-economics: as the demand curve for evaluation shifts outward, there should be a compensating movement along the supply curve to meet the new level of demand. Overall, this results in more evaluations.

We would expect increasing demand for evaluation to be associated with the advancement of the evidence-based policy movement. A key aim of this movement is to ensure that public programs, services, and policies are based on a strong foundation of empirical evidence (Haskins, 2018). If successful, this movement should result in greater demand for evaluation, as policy-makers and the public look for and apply evidence to their support for program authorization and public funding. Greater demand for evidence should accelerate the production of evidence, including program and policy evaluation.

In theory, the evidence-based policy movement should increase demand for evaluations of all types. However, in its quest to base policy on “what works”, it may particularly increase demand for impact evaluations. Causal inference around program or policy impact is often what proponents of evidence-based policy have in mind when they talk about “evidence.” In this sense then, advancement of the evidence-based policy movement should reduce aversion to and increase support for impact evaluations, and RCTs by extension.

An additional constraint on the demand for evaluation may be the timelines of the policy-making process itself. While every evaluation design is different, robust evaluations do take some time to run. If the results do not become available before they are needed to feed into a decision, then the evaluation’s usefulness is voided. This means that there is an incentive towards evaluations that produce quick results. In theory, this shouldn’t necessarily shift a public administrator towards any particular evaluation type, as any given process evaluation or impact evaluation could be of shorter or longer duration. In practice though, it likely leads towards process or impact evaluations where the outcomes are shorter-term. In other words, while this “timeliness” incentive shouldn’t increase RCT aversion, it probably means that public administrators will prefer RCTs that use shorter-term outcomes rather than longer-term. For example, an RCT that studies which outreach strategy effects how many people sign up for a program is likely to be preferred to an RCT that studies whether that program influences educational attainment.

In practice, however, the evidence-based policymaking movement’s success has been more limited. The field has acknowledged that despite increasing production of evidence since the 1950s, this evidence is not often used to guide public decision-making. To the extent that it is used, its often after the fact to legitimize decisions that have already been made. Shulock (1999) finds evidence for the view that policy analysis and evaluation may be used more as an important input in the democratic process than a problem-solving tool. In other words, rather than being used by public administrators to choose among alternative and competing policies, policy evaluation has value in framing discourse, providing justification for action, and serving as a symbol for legitimate processes. In essence, evaluations and evidence are used as a signaling technique to make policy arguments or rationalize decisions, rather than an input in traditional conception of a multi-stage policy-making process.

Overall, using evaluation and evidence for their signaling value should still increase demand, but it’s likely to change what gets evaluated. For programs and policies that have widespread support, and do not need additional legitimization, there is little value or need for evaluation. However, supporters of programs and policies that do face resistance or skepticism are likely to push for evaluation to win support. The result is that new or emerging programs or policies are the most likely to be evaluated, while longstanding and uncontroversial programs are not given a critical eye.

In summary, increasing demand for evidence is likely to reduce RCT aversion in two ways. First, the evidence-based policymaking movement increases the value of an RCTs outputs as an input to decision-making. A key factor in how valuable an RCT is in decision-making relies on the timing of results. This "timeliness" factor likely incentives RCTs with shorter-term outcomes than RCTs with longer-term outcomes. Second, RCT evidence has increasing value as a signal in winning policy arguments and increasing public support for a program. However, this signaling factor likely incentives the use of RCTs for emerging or innovative policies and programs that do not already have widespread support or legitimacy. Programs or policies that already maintain popular support are not likely to be leading candidates for evaluation generally, and therefore impact evaluations and RCTs.

Factor 5: Requirements

All else equal, the more evaluation is required in statutes, regulations, and funding provisions, the more likely the public administrator will use the RCT

Another way to increase the demand and supply of evaluations is through mandates. Mandates will often take the form of funding conditions or "conditions-of-aid", where grants or budget funds are provided on the condition that evidence is provided that dollars were used as intended (Massey & Straussman, 1985). Congress can attach mandatory evaluation provisions through the appropriations process. Mathematica and the Urban Institute's evaluation of the State Children's Health Insurance Program is one such example (Hill et al., 2003). External philanthropic funders can also make evaluation a condition or focus of their grant funding, like Arnold Ventures does to fund RCTs of social programs.

One example of the power of public mandates in incentivizing evaluation and RCTs is the Food and Drug Administration (White Junod, n.d.). The 1938 Food, Drug, and Cosmetic Act required that new drugs undergo evaluation for clinical safety prior to market approval. While this law did not direct the specific tests required for approval, regulators could negotiate the study and approval requirements. These requirements were tightened with the 1962 Drug Amendments in response to the worldwide thalidomide disaster, which had caused severe birth defects and infant deaths after samples had been given to mothers without a disclaimer that the drug was experimental. The new regulations increased the testing standards for new drug approvals and required substantial evidence of both efficacy and safety. Since 1962, large, well-controlled randomized trials have become the default for approval of new drugs and clinical practices.

Currently, there is no similar mandate or regulatory requirement for the evaluation of social welfare programs in the United States. The 2018 Foundations for Evidence-Based Policymaking Act (Evidence Act) is potentially a step in the direction of wider endorsement of program evaluations, with the requirement that major federal agencies appoint Chief Evaluation Officers and develop learning agendas that identify priorities and needs for evidence (Hart & Davis, 2017). This law should address barriers to evaluation at the federal level, leading to more evidence while protecting data confidentiality and privacy. In theory, this should lead to more impact evaluations and RCTs of federally-funded programs. How much impact the Evidence Act has in practice is yet to be determined.

Reporting and evaluation mandates can range in their specificity and the rigor required. A funding report that simply requires an account of how funds were spent will incentivize an evaluation with a process element, measuring inputs and outputs. A funding report that requires an account how funds have made a difference towards a specific goal or outcome incentivizes impact evaluations. Like all command-and-control regulations, an evaluation mandates sets the floor for the quantity and quality standards (Guesnerie & Roberts, 1984). Public administrators must produce at least as much evaluation to meet the requirement. However, once the standard outlined in the mandate is met, public administrators do not have any additional incentive to improve the evaluation. This risks making program evaluation a “box checking” activity, in which public administrators authorize evaluations to satisfy requirements, but do not actually use the products of the evaluation.

Additionally, mandates tend to lack flexibility and differentiation. In other words, the more specific the mandate, the more it risks inappropriately treating all evaluation contexts the same. For example, there may be some situations in which a quasi-experiment may be less expensive than an RCT, with relatively little tradeoffs in validity. A mandate requiring the use of the RCT may therefore be less optimal than a more flexible mandate simply requiring the use of an impact evaluation.

In summary, mandates such as regulatory or funding requirements can provide an external incentive for public administrators to conduct RCTs and overcome RCT aversion. The more specific the mandate is for impact evaluations and RCTs, the more we would expect RCTs to be used. Depending on the regulatory context, mandates can be an extremely powerful tool in setting a “market-wide” default. A downside to mandates is that they generally will only incentivize the lower bound and have a difficult time incentivizing additional innovation or improvements. They can also constrain flexibility, potentially reducing optimization at the individual level.

Factor 6: Transparency

All else equal, the greater the risk of being transparent about outcomes or impact, the less likely the public administrator will be to use the RCT

Discussion around previous hypotheses has looked at the positive incentives public administrators face to conduct more evaluation and generate additional evidence. Very often, there isn't a great deal of demand for RCTs outputs, and therefore, public administrators are not well-incentivized to conduct them. Competing with demand for RCTs may be a strong disincentive: what happens when programs or policies are revealed as being less impactful than previously thought.

Agency theory has long outlined the complex relationship between a principal and their agent, and how mis-aligned incentives or differences in goals and attitudes can lead to problems (Eisenhardt, 1989; Ross, 1973). As an example, let us make the Chief Performance Officer (CPO) for a city the principal. Their agent is the Head of the Department of Public Works (DPW). The CPO wants to direct the Head of DPW to achieve certain goals and delegates the work. However, the CPO faces an information asymmetry, and cannot directly observe the Head of DPW's actions or clearly measure their impact on resident outcomes. Additionally, the interests of the CPO and the Head of DPW may not be fully aligned. While both parties may favor improved city infrastructure and resident safety, the Head of DPW also wants to maximize their job security and preserve or grow the department's budget and influence.

The information asymmetry and misaligned incentives lead the CPO and the Head of DPW to have different risk tolerances when it comes to program evaluation. Should an evaluation result in negative findings of a program or policy, the CPO will want to re-direct funds towards alternative uses within the city. Over time, this should result in better outcomes for the public and higher social welfare. The CPO, and other centralized city functions, generally stand to gain from more evaluation, as it reduces the information asymmetry around the outputs and outcomes of work, and provides a helpful input into funding and policy decisions.

However, for the Head of DPW, there is more risk involved and whether they stand to gain or lose depends on the evaluation's findings. Positive evaluation findings may lead to gains in the department head's influence or funding, as programs under their jurisdiction expand in scope and scale. However, negative evaluation findings may lead to losses for the public administrator, as their programs are curtailed or defunded. As explained previously, prospect theory holds that public administrators will be loss averse. Therefore, through a desire to avoid losses, public administrators are also likely to be risk averse and potentially evaluation-avoidant.

It seems reasonable to expect that the public administrator's risk aversion around negative evaluation findings is likely to be more significant towards impact evaluations than other types of program evaluation. Providing clear information about a program's inputs and outputs is more directly within a public administrator's control and influence. The impact around social welfare outcomes is considerably more complex. Despite good intentions and hard work, a perfectly implemented program may turn out to be ineffective in changing outcomes such as health and wellbeing. Impact evaluations, therefore, are likely to inherently carry more risk for public administrators. This, therefore, is likely to exacerbate impact evaluation aversion and result in fewer RCTs.

One way to overcome the inherent tension between the principal's desire for accountability and the agent public administrator's incentive to limit the risk of transparency is to better align incentives. For example, funders can increase the rewards for conducting evaluations, and impact evaluations in particular, by awarding larger budgets to programs with an evaluation component. Providing incentives upfront for simply implementing an evaluation, may help counterbalance the perceived risk of getting a negative finding. In other words, funders can help “bring forward” the benefits of evaluation in the public administrator’s risk calculation. In a study on agency controls and task performance, Fong and Tosi found incentive alignment to be more effective than monitoring in increasing work performance (2007). In other words, it is likely to be more effective to try and better align the interests of the principal and agent, rather than attempting to overcome the information asymmetry through more monitoring of activities.

Overall, there is an interesting balance to be struck between increasing accountability of public administrators and limiting the risk they face in being transparent about programmatic outcomes and impact. If agent public administrators are a single source of information to a principal accountability or performance officer on how public programs and policies are performing, the greater their control in what evaluations are conducted and when. If the principal’s decision-making about future investments is likely to be highly influenced by negative evaluation results, the greater risk the public administrator faces and the greater the disincentive to evaluate at all. Whether it’s a loss to professional reputation or programmatic funding, public administrators are likely to weigh losses more than equivalent gains. Risk aversion is therefore likely to lead to evaluation aversion of all types, and fewer RCTs overall.

Factor 7: Optimism

All else equal, the less the public administrator is overconfident about the intervention’s impact, the more likely the public administrator will be to use the RCT

As explained previously, public administrators, like all people, are likely to be risk averse. The hypotheses above explored how risk aversion around an evaluation’s results may contribute to RCT aversion. However, a public administrator’s risk calculus may also be affected by a miscalibration around the potential benefits or harms of interventions. In general, I assume that while the public administrator's interests will not be fully aligned with every public stakeholder, it is not in the public administrator's interest to cause the public injury. Public service motivation, and its related ethics and values, should result in the public administrator's desire to maximize potential benefits to the public and minimize harms.

Optimism bias may potentially lead public administrators to be overconfident about the potential benefits of their interventions. The role of overconfidence in decision-making has been noted as far back as Adam Smith, who wrote in The Wealth of Nations,

> "The over-weening conceit which the greater part of men have of their own abilities is an ancient evil remarked by the philosophers and moralists of all ages. Their absurd presumption in their own good fortune has been less taken notice of. It is, however, if possible, still more universal. There is no man living who, when in tolerable health and spirits, has not some share of it. The chance of gain is by every man more or less overvalued, and the chance of loss is by most men undervalued, and by scarce any man, who is in tolerable health and spirits, valued more than it is worth" (Smith, 2005, p. 93).

The modern concept of optimism bias comes from psychology and that predicts that people consistently overestimate their chances of success compared to failure (Sharot, 2011; Weinstein, 1980). It's hypothesized that this stems from the greater positive emotions around feeling in control over outcomes that affect us. Research has shown that people are more sensitive to evidence of success, updating beliefs more in response to positive information. Sharot et al. (2011) asked participants to estimate the likelihood of several different bad life events, and then were given the true statistical likelihood of the event and asked to re-estimate their personal likelihood. They found that if their initial personal estimate was lower than the statistical likelihood, the revised estimates would change very little. On the other hand, if their initial estimate was higher than the statistical likelihood, their revised personal estimate would decrease.

When applied to public administrators, optimism bias potentially leads them to be over-confident about the positive impact their programs and policies are having and less sensitive to negative information about potential harms or downside risks. This phenomenon is likely to have an interesting interaction with evaluation and RCT aversion. First, this may lead public administrators to put less value on the information coming from evaluations overall. If public administrators see evaluations as simply confirming what they already believe to be true, then there is little value added from conducting the evaluation. Research ethics hold that an RCT should only be used if there is substantial uncertainty about the relative value of one treatment versus another- known as the uncertainty principle (Djulbegovic et al., 2000). If public administrators don't think there is uncertainty about the intervention being superior, they won't view the RCT as an ethical evaluation design.

At a system level, fewer evaluations overall can reinforce the idea that interventions are generally helpful, as there is little evidence to counteract this belief. Selection around what evidence is shared publicly can further exacerbate this positive feedback loop. It has been widely noted that there is likely a publication bias around scientific findings, with only large and statistically significant findings making their way into academic journals (DellaVigna & Linos, 2022; Franco et al., 2014). Politically, public administrators face a disincentive from sharing evidence that programs are not achieving their goals, as noted in previous sections. Overall, this can reinforce the expectation that most social welfare programs are beneficial, which then feeds over-confidence further.

Risk aversion potentially adds an additional layer of complexity. While public administrators may be over-confident about the potential success of their interventions, they are also likely to be risk averse about potential innovations. Social welfare interventions that are clearly potentially risky, in that there is both an obvious chance of success and failure, are at a disadvantage compared to social welfare interventions where there is less perceived downside risk. Overall, this can create a dynamic where the programs and policies put forward by public administrators are safe but not optimally efficacious. In other words, risk aversion incentivizes public administrators to choose interventions where the perceived risk of harm is low, but the potential for maximizing gains and impact may also be low. This could potentially lead to a general mediocrity in what social welfare policies and programs are attempted, and further limits the perceived value of impact evaluations to demonstrate efficacy.

In summary, optimism bias and the miscalibration around the potential effects of social welfare interventions are likely to exacerbate RCT aversion. However, this is likely through decreasing the perceived value of evaluations and impact evaluations overall, rather than being targeted at RCTs specifically. While optimism bias likely makes the public administrator less fearful of negative evaluation results, it also leads them to believe that the evaluation will simply confirm what they already know. Fewer evaluations and less public information about potential null or negative findings worsen the problem, by reinforcing the expectation that most social welfare interventions are likely to be positive. Risk aversion further incentivizes public administrators to be wary of innovative interventions where less is known about the potential impact, and where the upside and downside risks could be large. Overall, this potentially creates a dynamic where RCTs are not valued and are therefore avoided.

Factor 8: Professional incentives

All else equal, the greater the professional rewards, the more likely the public administrator will be to use the RCT

Previous hypotheses have discussed the instrumental benefits and risks a public administrator may face in deciding whether or not to conduct an evaluation, impact evaluation, or RCT. Another group of incentives is around the professional rewards they may receive, contingent on running an RCT. The literature on work motivation distinguishes between two types of professional rewards: extrinsic and intrinsic.

Extrinsic rewards are incentives from external sources and are contingent upon a level of performance and achievement or accomplishment of a particular behavior (Bénabou & Tirole, 2003; Deci et al., 1999). They may be tangible, such as a performance bonus, or intangible, such as praise or recognition from peers or supervisors. Extrinsic rewards also include the avoidance of a disincentive or sanction, such as criticism for poor performance. With regard to evaluation, extrinsic rewards for the public administrator may look like a boost to their reputation, potentially with a news profile or being featured as a speaker at a conference. Depending on the performance management approach of the organization, their advancement may also be accelerated if they prove to be accomplishing key goals and objectives.

In reality, it is not likely to be common that a government organization has clear incentives for public administrators to evaluate their programs and services. Public service motivation theory and the New Public Management movement describe how difficult it is to create clear, objective criteria for performance evaluation in the public sector (Boruvka & Perry, 2020; B. E. Wright, 2001). Ideally, public administrators would be held accountable for programmatic impact, but most performance management is constrained to measure inputs and outputs.

One approach is to incentivize public administrators to simply conduct evaluations. The What Works Cities program is an example of this method (Bloomberg Philanthropies, 2017). While the public administrators of participating cities do not receive monetary compensation, they are able to access a network of peers and work towards levels of certification in several areas of data-informed governance, including evaluation. Achieving certification is viewed as a win for mayors, city managers, and other senior city officials, who are then able to celebrate the city's advancement with the public. Certification is seen as providing legitimacy and credibility to the city administration and signals national leadership in advancing data and evidence-based decision-making.

I hypothesize that as public organizations and community stakeholders increase the extrinsic rewards for public administrators to conduct evaluation, the more public administrators will use evaluations, impact evaluations, and RCTs to study social welfare interventions. The more targeted these rewards are towards impact evaluations or RCTs, such as incentivizing a certain standard of evidence around causal impact, the more the public administrator is likely to choose the RCT. The absence of these professional rewards means that the public administrator faces fewer gains in choosing the RCT approach.

Intrinsic rewards are incentives that emerge internally. As compared to extrinsic rewards, they are completely intangible and are generally linked to a sense of mastery and self-determination (Deci, 1976). As described by Ryan and Deci, intrinsic motivation refers to the tendency "to seek out novelty and challenges, to extend and exercise one's capacity, to explore, and to learn" (Ryan & Deci, 2000, p. 70). Someone who chooses to engage in an activity simply because they find it interesting, rewarding, pleasurable, or satisfying are intrinsically motivated.

Whether a public administrator will be intrinsically motivated to pursue evaluations, impact evaluations, and RCTs is likely to be influenced by many personal factors. Personal development, past experiences, and personality will all play an important role in determining whether evaluation will be a passion or interest. As described previously, training and education around program evaluation, causal inference, and the scientific method generally is likely to increase a public administrator's perceived value of evaluation. This exposure may also spark a public administrator's intellectual interest in the activity, or a poor educational experience may turn them off entirely.

Personality is likely to have a role in determining whether a public administrator will find evaluation intrinsically motivating. For example, the Big Five or Five Factor Model is currently the leading representation of personality traits or characteristics. It has been used to study the role of personality in a range of opinions, behaviors, and outcomes, such as political affiliation and academic achievement (Caspi et al., 2005; Costa & McCrae, 1992; Gosling et al., 2003). The model categorizes personality traits into five groups: conscientiousness, emotional stability, extraversion, openness to experience, and agreeableness (McCrae & Costa, 2008).

With regard to evaluation, one might expect that those who are high on "openness to experience," who are generally curious and eager to try new things, might find that evaluations help them satisfy a drive to learn and innovate. Similarly, highly "conscientious" individuals, who tend to be careful and methodical, might find that evaluations satisfy a personal interest in validating and ensuring their work is as effective as possible. Those low on these personality dimensions might find little satisfying or attractive about evaluations or RCTs. This doesn't necessarily mean they will be evaluation or RCT averse, but rather they will need to be motivated by other factors.

Self-determination theory suggests there are three basic psychological needs that drive intrinsic motivation: autonomy, competence, and relatedness (Ryan & Deci, 2017). Autonomy is the feeling of choice, integrity, and control over one's behavior, rather than being externally compelled or pressured. Competence is feeling of efficacy, experiencing increasing mastery of challenges, and developing one's capabilities. Relatedness is the feeling of connection and belongingness with others. When these three needs are supported, intrinsic motivation can follow. If these needs are undermined, intrinsic motivation is likely to suffer (Di Domenico & Ryan, 2017).

As applied to evaluation and RCT aversion, self-determination theory predicts that public administrators will be intrinsically motivated to evaluate if they are given a sense of choice over whether to evaluate, if conducting the evaluation is seen as an optimal challenge that is developing their capabilities, and if it supports their sense of connection with others, especially peers. This suggests that external incentives and requirements that coerce or compel public administrators to evaluate may come at a cost to intrinsic motivation. If public administrators feel that evaluation and RCTs are far beyond their current competence level, their intrinsic motivation is likely to suffer. Finally, evaluations that require teams to work collaboratively, rather than independent activity, are likely to increase a sense of relatedness and increase intrinsic motivation.

Overall, professional rewards are likely to be a strong motivator for public administrators to conduct more evaluations, impact evaluations, and RCTs. Extrinsic rewards, whether it is recognition and praise or advances in progression, can be used to incentivize the use of RCTs and overcome RCT aversion. However, extrinsic rewards need to be deployed carefully to ensure they do not become overly coercive and damage the public administrator's sense of autonomy, and therefore intrinsic motivation. Aspects of personality and experience are also likely to influence a public administrator's curiosity, interest, and intellectual inclinations.

Factor 9: Ethical norms

All else equal, the greater the perceived conflict with public values, the less likely the public administrator will be to use the RCT

Previous hypotheses have explored factors that may lead the public administrator to either prefer the RCT as an evaluation approach, or to contribute to greater RCT aversion. Most of these factors are not unique to the RCT methodology, as many will also lead public administrators to be evaluation or impact evaluation averse. In other words, what may look like opposition towards RCTs may be rooted in a reluctance to evaluate in general, or to be skeptical about impact evaluations in particular. However, one factor that is likely to contribute uniquely to RCT aversion is a perceived conflict with public values, and specifically distributive justice norms.

Social norms are the set of rules that govern the behavior of groups and societies, and generally fall into two common types: descriptive and injunctive (Bicchieri & Muldoon, 2011; Cialdini et al., 1991). Descriptive norms are the shared perceptions of what people do, or what is typical or usual. Injunctive norms are the shared beliefs of what people ought to do, or what should be expected or ideal. They are the rules we expect others and ourselves to follow, even if those rules may hinder our own immediate interest. Bicchieri proposes that injunctive social norms are maintained through external sanctions, by the threat of social punishment, and internal motivation, in which people develop a desire or need to conform (Bicchieri, 2005; Bicchieri & Muldoon, 2011).

Distributive justice norms are a particular type of social norm, pertaining to what people believe is fair and just behavior. They are specifically focused on how people believe goods, rewards, or costs should be shared or distributed across members of a group. While principles are likely to vary in importance depending on the context, prior work has identified a core set of principles that are applied across domains.

Scott et al. (2001) identify these four main allocation principles as contribution, equality, need, and efficiency in their review of the prior literature. Contribution (or merit) holds that allocation of goods should be proportionate to individual contributions. Equality holds that allocation should be absolutely equal across the group. Need holds that allocation should be based on individual need, whether relative or absolute. Finally, efficiency holds that allocations should maximize the overall amount of good holding input constant.

Decision-making about how goods, rewards, or costs should be allocated among a group can be challenging if principles are perceived to conflict or compete with one another (Lamm & Schwinger, 1980). For example, in making decisions around how income ought to be distributed, individuals must make tradeoffs between equality, equity (or merit) and efficiency. Scott et al. (2001) find evidence that people tend to prefer more equality over less, even at the expense of efficiency.

With regard to RCT aversion, public administrators may perceive that the RCT design presents a challenge to distributive justice principles. First, the control group element may be perceived to violate equality principles by deliberately treating equivalent people differently. While public administrators are often required to treat similar people differently in their jobs, this is typically caused by administrative factors or policy decisions. For example, an implementation date for a program or a sign-up deadline may be equally arbitrary as randomization but more widely accepted reasons for differential treatment. Public administrators may perceive evaluation to be a less satisfactory rationale. In the end, every policy or program will need an implementation date, even if the decision between the last day of one month and the first day of the next could be as good as random. But not every policy or program needs an evaluation.

Secondly, the random assignment element may be perceived to circumvent distributive justice principles in general. Random assignment is, by definition, an arbitrary allocation mechanism. Its methodological strength lies in it being unrelated to other contextual, environmental, or individual factors. When applied to the distribution of a public good or access to a public service, public administrators may find this unrelatedness to be difficult to justify. Even though the intervention may be unproven, public administrators may find random allocation to violate distributive justice principles and norms.

Alternative evaluation approaches do not require the use of control groups or random assignment. For example, quasi-experimental approaches may use a comparison group as a means of estimating counterfactual outcomes (as in a difference-in-differences design) or use thresholds as allocation mechanism (as in regression discontinuity designs.) Many other evaluation designs are observational rather than experimental, allowing implementation to proceed naturally and without the intervention of the public administrator. Therefore, public administrators may not perceive these approaches to challenge distributive justice norms as strongly as the RCT, where they must deliberately withhold treatment from an eligible group based on chance assignment. I hypothesize that this tension between the requirements of the RCT design and moral or ethical principles about how public programs and services ought to be distributed is a specific contributor to public administrator RCT aversion.

Public service motivation may either exacerbate or lessen this perceived tension between principles or norms. Perry and Hondeghem (2008) define public service motivation (PSM) as the “motives and action in the public domain that are intended to do good for others and shape the well-being of society" (p. 3). If public administrators are more attentive to distributive justice principles, PSM may make the RCT's perceived violation of these norms even more problematic. However, if public administrators are more attentive to other injunctive norms, they may perceive there to be less tension with the RCT. For example, The American Society for Public Administration promotes a Code of Ethics that lists eight principles to guide the behavior of its members. These principles include "strengthen social equity" but also to "fully inform and advise." If a public administrator believes that the RCT is the strongest evaluation design and will best inform decision-making, PSM may lessen perceived tension between the RCT design and ethical norms.

In summary, if public administrators perceive that the requirements of the RCT are in competition with norms of distributive justice, there is a disincentive to conduct them. The greater this perceived conflict, the stronger the RCT aversion. Other impact evaluation methods, and evaluation more generally, are not likely to present the same tradeoffs, and therefore, public administrators are not likely to see these other methods as challenging injunctive norms. Therefore, tension with public values around how goods and services should be distributed is likely to affect RCTs in particular and create an incentive to substitute towards alternative evaluation methods. If a wider ethical perspective increases the perceived alignment between the RCT and the public administrator's duty and responsibility, then RCT aversion may be lessened.

Conclusion

This chapter has explored the public administrator's decision-making process as they contemplate whether to engage in program evaluation, whether to choose an impact evaluation, and whether that impact evaluation should be an RCT. The public administrator is likely to choose the option that best maximizes their utility. However, their decision-making process around utility maximization is likely to be more of a satisficing rather than optimizing approach. They are also likely to consider the welfare of the collective alongside their individual best interests. Uncertainty is also likely to make the public administrator hesitant to choose options that deviate from the status quo and bear some downside risk.

The nine factors described above demonstrate that there are very few clear incentives to conduct evaluation in general, let alone impact evaluations and RCTs. Organizations that have a previous history conducting RCTs, complete with the capability, infrastructure, practices, and norms that support evidence-generation, are more likely to favor RCTs in the future. At the same time, there are several clear disincentives to conducting evaluation, impact evaluation, and RCTs. The realities of the policy process, the risks of transparency around programmatic outcomes and impact, general overconfidence about the likely benefits of social welfare interventions, a lack of explicit professional rewards, and perceived conflicts in public values are all likely to contribute to greater RCT aversion.

Risk aversion makes this dynamic worse, as it leads public administrators to weigh the disincentives more than the incentives as they make their decision. In other words, the costs will be more influential than equivalent benefits. To move forward with the RCT, the public administrator must believe that the benefits far outweigh the potential risks. Therefore, RCT and evidence-based policy advocates likely face an uphill, but not insurmountable, task in boosting the incentives to conducting impact evaluations and RCTs and minimizing the disincentives.

It is important to note that some of these factors are more exogenous to the public administrator than others. When understanding their preferences and making a choice around an evaluation option, the public administrator may face constraints or an environment that has been shaped by others. For example, costs, the current investments and capabilities of the wider organization, and the demand for evaluation and evidence are all external to the public administrator. Other factors are more endogenous: internal motivations, beliefs, and overconfidence. These factors may also affect one another over time. For example, the endogenous factor of internal motivation is likely to determine how the public administrator invests in the organizational infrastructure and their own evaluation capabilities, which then influences evaluation costs in a later period. Isolating the factors from one another is far from simple in practice.

While this chapter has pulled from the wider literature on the public sector and decision-making, there is still much we do not know about how public administrator's decide to engage in program evaluation. The outlined hypotheses set out a potential research agenda for the public administration field to better understand the role of evaluation in public organizations and the public administrator's evaluation decision-making process. Advancing this research agenda will help clarify which of the outlined factors are most influential in the public administrator's evaluation decisions, and the degree to which these factors are specific to individual situations or are part of a wider organizational or collective context. If we believe the claims of the evidence-based policy movement, and think that more programs and policies should be based in evidence around what works, understanding the public administrator's perspective and decision-making process is key to unlocking more evaluation, impact evaluation, and RCTs in public organizations.

Chapter 3. Survey Experiment

Introduction

This dissertation examines whether public administrators are averse to randomized controlled trials (RCTs). More specifically, are public administrators reluctant or opposed to conducting RCTs for the evaluation of social welfare programs in situations where they would both be beneficial (i.e., would answer an important or useful policy question) and feasible (i.e., ethical and practical to implement)? If they are RCT averse, what are the underlying reasons?

Previous chapters have explored the role of RCTs in the evidence-based policy movement, our current understanding of experiment aversion or the A/B effect, and a list of factors for RCT aversion among public administrators. This chapter turns to an empirical test of one of these hypotheses: that one source of RCT aversion lies in the challenge it presents to distributive justice norms.

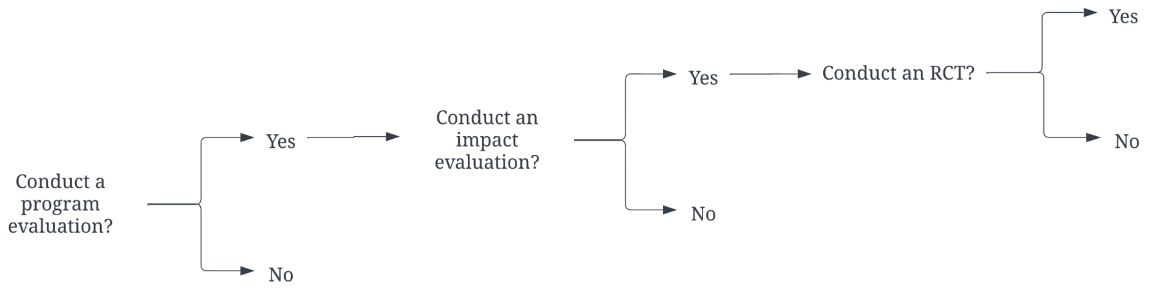
First, a return to the evaluation decision tree. As previously noted, this model is simplistic, in that it only presents the line of decision-making that leads to the RCT, rather than all of the evaluation options open to a public administrator. Similarly, it is likely that no public administrator follows this decision tree systematically, making separate and sequential choices. Rather, decision-makers are often likely making joint decisions when selecting an evaluation path.

Regardless, the evaluation decision tree is useful for setting up clean contrasts for testing the sources of RCT aversion. A public administrator may be averse to running an RCT for the same reasons they are averse to using a quasi-experimental design. I am interested in understanding whether people, and public administrators specifically, are averse to the RCT in particular, and why.

As a brief summary, the choice set facing a public administrator in deciding whether and how to conduct program evaluation can be modeled as follows:

- Choice 1: Whether to conduct program evaluation, or no evaluation
- Choice 2: Whether to conduct an impact evaluation, or a different type of evaluation
- Choice 3: Whether to conduct a randomized controlled trial, or a different type of impact evaluation - such as a quasi-experiment

Figure 2: RCT Decision-Tree



This empirical study focuses on the third node of the decision tree: the public administrator has decided to conduct a program evaluation, and the purpose of the evaluation will be to study the program's impact. The public administrator now faces a decision between conducting a quasi-experiment and a randomized controlled trial. If the public administrator demonstrates a consistent preference for the quasi-experiment, this is evidence for RCT aversion.

As discussed in the previous chapter, one potential explanation for RCT aversion is that individuals perceive the randomized controlled trial design to violate distributive justice norms. While the definition may vary across academic disciplines, social norms are generally taken to mean the set of rules that govern the behavior of groups and societies (Bicchieri & Muldoon, 2011). Cialdini et al. (1991) describe two common types of social norms: descriptive and injunctive. Descriptive norms are the shared perception of what people commonly do. Injunctive norms are the shared beliefs of what people ought to do. In other words, they are the rules that people expect others to follow, even if it runs counter to their own immediate interests. Bicchieri (2005) theorizes that injunctive norms are maintained both by the threat of sanction, or social punishment should the norm be violated, and by internalization, in which people develop a psychological motive to conform (Bicchieri & Muldoon, 2011).

Distributive justice norms are a particular type of social norm, and govern the domain of what people believe to be fair and just. They specifically concern perceived fairness for how goods, rewards, or costs are shared across members of a group. Recent theoretical and empirical work has found evidence for a pluralistic view of justice (Scott & Bornstein, 2009). These frameworks propose that individuals draw on several distinct principles when making allocation or distribution decisions, and apply these principles according to distinct and predictable factors. Walzer's "spheres of justice" is the most influential/well-known of these pluralistic theories (2008). He proposes that there are different spheres of society in which different distributive justice norms should rule. Elster's "local justice" approach (1993) and Miller (1991) expand on this framework, hypothesizing that in addition to the type of good being distributed, policy arena and mode of human relationships also contribute to decision-making. What all pluralistic theories have in common is the idea that views of "what is morally right" is likely to be context-dependent.

Most prior work has identified a core set of principles as most important across domains. Scott et al. (2001) identify four main allocation principles in their review of the prior literature. Contribution (or merit) holds that allocation of goods should be proportionate to individual contributions. Equality holds that allocation should be absolutely equal across the group. Need holds that allocation should be based on individual need, whether relative or absolute. Finally, efficiency holds that allocations should maximize the overall amount of good holding input constant.

Empirical work has documented several contextual factors that are likely to influence allocation decision-making. These include the relationship between parties, expectations of future interaction, the allocation objective, the environment or institutional setting, and whether the allocation decision is going to be disclosed (Scott & Bornstein, 2009). In addition to context, the nature of the good being distributed and individual factors, such as gender and ideology, have been found to matter in allocation decision-making (Scott et al., 2001).

Allocation decision-making can be challenging for people when principles conflict or compete with one another (Lamm & Schwinger, 1980). For example, in making decisions around how income ought to be distributed, individuals must make tradeoffs between equality, equity (or merit) and efficiency. Scott et al. (2001) find evidence that people tend to prefer more equality over less, even at the expense of efficiency.

I hypothesize that individuals may be RCT averse because it presents a challenging tradeoff in distributive justice principles. First, in the use of a control group, it violates equality principles by deliberately treating equivalent individuals differently. Creating a parallel control group requires withholding a new policy intervention from some people. Even if this treatment is unproven (and could potentially have negative consequences), public administrators may dislike treating people differently. If they are forced to treat people differently, they may feel they should not do so solely for evaluation purposes, rather than for other administrative or policy reasons.

Secondly, the use of random assignment is, necessarily, an arbitrary allocation mechanism. Rather than allocating a public good or service according to merit or need, access is instead determined by random chance. Randomness is important specifically because it is entirely unrelated to environmental or individual characteristics. Yet this “unrelatedness” may be perceived to violate distributive norms. In other words, the very reason that the RCT is so strong as an evaluation approach is potentially the same reason it is viewed as distasteful in public policy. While researchers may argue that the benefit to the quality and integrity of the evidence outweighs the harm, especially for only a short period of time, public administrators and the general public may not agree. This may be because they do not appreciate the degree to which the RCT design can increase confidence in the results, or they may not find the break in distributive justice norms to be as tolerable.

Quasi-experimental designs share many of the same features as RCTs. Quasi-experiments also make comparisons across groups, in which one group has received the policy intervention and the other has not. The strongest quasi-experimental designs use selection mechanisms that are equally arbitrary to achieve “as-if” randomization. However, a key difference between the two approaches is that the “violations” in the distributive justice norms for quasi-experiments are either less salient, are the result of factors that are unrelated to the evaluation, or are not under the control of the policy maker or administrator. For example, while the RCT explicitly calls for a control group, quasi-experiments do not require the public administrator to deny an intervention to an equivalent group of people. Instead, quasi-experimental designs are able to leverage comparisons across time (as in a time-series design) or between groups that differ on some dimensions (as in a difference-in-differences or regression discontinuity design.)

This leads to my first hypothesis: given the choice between an RCT and a quasi-experiment for the evaluation of a social welfare intervention, people will prefer the quasi-experiment. This is holding the evaluation’s cost, number of people included and treated, and time period constant. Specifically, when the allocation mechanism and comparison group composition are the only two factors that differ between the evaluation approaches, people will prefer the evaluation approach that best aligns with distributive justice norms - a design that prioritizes allocation on the basis of merit, need, or equality rather than random chance.

My second hypothesis is related, and predicts that RCT aversion will be greater in scenarios where the gap in treatment between the "haves" (the treated group) and the "have nots" (the control or comparison group) are greater. In other words, the preference for non-randomized evaluation designs for the study of social welfare interventions will become more pronounced as the "treatment gap" between groups is exacerbated, or becomes wider. Contextual factors that cause similar people to be treated differently are likely to violate distributive justice norms to an even greater extent. Additionally, these contextual factors may make arbitrary allocaction mechansims, the process determining will be the "haves" and "have nots," even less tolerable.

Similar to findings by Arno et. al (2020; Scott & Bornstein, 2009), I predict that this perceived gap will be influenced by features of the policy environment. For example, how intensive the new intervention is, whether there is a current intervention already in place, and how quickly the new intervention could be scaled after the results come in. The more the public administrator perceives the control group to be disadvantaged in comparison to the treatment group, the greater their potential objections to the RCT and preference for the quasi-experiment.

Finally, my third hypothesis is that RCT aversion is likely to be influenced by individual-level factors. Aligned with previous research, I predict the preference for the RCT to vary systematically by personal characteristics, such as educational attainment (Lamm & Schwinger, 1980; Scott et al., 2001). Specifically of interest is whether public administrators have different preferences for the RCT or quasi-experiment compared to their non-public administrator peers.

Differences between public administrators and non-public administrators may result from two sources. First, there may be systematic differences between the two groups based on demographic or socioeconomic characteristics, experience, or personality. For example, public administrators may be more likely to have greater educational attainment or personal experience with evaluations. If public administrators are more likely to have been exposed to policy evaluation - either through their schooling (such as attending a masters program in policy or public administration) or through their jobs (such as participating in an evaluation), then they may be less RCT averse (or more RCT inclined). While they may share this tendency with equivalent peers in the private sector, public administrators may appear to be less RCT averse as a group on average compared to the general public.

A second source of potential differences between public administrators and non-public administrators (or "the general public") may be the result of different attitudes or distributive justice norms. Public service motivation theory hypothesizes that the work behavior and performance of individuals in public organizations is likely to be more influenced by self-determined motivational factors, such as intrinsic motivation and moral obligation, then external factors like monetary incentives (T.-M. Wang et al., 2020). Perry and Wise initially defined the concept as "an individual's predisposition to respond to motives grounded primarily or uniquely in public institutions and organizations" (1990, p. 368). This conception has been broadened in the decades since, with Perry and Hondeghem (2008) defining it as "motives and action in the public domain that are intended to do good for others and shape the well-being of society" (p. 3). Ritz et al. (2016) define public service motivation as "the desire to behave in accordance with motives that are grounded in the public interest in order to serve society" (p. 9).

Much of the literature has been focused on understanding how public service motivation may lead to differences between public-sector employees than their private-sector counterparts (Crewson, 1997). If public administrators have a greater internalization for the desire to act in the public interest, this may lead public administrators to feel that the RCT violates principles of distributive justice even more strongly than the general public. This potentially can lead to biased judgments. For example, in a lab experiment, Prokop and Tepe (2020) found that individuals with higher attraction for public service used unnecessarily excessive sanctions to enforce norms around fairness. If this is the case, we may expect to see that public administrators are more RCT averse (and less RCT inclined) than their non-public administrator peers.

In summary, this study aims to three research questions consistent with these hypotheses:

- First, do participants prefer a quasi-experiment to the RCT?
- Second, do participants prefer quasi-experiments to RCTs more when features of the policy environment create a greater perceived difference in treatment between groups?
- Third, do preferences for the RCT differ by certain characteristics? Specifically, do public administrators have different evaluation preferences compared to their non-public administrator peers?

Methodology

Study Design

I use an online survey experiment to directly test these hypotheses. This experiment asks participants to perform a task similar to that of a policy maker or administrator, determining which evaluation approach to select in different contexts. Participants are given a policy scenario, in which a social welfare intervention has been proposed and requires an evaluation. After reading about two proposed evaluation options, participants are asked to select which is their preferred approach: the randomized controlled trial, or the quasi-experiment. Elements of the policy scenario are experimentally manipulated to see whether the "widening" of the treatment gap between groups causes changes in evaluation preferences, on average. The survey also collects demographic and socioeconomic information on the participants, to shed light on whether evaluation preferences are influenced by individual-level characteristics. By using two samples of interest, public administrator and "general public," allows comparisons between the two groups, providing evidence on how both public managers and citizens approach the evaluation decision.

This study uses a vignette factorial survey experimental design. A type of stated preference experiment, the aim of the approach is to uncover the systematic principles by which humans make judgements about social objects (Rossi & Nock, 1982). The setup is relatively simple. Participants are asked to choose from multiple descriptions of objects that vary along set dimensions. These dimensions or attributes are hypothesized to be important factors in the decision. Each of these dimensions has a set of values or levels, which are randomly varied across respondents. This allows the researcher to estimate the relative importance of each dimension to the resulting decision (Wallander, 2009).

The specific design of this study is closest to a forced choice paired conjoint design. Participants are presented with two evaluation options: an RCT design and a quasi-experimental design. These options appear next to each other in a "conjoint table," enabling easy comparison across dimensions (Hainmueller et al., 2014, 2015). Participants are then forced to choose which of the options they most prefer.

The strength of the design lies in combining the advantages of the survey with the realism of a decision task. Thus, it "provide(s) an effective, low-cost, and widely applicable tool to study human preferences and decision making" (Hainmueller et al., 2015, p. 2395). Respondents in factorial survey studies are generally not aware of the experimental manipulation, or the randomized elements of the vignettes. Therefore, there is less risk that the results will suffer from social desirability bias (Alexander & Becker, 1978). Additionally, people are not always aware of how certain factors affect their judgments. The factorial survey design allows researchers to observe the influence of carefully manipulated variables on the survey participants' resulting decision-making (Alexander & Becker, 1978).

Thus, while the scenarios included in this experiment depart from some of the realism of decision-making in true policy environments, it allows me to isolate how particular contextual factors influence evaluation choices. I can test for RCT aversion in the abstract, without the complicating factors of feasibility, political dynamics, or organizational constraints. The question answered by this study is arguably, therefore, a "purer" indication of the individual's preference: how the participant would choose in the absence of real-world pressures and influence.

A key criticism of survey experiments is that they lack external validity and suffer from many of the response biases as other survey research. Hainmueller et al. (2015) examine the external validity of the design directly, comparing results from conjoint and vignette analyses to data from natural experiment in the context of support for naturalization in Switzerland. They find that the survey experiments perform both qualitatively and quantitatively well, with the paired conjoint design coming closest to the natural experiment results. This performance is despite the paired conjoint design being the least similar to the real-life informational leaflets used in the referendums studied. The authors hypothesize that this is because the paired designs cause participants to be more engaged and less likely to satisfice. Fortunately, for my purposes, the paired conjoint design can replicate the real-life decision-making process of managers and administrators, who are often presented with different proposals set in contrast to one another.

A second benefit of the survey experiment design is that it allows me to capture typically unobserved characteristics, such as personality, that may be influential in the evaluation choice and conceptions of distributive justice norms. This allows me to explore potential mediating factors for RCT aversion. While it may be possible to assemble an administrative data set of evaluation decisions, it is difficult to link that to a rich dataset of decision-maker characteristics.

Finally, the online experiment allows access a large and diverse pool of participants in a cost-effective manner. Evans and Mathur (2018) document many of the strengths of online survey research, including flexibility, speed, ease of data entry, and control of answer order and completion. One of my central research objectives is to understand whether public administrators have different evaluation preferences than the general public. Sub-samples of sufficient size are required to enable this comparison. Conducting the experiment online allows me to access a panel of survey participants, screen them quickly into these two sub-samples of interest, and provide compensation for the task quickly and efficiently. Of course, online surveys also come with weaknesses, including the skewed attributes of the online population and potential inattention (Evans & Mathur, 2018).

Experimental Manipulation

The key component of factorial surveys are the vignettes, or the fictional descriptions of "social objects," usually either people or social situations (Alexander & Becker, 1978). These vignettes are composed of randomized combinations of values from different attributes or dimensions (Rossi & Nock, 1982).

In this experiment, survey participants are randomized to see one of eight (8) different policy scenarios or vignettes, each proposing a new policy intervention that must be evaluated for impact. This study features two potential lead screening and abatement interventions. I selected lead abatement as the social welfare policy area because lead abatement has been shown to have significant social welfare implications, with substantial returns to investing in targeted early interventions in communities most at risk (Gould, 2009). Additionally, the interventions are realistic for a city to implement and the risk of harm is likely to be minimal, but there is not evidence about the cost-effectiveness of these specific approaches (Results for America, 2020a, 2020b).

I test the influence of three different policy scenario attributes or dimensions on the likelihood of selecting the RCT as the preferred evaluation option. Each of the three policy scenario elements have two levels:

- Attribute 1: The "business-as-usual condition"
    - Level 1: The government has a current lead testing program in place that it will continue to offer to residents who don't receive the intervention
    - Level 2: The government does not have a current lead testing program in place, and therefore residents who don't receive the intervention are not offered anything comparable.
- Attribute 2: The intervention intensity
    - Level 1: The government is testing a lower-cost, "implementation-light" intervention. In this experiment, the low-cost intervention is mailing two free instant lead test kits to homes.
    - Level 2: The government is testing a higher-cost, "implementation-intense" intervention. In this experiment, the high-cost intervention is a "Healthy Home Assessment" visit by community health workers who conduct the testing for residents (RFA).

- Attribute 3: The scaling path
    - Level 1: If the intervention proves successful, the government will be able to offer the intervention to everyone immediately upon concluding the evaluation.
    - Level 2: If the intervention proves successful, the government will need to apply for more funding and resources to then be able to offer the intervention to everyone.

These three policy elements, with two levels, are matrixed to result in eight unique vignettes (2 x 2 x 2). Across all three dimensions, the two attribute levels are designed to test whether a larger treatment gap is associated with greater RCT aversion. In other words, there is a smaller difference in how the control (or comparison) group is treated compared to the treatment group in Level 1 attributes than in Level 2 attributes.

Table 1: Experimental Attributes

Smaller treatment gap (Level 1 Attributes)	Larger treatment gap (Level 2 Attributes)
There is a current program offered, control/comparison group gets something	There isn’t a current program offered, control/comparison group gets nothing
Less intense intervention	More intense intervention
The new program could be scaled immediately	The new program couldn’t be scaled immediately

Participants are randomly assigned to one of these eight vignettes using a random number generator on the survey platform. This results in simple randomization, where each participant has an equal likelihood of being assigned to each vignette. The limitation of simple randomization is that it can result in different observation numbers across scenarios, and it cannot ensure balance on covariates. The unit of assignment is the unique participant platform identification number.

Table 2: Experimental Scenario Vignettes

Scenario vignette	Attribute 1: Business-as-usual	Attribute 2: Intervention intensity	Attribute 3: Scaling path
1	Current program	Healthy home assessment	Offer to everyone
2	No program	Healthy home assessment	Offer to everyone
3	Current program	Healthy home assessment	Apply for more funding
4	No program	Healthy home assessment	Apply for more funding
5	Current program	Lead test kits	Offer to everyone
6	No program	Lead test kits	Offer to everyone
7	Current program	Lead test kits	Apply for more funding
8	No program	Lead test kits	Apply for more funding

All participants receive the same policy introduction. This text introduces Cityville and explains why lead is a policy priority and what causes residents to be at risk for lead exposure.

Table 3: Experiment Policy Introduction Text

Cityville wants to use federal funds to improve resident health.

Cityville housing is old, with most houses built between 1900 and 1950. Older homes are at higher risk for various environmental health hazards. These include lead, radon, and mold.

Lead is difficult to detect because it is odorless and invisible. It is also dangerous to children. It can cause:

- Damage to the brain and nervous system
- Slowed growth and development
- Learning and behavior problems
- Hearing and speech problems

Lead exposure is preventable. Residents need to first identify lead sources, and then either control or remove them.

Governments can help by providing access to education, testing, and resources. If implemented well, these lead programs work to improve resident health. They can also result in higher academic achievement and reduced crime.

Depending on their randomized assignment, participants then see different text explaining Cityville's current lead screening program and the intervention they hope to test.

Table 4: Experiment Program and Intervention Text

Business-as-usual: Current program Scenarios: 1, 3, 5 and 7 While Cityville has a lead screening program, it's not having much impact. The current program requires residents to submit an application for free testing, and less than 1 percent of Cityville households have applied.	Business-as-usual: No program Scenarios: 2, 4, 6, and 8 Cityville doesn't currently have a lead screening program.
Intervention: Mailed Test Kits Scenarios: 5, 6, 7, and 8 So Cityville wants to try mailing free test kits to residents. Families would receive two instant lead paint testing kits. These kits allow families to test surfaces and detect whether there is lead in seconds. Cityville hopes that this new program will reduce the number of children exposed to lead. They predict that it will cost an average of $25 per household.	Intervention: Healthy Home Assessment Scenarios: 1, 2, 3, and 4 So Cityville wants to try a new service, called a "healthy home assessment." Community health workers will visit resident homes and conduct free screening for health hazards, including lead. If the worker identifies an issue, they will also contain or fix it immediately. Cityville hopes that this new program will reduce the number of children exposed to lead. They predict that it will cost an average of $125 per household.

Participants then see the same evaluation context, which explains why Cityville needs to evaluate and what they hope to see the interventions achieve. It also explains that the evaluation options are expected to cost the same and include the same number of residents in the intervention or treatment groups.

Table 5: Experiment Evaluation Context Text

The federal funding requires Cityville to measure the program's impact. The City Manager wants to know whether the program worked to: ● Improve resident access to quality housing ● Improve resident health outcomes The City Manager's staff have developed two evaluation options. They are expected to cost the same and offer the new service to the same number of residents.

Depending on their vignette assignment, participants then see two different evaluation options. One is a randomized controlled trial, and the other is a differences-in-differences quasi-experiment. Both evaluation options start with the same sample: addresses taken from Cityville tax records. They also will take the same amount of time to conduct (1 year) and will use the same health outcomes. Vignette assignment determines whether the control group will continue to have access to a current lead abatement program or not, and whether the intervention being tested is the mailed test kits or healthy home assessment.

Psychological research has demonstrated that decision-making can be significantly influenced by small changes in framing (Tversky and Kahneman, 1981). With regard to survey research, this means that the presentation or wording of questions needs to be taken with care (Schuman and Presser, 1981). In particular, primacy effects can increase the likelihood that survey respondents select items at the beginning of a list (Krosnick and Alwin). To avoid this bias, I create survey variants that switch the order of the evaluation options. Survey variant 1 shows the evaluation options ordered as presented below with the RCT as "Option A." Survey variant 2 switched the order, presenting the quasi-experiment as "Option A" instead. There is equal likelihood that a survey participant is randomly assigned to survey variant 1 or 2.

Table 6: Experiment Evaluation Options Text

Business-as-usual: Current program Intervention: Healthy Home Assessment Scenarios: 1 and 3	
Option A ● Start with all addresses from tax records. ● **Randomly pick** one half to a treatment group and the other to a control group. ○ The treatment group will be offered the new healthy home assessment. ○ The control group won't be offered the new healthy home assessment. ● Continue to provide everyone access to the current lead abatement program: Any resident can apply for free testing. ● After one year, compare health outcomes for both groups.	**Option B** ● Start with all addresses from tax records. ● **Select neighborhoods** with the greatest number of older homes. ○ Selected neighborhoods will be offered the new healthy home assessment. ○ All other Cityville neighborhoods will be the comparison group. They won't be offered the new healthy home assessment. ● Continue to provide everyone access to the current lead abatement program: Any resident can apply for free testing. ● After one year, compare health outcomes for the different neighborhoods.
Business-as-usual: No program Intervention: Healthy Home Assessment Scenarios: 2 and 4	
Option A ● Start with all addresses from tax records. ● **Randomly pick** one half to a treatment group and the other to a control group. ○ The treatment group will be offered the new healthy home assessment. ○ The control group won't be offered the new healthy home assessment. ● After one year, compare health outcomes for both groups.	**Option B** ● Start with all addresses from tax records. ● **Select neighborhoods** with the greatest number of older homes. ○ Selected neighborhoods will be offered the new healthy home assessment. ○ All other Cityville neighborhoods will be the comparison group. They won't be offered the new healthy home assessment. ● After one year, compare health outcomes for the different neighborhoods.
Business-as-usual: Current program Intervention: Mailed test kits Scenarios: 5 and 7	
Option A ● Start with all addresses from tax records. ● **Randomly pick** one half to a treatment group and the other to a control group. ○ The treatment group will be	**Option B** ● Start with all addresses from tax records. ● **Select neighborhoods** with the greatest number of older homes. ○ Selected neighborhoods will be

mailed the free test kits. ○ The control group won't be mailed the free test kits. ● Continue to provide everyone access to the current lead abatement program: Any resident can apply for free testing. ● After one year, compare health outcomes for both groups.	mailed the free test kits. ○ All other Cityville neighborhoods will be the comparison group. They won't be mailed the free test kits. ● Continue to provide everyone access to the current lead abatement program: Any resident can apply for free testing. ● After one year, compare health outcomes for the different neighborhoods.
Business-as-usual: No program Intervention: Mailed test kits Scenarios: 6 and 8	
Option A ● Start with all addresses from tax records. ● **Randomly pick** one half to a treatment group and the other to a control group. ○ The treatment group will be mailed the free test kits. ○ The control group won't be mailed the free test kits. ● After one year, compare health outcomes for both groups.	**Option B** ● Start with all addresses from tax records. ● **Select neighborhoods** with the greatest number of older homes. ○ Selected neighborhoods will be offered the new healthy home assessment. ○ All other Cityville neighborhoods will be the comparison group. They won't be offered the new healthy home assessment. ● After one year, compare health outcomes for the different neighborhoods.

Finally, participants see language that explains how Cityville plans to scale the intervention, if the evaluation proves it to be successful.

Table 7: Experiment Program Scaling Text

Intervention: Healthy Home Assessment Scaling: Offer to everyone If the healthy home assessments are successful, Cityville will offer them to everyone.	Intervention: Healthy Home Assessment Scaling: Apply for more funding If the healthy home assessments are successful, Cityville will apply for more funding to keep the program running.	Intervention: Mailed test kits Scaling: Offer to everyone If the two free test kits are successful, Cityville will mail them to everyone.	Intervention: Mailed test kits Scaling: Apply for more funding If the two free test kits are successful, Cityville will apply for more funding to keep the program running.

Participants are then asked to select a preferred evaluation option. This is a forced choice, as participants are not given the option to not select an evaluation option.

Table 8: Experiment Evaluation Decision Text

Which evaluation option would you choose?

- Option A
- Option B

Sample and Setting

I aimed to recruit at least 4,000 total participants, with approximately 500 participants assigned to each survey arm. Within this 4,000, I aimed for two main subsamples: a subsample of approximately 2,000 public administrators, and 2,000 general public participants. The sample size was set to be powered to detect a 5 percentage point difference between the policy scenario elements. This assumed a baseline of 50%, or that participants would be equally likely to choose the RCT as the quasi-experiment, and conventional power of 80%. Alpha was set at .016 to adjust for the three comparisons.

This experiment uses a quota sampling method, a type of nonprobability sampling often used in market research and online surveys. Respondents are invited to participate in the survey by registering on an “opt-in” panel (Brick, 2014). The eligible population is divided into strata, or “quota controls”, such as gender, age, education, that are chosen by the researcher depending on the topic of study (Yang & Banamah, 2014). Each quota or strata is given a target percentage, which is then used to ensure that the sample meets that pre-set target numbers of individuals (Brick, 2014). This avoids issues of low response rates or frame undercoverage in probability sampling methods, as people who are unwilling to participate in the survey are simply replaced

(Yang & Banamah, 2014). While the quota sampling methodology may have limitations for the generalizability of sample statistics to a broader population, it should not impact the internal validity of the experiment. Following the "fit-for-purpose" framework proposed in Baker et al. (2013), this study is primarily concerned with estimating sample average treatment effects, for which convenience samples may be acceptable (Coppock & McClellan, 2019).

To be included in the general public subsample, participants needed to be registered on the online panel, have access to the internet, have a device that enabled them to take part in online experiments, reside in the United States, and be over 18 years of age. Participants also needed to consent to participate in the survey and provide answers to the quota screening questions.

To be considered a public administrator, there were three additional criteria:

1. Participants needed to be employed full or part-time,
2. Participants must have final or significant decision-making authority or influence within their organization, and
3. Participants must work either for a non-profit or governmental organization.

This experiment was conducted on Predictiv (www.predictiv.co.uk), an online platform for running behavioral experiments built by the Behavioral Insights Team. Predictiv provides access to a large panel, including over 2 million individuals in the US, as well as the functionality to run a range of online experiments, such as choice simulations and comprehension tests.

Participants were recruited from the general public via a network of more than 40 panel providers to the Predictiv platform. These participants have already agreed, with their respective panel provider, to take part in online research and be contacted about studies for which they are eligible. Participants are typically recruited via online and offline advertisements and referral programs. Providers and participants go through a host of security and quality checks to ensure that data collected through the network is reliable.

Participants agreed to participate specifically in this 6-8 minute online survey. They either selected this survey through the provider's portal or were invited directly through a notification from the panel provider. These notifications include targeted email invitations, offerwalls, SMS or in-app messaging. Participants are given some high-level information about the study to help them decide if they want to take part (e.g., name, modality, time.) This information is kept broad to avoid participants self-selecting or preparing for the experiment.

Predictiv is able to guarantee a representative quota sample of the US population on the characteristics of gender, age, race/ethnicity, and regional geographical location. Quotas are used at the front-end of the experiment during participant screening, which allowed me to cap participants if the sample target for certain groups had been reached. The general public subsample quotas were set to the platform's "US census representative" defaults. The public administrator subsample quotas were modeled off of the U.S. Bureau of Labor Statistics estimates for the Public Administration industry (2021).

Table 9: Experiment Sample Quotas

	General Public Subsample	Public Administrator Subsample
Gender	● Female: 51% ● Male: 49%	● Female: 46% ● Male: 54%
Age	● 18 - 24: 12% ● 25 - 34: 18% ● 35 - 44: 16% ● 45 - 54: 16% ● 55 - 54: 17% ● 65+: 21%	Not applicable

Race	● White: 75% ● Black: 13% ● Another: 12%	● White: 73% ● Black: 18% ● Another: 9%
Ethnicity	● Hispanic: 16% ● Not Hispanic: 84%	● Hispanic: 13% ● Not Hispanic: 87%
Region	● Northeast: 17% ● South: 38% ● Midwest: 21% ● West: 24%	Not applicable

Participants are compensated for their time in completing the survey. The range for a 6-8 minute survey is between $1 and $3. Participants know this rate in advance, and already have an existing financial connection to their panel provider. These panel providers each have unique incentive programs, but most provide loyalty reward points, gift cards, or cash payments. Incentives are only provided to participants who fully complete the survey. I was not able to control the incentivization model.

To be included in the final sample, participants also needed to pass an attention check at the start of the survey experiment. The purpose of the attention check is to screen out participants who are not reading closely and computer programs that are trained to complete surveys at high speed. Because online surveys are self-administered, there is a risk of “extreme forms of satisficing,” or careless responding, in which respondents may fail to consider the questions carefully, but also fail to read the question at all (Anduiza & Galais, 2016; Ward & Meade, 2018). Attention or manipulation checks are used to detect respondent disengagement. These checks have shown to be effective, and are recommended for increasing data reliability (Alvarez et al., 2019; Berinsky et al., 2014; Oppenheimer et al., 2009).

The text of the attention check included in this experiment follows the example of Aronow et al. (2020). Participants are given two opportunities to answer the attention check correctly. After two failures, their survey is terminated, and they are referred back to their panel provider without payment.

Table 10: Experiment Attention Check Text

For our research, careful attention to survey questions is critical! To show that you are paying attention, please select "I have a question."

a. I understand
b. I do not understand
c. I have a question

[New Screen: If a, b] You didn’t select the correct answer to our last question. Your attention to the survey questions is very important for our research, so we’d like to give you another chance to respond. To show that you are paying attention, please select “I have a question.”

a. I understand
b. I do not understand
c. I have a question

[New Screen: If a, b] You have answered our questions incorrectly. We can only accept surveys from people who are paying close attention, so we have ended this survey early. Please click ‘Next’ to return to your panel website.

It is possible that survey participants may differ systematically from the population-of-interest in several ways. First, this survey requires participants to be internet users. According to the Pew Charitable Trusts, approximately 7 percent of US adults do not use the internet (Perrin & Atske, 2021). Age, educational attainment, and income are indicators of a person's likelihood to be offline: older people (over 65) and those with less education (HS degree or less) and income (less than $30,000 per year) are more likely to be offline (Perrin & Atske, 2021).

Second, participants may have higher favorability towards evaluation. Panel participants have already agreed to participate in evaluations and market studies, indicating higher levels of familiarity with and interest in evaluations more generally. However, it is not clear whether a greater familiarity or interest in evaluation influences preferences for the RCT versus the quasi-experiment. To investigate a potential relationship between evaluation experience and RCT aversion, I ask participants about whether they have previously participated in designing or implementing an evaluation and use this as a covariate in my analysis.

Finally, the public administrator subsample may be a bit broader than ideal. The public administrator population-of-interest are individuals with management authority over public goods, services, or programs intended to improve social welfare. However, I was only able to feasibly screen on organization type, or whether a participant worked for the government or a non-profit. Therefore, it is possible that my sample both misses some people whom I would ideally include (for example, private sector managers but contracted to work for the government), and includes some people whom I would ideally exclude (for example, managers working in the non-profit sector but not overseeing the delivery of public goods and services.)

Experimental Survey Structure

Before entering the experiment, survey participants provided answers to several screening questions to check their eligibility. These questions were also used for the survey quotas to ensure demographic representation. After completing the survey screener through the panel provider, participants were randomized to a policy scenario arm upon entering the survey. The first survey page was an introduction to the survey, which provides high level information about the task and duration.

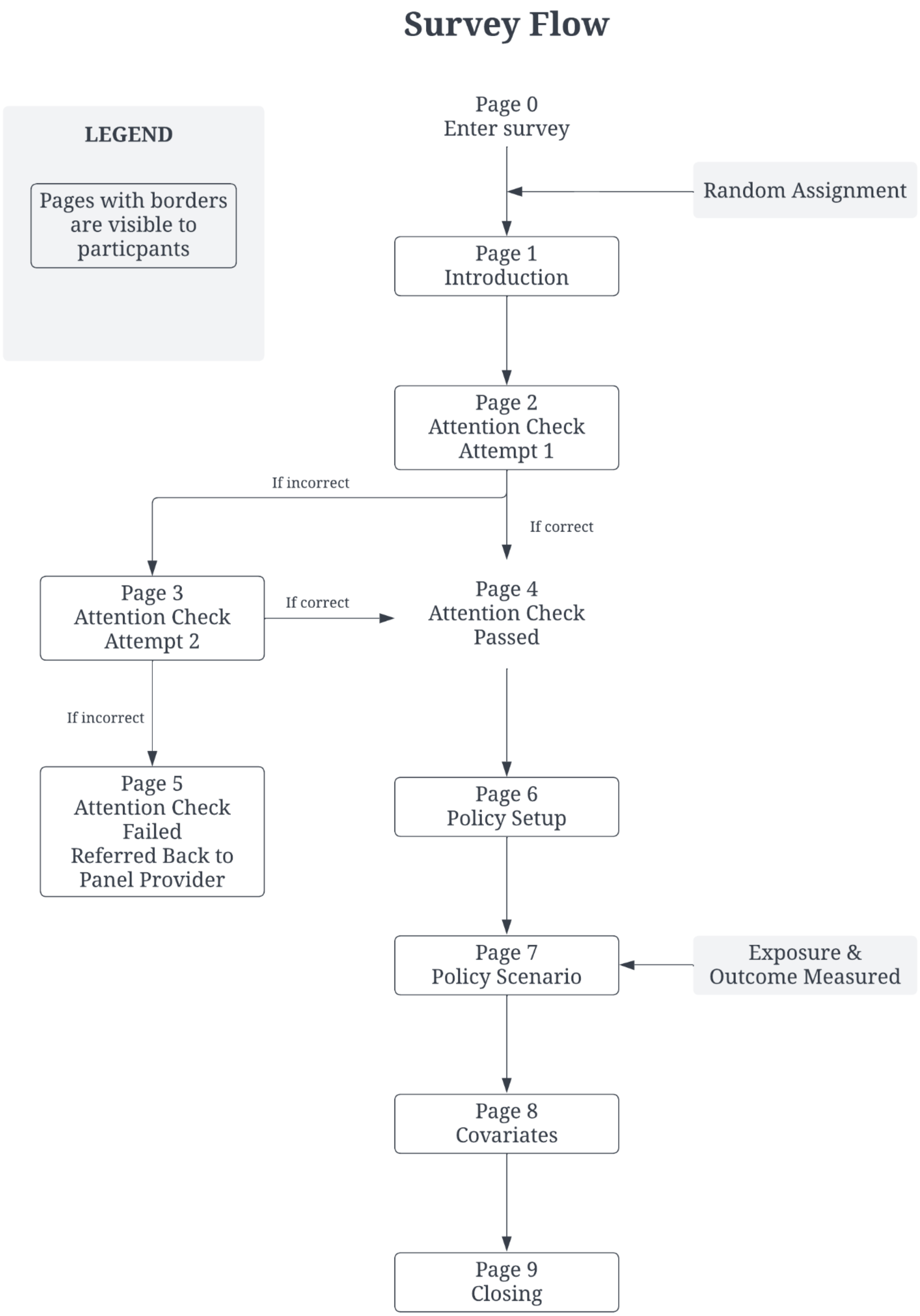
If they clicked ‘continue,’ participants then entered the attention check section of the survey. If they failed the attention check twice, their survey was terminated and they were referred back to their panel provider. If they passed on the first or second attempt, they continued to the survey page that provided the policy setup. Conditional on their random assignment, they then saw one of eight policy scenario pages and selected which evaluation option they preferred.

After completing the policy scenario portion of the survey, participants were then asked to answer six additional survey questions covering ideology, political affiliation, a short personality assessment, their educational attainment, work industry, and evaluation experience. These are factors that may be associated with underlying preferences towards evaluation and policy choices, and were used as covariates in my analysis to improve estimation precision.

Before exiting the survey, participants saw a final screen that thanked them for their participation and provided an opportunity to provide feedback.

This study was reviewed by Syracuse University's Institutional Review Board. They determined that it did not meet the criteria for human subjects research, as no personally identifying information was shared with Syracuse University researchers.

Figure 3: Survey Participant Flow



Analytical Strategy

In its basic form, my analytical approach was to estimate the following model:

$$Pr\ (ChooseRCT)_i\ =\ \alpha noProgram_i\ +\ \beta healthyHome_i\ +\ \gamma additionalFunding_i + \delta orderRCTfirst_i +$$
$$\lambda X_i +\ \varepsilon_i$$

where the outcome is a binary measure equal to one if participant i selected the randomized controlled trial as their preferred evaluation option, and zero if they chose the quasi-experiment. $noProgram_i$ represents a dummy variable equal to 1 if the participant was randomized to the scenarios where the business-as-usual condition was no lead testing program (scenarios 2, 4, 6, and 8), and 0 if not (scenarios 1, 3, 5, 7). $healthyHome_i$ represents a dummy variable equal to 1 if the participant was randomized to the scenarios where the intervention being tested was the Healthy Home Assessment (scenarios 1, 2, 3, and 4), and 0 if not (scenarios 5, 6, 7, and 8). $additionalFunding_i$ represents a dummy variable equal to 1 if the participant was randomized to the scenarios where the scaling path was to apply for additional funding (scenarios 3, 4, 7 and 8), and 0 if not (scenarios 1, 2, 5, and 6). In this model, the $\alpha, \beta, and\ \gamma$ coefficients represent my parameters of interest, and capture the average difference between the two scenario conditions for each of the three policy element dimensions.

More specifically, these coefficients correspond to three hypotheses:

- α tests the hypothesis that, on average, participants will prefer RCTs less when people assigned to the control group will get nothing versus when people assigned to the control group will have access to a current program,
- β tests the hypothesis that, on average, participants will prefer RCTs less when the policy intervention is more substantial or intense (e.g., a healthy home assessment versus mailed lead paint testing kits), and

- γ tests the hypothesis that, on average, participants will prefer RCTs less when people assigned to the control group have to wait for additional resources or approvals to receive the new program, versus automatically getting access if the evaluation shows it is successful.

If the estimated values of $\alpha, \beta, and\, \gamma$ are statistically significant and negative, we can conclude that being exposed to the "greater treatment gap" scenario policy elements increases aversion to choosing the RCT relative to the "smaller treatment gap" scenario policy elements on average.

All models include a binary covariate for whether the participant was randomized to the survey variant where the RCT evaluation option was ordered first, or not. The coefficient δ, therefore, represents the average ordering effect for the evaluation options. The vector ϵ includes all coefficients for the control variables, including the intercept, and ε is a random error term. The X_i vector includes indicators grouped into two categories. Covariate group A includes indicators for educational attainment, organizational type, level of organizational decision-making authority, ideology, and political affiliation. It also includes a short, 10-item rating instrument for the Big-Five personality dimensions, called the Ten Item Personality Inventory or TIPI (Gosling et al., 2003). The Big-Five model of personality classifies individual differences in personality traits into five broad domains: extraversion, agreeableness, conscientiousness, stability, and openness. The TIPI includes two items for each of these five domains, which are averaged together to produce an individual score for that dimension. Covariate group B includes indicators for age, male gender, race, hispanic ethnicity, and employment status.

My primary model specification is a linear probability model, estimated using ordinary least squares. However, I also report results for this specification using a logit model as a robustness check. I report a set of three variations on the primary model specification in the primary results. Model 1 does not include covariate adjustment, and is run only with the variables representing the randomized elements for scenarios and survey order variant. Model 2 includes group A covariates, and Model 3 includes group A and group B covariates. Therefore, the scenario coefficients from Models 2 and 3 represent conditional average treatment effects.

I use the Benjamini-Hochberg procedure to control the false discovery rate in making multiple comparisons. The Benjamini-Hochberg procedure is preferred because it represents a "middle road" between being too conservative (compared to the Bonferroni corrections) and too liberal (Ferreira & Zwinderman, 2006; Thissen et al., 2002). It is also easy to implement. The procedure works as follows: take the p-values from each comparison and arrange them in ascending order. Then compare these p-values with a linearly increasing vector from alpha/k (where k is the number of comparisons) to alpha. Once a comparison is found not significant, all remaining comparisons are also classified as non-significant. The table below illustrates the significance thresholds for my three treatment coefficients, using an original significance threshold of 0.05 for alpha.

Table 11: Significance thresholds using the Benjamini-Hochberg Procedure

Comparison #	p-value	Original significance threshold	New significance threshold
1	smallest	0.05	0.05*(1/3) = 0.0167
2	medium	0.05	0.05*(2/3) = 0.0333
3	largest	0.05	0.05*(3/3) = 0.05

Table 12: Data structure

VARIABLE	TYPE	SOURCE	MEASUREMENT

choseRCT	Dependent	Constructed from survey question “evalOptionChoice” Option A(1) if variant 1 Option B(2) if variant 2	Binary, individual level 0 “Chose QE” 1 “Chose RCT”
noProgram	Independent	Constructed from survey arm	Binary, individual level 0 if scenarios 1, 3, 5 and 7 1 if scenarios 2, 4, 6, and 8
healthyHome	Independent	Constructed from survey arm	Binary, individual level 0 if scenarios 5, 6, 7, and 8 1 if scenarios 1, 2, 3, and 4
additionalFunding	Independent	Constructed from survey arm	Binary, individual level 0 if scenarios 1, 2, 5, and 6 1 if scenarios 3, 4, 7, and 8
orderRctFirst	Covariate	Constructed from survey arm	Binary, individual level 0 if scenarios variant 2 1 if scenarios variant 1
eduAttainmentCat	Covariate - Group A	Survey question	Categorical, dummy coded, individual level 0 “Less than high school” 1 “High school graduate” 2 “Some college” 3 “College graduate” 4 “Masters or doctorate degree” 5 “None of the above” ORIGINAL: 0 "Less than College" 1 "College graduate" 2 "Masters or doctorate degree"
orgTypeCat	Covariate - Group A	Screener question	Categorical, dummy coded, individual level 0 "Government" 1 "Nonprofit" 2 "Private and other"
orgAuthorityCat	Covariate - Group A	Screener question	Categorical, dummy coded, individual level 0 "Final or significant decision-making authority"

			1 "Minimal or no decision-making authority"
ideology	Covariate - Group A	Survey question	Categorical, dummy coded, individual level 0 "Neither conservative nor liberal" 1 "Conservative" 2 "Liberal”
politicalAffiliationCat	Covariate - Group A	Survey question	Categorical, dummy coded, individual level 0 "Independent" 1 "Republican" 2 "Democrat" 3 “Prefer not to answer”
tipiExtraversion	Covariate - Group A	Constructed from survey question “personality” (1 + 6Reverse)/2	Continuous, individual level
tipiAgreeableness	Covariate - Group A	Constructed from survey question “personality” (2Reverse + 7)/2	Continuous, individual level
tipiConscientiousness	Covariate - Group A	Constructed from survey question “personality” (3 + 8Reverse)/2	Continuous, individual level
tipiStability	Covariate - Group A	Constructed from survey question “personality” (4Reverse + 9)/2	Continuous, individual level
tipiOpenness	Covariate - Group A	Constructed from survey question “personality” (5 + 10Reverse)/2	Continuous, individual level
evalExperienceCat	Covariate -	Survey question	Categorical, dummy coded,

	Group A		individual level 0 "No or Don’t Know" 1 "Yes"
ageCat	Covariate - Group B	Screener question	Categorical, dummy coded, individual level 0 "18-24" 1 "25-34" 2 "35-44" 3 "45-54" 4 "55-64" 5 "65+"
male	Covariate - Group B	Constructed from screener question “gender”	Binary, individual level 0 “Female or Another” 1 “Male”
race	Covariate - Group B	Screener question	Categorical, dummy coded, individual level 0 "White" 1 "Black" 2 "Another race"
hispanic	Covariate - Group B	Screener question	Binary, individual level 0 “Not hispanic” 1 “Hispanic”
employmentCat	Covariate - Group B	Screener question	Categorical, dummy coded, individual level 0 "Employed" 1 "Self-Employed" 2 "Another employed"

Results

Recruitment and Survey Attrition

I recruited a total of 5,407 unique online survey participants from February 12, 2022 through March 12, 2022. Table 1 below shows how many unique participants started the survey by day.

Table 13: Survey participants by survey data

Survey Date	Unique survey participants		
	Count	Percent	Cumulative
2022-02-12	109	2.02%	2.02%
2022-02-13	7	0.13%	2.15%
2022-02-17	468	8.66%	10.80%
2022-02-18	372	6.88%	17.68%
2022-02-19	306	5.66%	23.34%
2022-02-20	245	4.53%	27.87%
2022-02-21	380	7.03%	34.90%
2022-02-22	236	4.36%	39.26%
2022-02-23	17	0.31%	39.58%
2022-02-24	381	7.05%	46.62%
2022-02-25	387	7.16%	53.78%
2022-02-26	181	3.35%	57.13%
2022-02-27	35	0.65%	57.78%
2022-02-28	582	10.76%	68.54%
2022-03-01	127	2.35%	70.89%
2022-03-02	15	0.28%	71.17%
2022-03-03	7	0.13%	71.30%
2022-03-04	19	0.35%	71.65%
2022-03-05	313	5.79%	77.44%
2022-03-06	336	6.21%	83.65%
2022-03-07	137	2.53%	86.18%
2022-03-08	223	4.12%	90.31%
2022-03-09	158	2.92%	93.23%
2022-03-10	206	3.81%	97.04%
2022-03-11	101	1.87%	98.91%
2022-03-12	59	1.09%	100.00%
Total	5407	100.00%	

Of the 5,407 unique participants recruited, a total of 4,110 completed the survey, for a completion rate of 76.01%.

There were two sources of attrition in the experiment. The first source was participants choosing to exit the survey prior to completion. A total of 506 participants terminated the survey early, for a dropout rate of 9.36% (506/5407).

The second source of attrition was participants failing the attention check. As described above, participants were given two chances to provide a correct answer to the attention check question. A total of 791 participants ultimately did not provide a correct answer to either question, for a failure rate of 14.65% (791/5407). These participants had their survey terminated early, and were forwarded back to their panel provider without compensation. This rate of attention-failure is slightly less than many recent estimates, which have put estimates at about one third (Aronow et al., 2020; Alvarez et al., 2019).

Figure 4 below shows the overall participant flow of the survey experiment, and the attrition at each stage of the survey. The largest points of exit were on page "0," which indicates that participants exited before the introduction page loaded, and on page 6, where participants were given the overall policy scenario. While participants were randomized upon entry to the survey, they were not exposed to treatment until page 7.

Table 14 shows the attrition, attention check failure, and total survey completion counts by policy scenario condition. Descriptively, attrition rates are fairly similar across scenario conditions.

I then conduct statistical tests for differential attrition by scenario condition. I ran a regression with "Complete" as our outcome, which is a binary variable equal to 1 if the participant completed the survey. Completion includes participants who passed the attention check and progressed to the final page of the survey, as well as participants who failed the attention check and did not exit the survey before being referred back to their panel provider. The independent variables are indicators for scenario condition. Figures 5 through 7 graph these results.

Figure 4: Survey participant completion and attrition

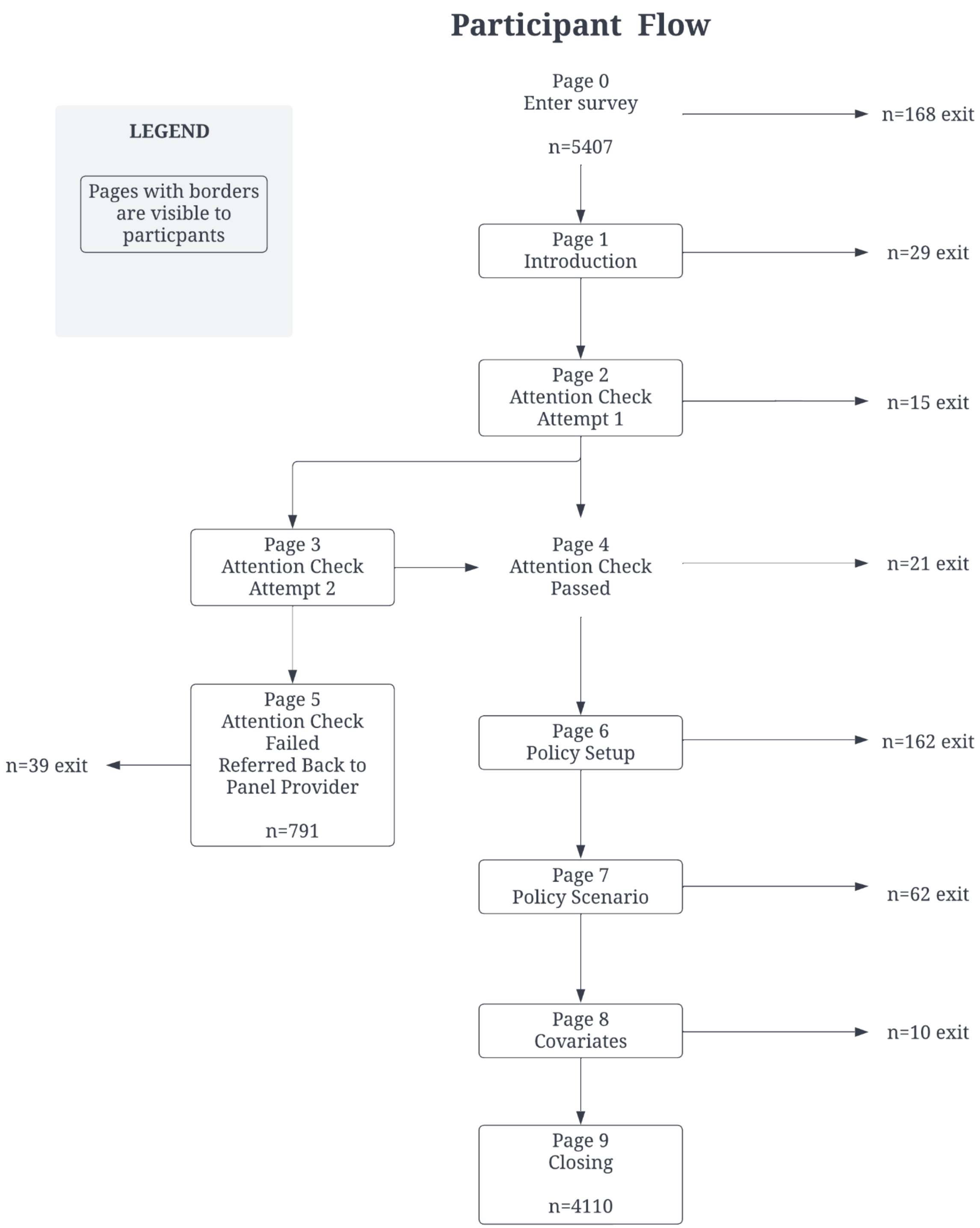
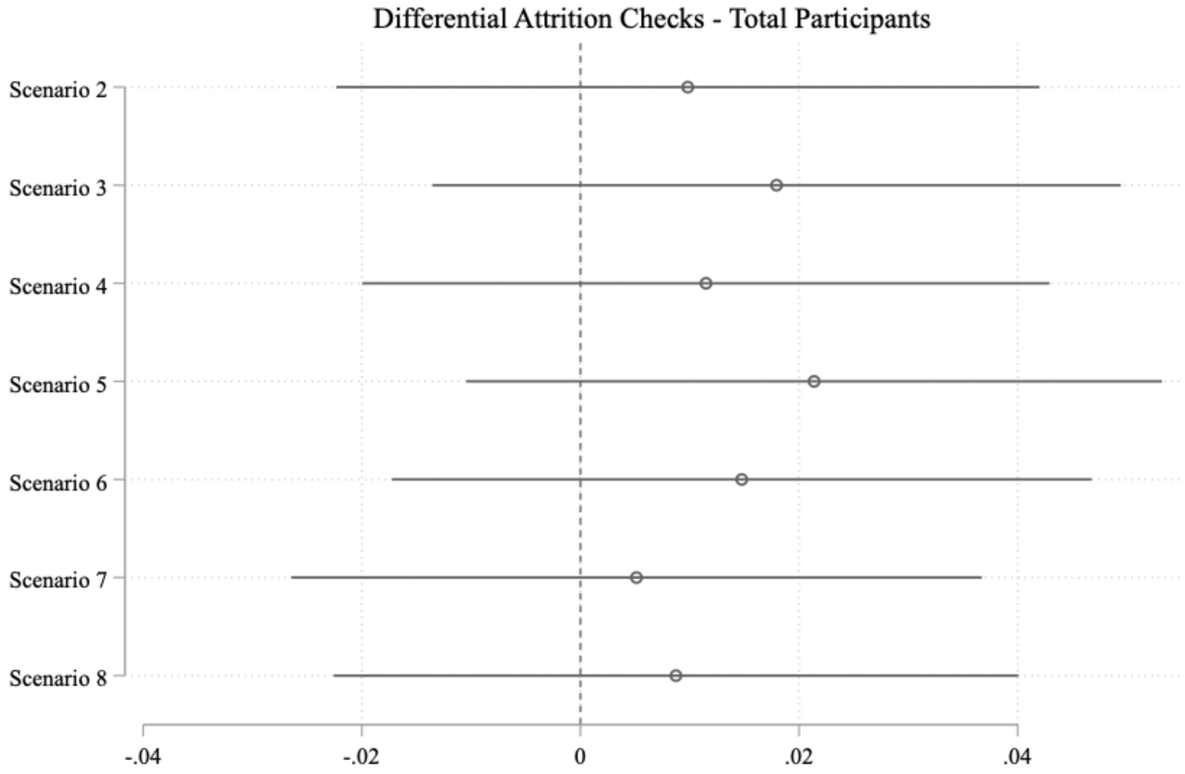


Table 14: Survey participant completion and attrition

	Attrited from survey									Failed attention check	Completed survey	Total
	Page 0	Page 1	Page 2	Page 3	Page 4	Page 5	Page 6	Page 7	Page 8			
Scenario 1	22	2	2	0	4	6	20	7	2	78	477	620
Scenario 2	21	5	0	0	1	5	17	9	3	110	471	642
Scenario 3	17	4	1	0	3	3	28	5	0	102	539	702
Scenario 4	22	5	3	0	1	10	17	8	0	99	542	707
Scenario 5	21	3	2	0	1	4	20	5	0	103	512	671
Scenario 6	19	3	2	0	1	1	19	12	2	113	483	655
Scenario 7	24	1	3	0	2	8	21	9	1	82	541	692
Scenario 8	22	6	2	0	8	2	20	7	2	104	545	718
Total	168	29	15	0	21	39	162	62	10	791	4110	5407

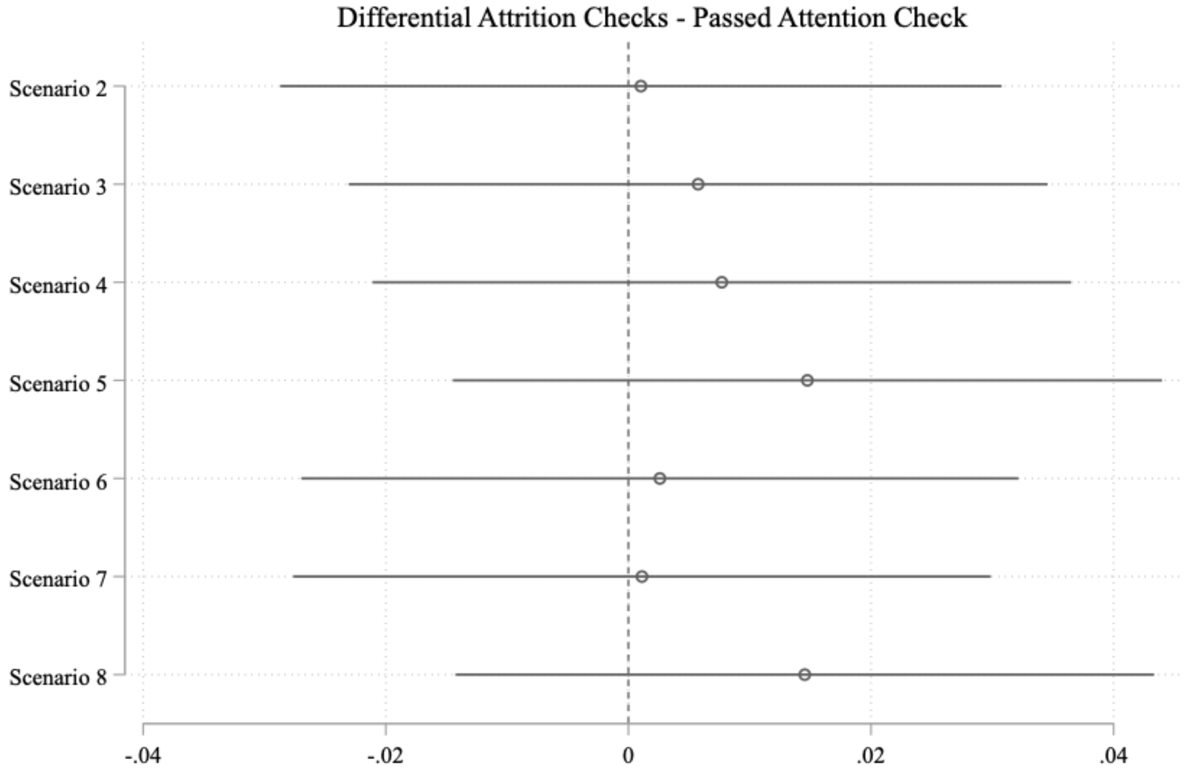
Figure 5: Differential Attrition Checks for full sample



Note: N = 5407

Figure 5 represents the results from this regression for the full sample of unique participants who began the survey. No scenario coefficient is statistically significant. I conclude that scenario assignment is not correlated with the likelihood that a participant will complete the survey or leave early.

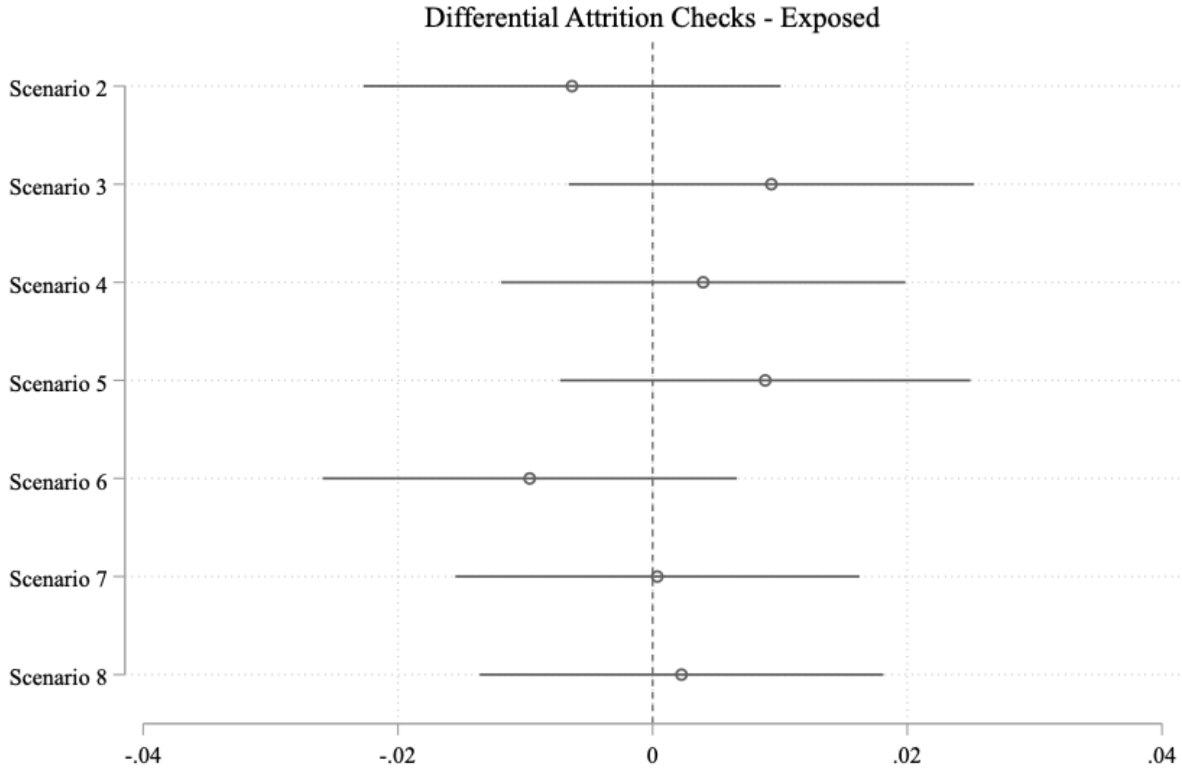
Figure 6: Differential Attrition Checks for participants passing the attention check



Note: N = 4383

Figure 6 represents the results from same regression, but for the sample of participants who passed the attention check. Again, no scenario coefficient is statistically significant. There appears to be no relationship between scenario assignment and whether the participants who passed the attention check will complete the survey.

Figure 7: Differential Attrition Checks for participants exposed to treatment scenarios



Note: N = 4182

Finally, Figure 7 represents the regression results for the survey participants who progressed far enough to be exposed to the treatment scenarios. Similar to the previous results, no scenario coefficient is statistically significant. Treatment assignment is not related to whether participants who saw the different scenarios would go on to complete the survey in full.

Table 15 shows the final complete survey counts by policy scenario elements. While the observations are fairly balanced across conditions, some variation is expected due to simple random assignment.

Tables 16 and 17 show complete survey counts by the sub-samples of interest: general population and public administrator. Again, while the observations are not equal across scenarios or policy elements, they are fairly well balanced.

The average length of time to complete the survey was 4 minutes and 47 seconds, with a standard deviation of 7 minutes and 49 seconds.

Table 15: Count of survey participants by scenario

	Business-as-usual condition		**Intervention**		**Scaling path**	
	Current program	No program	Free mailed test kits	Healthy Home visits	Scale immediately	Seek additional funding
Scenario 1	477	0	0	477	477	0
Scenario 2	0	471	0	471	471	0
Scenario 3	539	0	0	539	0	539
Scenario 4	0	542	0	542	0	542
Scenario 5	512	0	512	0	512	0
Scenario 6	0	483	483	0	483	0
Scenario 7	541	0	541	0	0	541
Scenario 8	0	545	545	0	0	545
Total	2069	2041	2081	2029	1943	2167

Table 16: Count of survey participants by scenario and sub-sample

	General Population	Public Administrator	Total
Scenario 1	248	229	477
Scenario 2	234	237	471
Scenario 3	275	264	539
Scenario 4	279	263	542
Scenario 5	253	259	512
Scenario 6	240	243	483
Scenario 7	269	272	541
Scenario 8	262	283	545
Total	2060	2050	4110

Table 17: Count of survey participants by policy element and sub-sample

	General Population	Public Administrator	Total
Business-as-usual			
Free testing	1045	1024	2069
No program	1015	1026	2041
Intervention			
Free mailed test kits	1024	1057	2081
Healthy Home visit	1036	993	2029
Scaling path			
Scaled immediately	975	968	1943
Seek additional funding	1085	1082	2167
Total	2060	2050	4110

Sample

Overall, demographics for the survey participants align fairly closely to expectations. Table 18 shows the demographic breakdown for each of our samples of interest: the full sample, the general population sub-sample, and the public administrator sub-sample.

For the general population sub-sample, I was able to adhere quite closely to the target quotas. The age and gender distribution are quite close to the US census levels. The final sample is slightly more diverse compared to the US census on race and ethnicity dimensions. The sample has a greater proportion of respondents who indicated “another race,” and slightly more respondents who indicated “Black” and “Hispanic” identity.

For the public administrator sub-sample, I was slightly off our target gender quotas, as I recruited more women than men. This subsample also has a greater proportion of respondents who indicated “another race,” and slightly fewer respondents who indicated “white” or “Black” identity.

In comparing demographics across sub-samples, the results align fairly closely to expectations. The public administrator sample has a greater proportion of respondents who indicated advanced degrees, while the general population subsample has a greater proportion of respondents who had completed “some college.” This category includes individuals who had completed some college but did not have a degree, those who had obtained an associate's degree, and individuals who had completed other post-high school vocational training.

Differences between the two samples by organization type, organizational decision-making authority, and employment status are likely due to how the sub-samples were constructed. However, individuals in the public administrator sub-sample were more likely to have some prior evaluation experience, at 49.32% compared to 21.31% in the general population sub-sample.

Both samples have a good representation of individuals from across the ideological and political spectrum. Between 17-20% of individuals in both samples are clustered at the far ends ideologically, and then the remaining 60% are distributed across the middle range. The breakdown of political affiliation is also fairly similar to one another across both samples, with public administrators having slightly fewer independents and slightly more democrats by roughly 2 percentage points.

As might be expected given the conditions for being considered a "public administrator," the distribution of ages does differ across the two sub-samples. Public administrators have more participants clustered in the 25-44 age ranges (core working years), while the general population has more participants over 65. These participants are likely to be retired, and therefore wouldn't have met the "employed" criteria for the public administrator sub-sample.

While I was not able to recruit our target quota percentage for men for the public administrator sub-sample, and therefore over-represent women slightly compared to the benchmark, the breakdown by gender is very similar across the two sub-samples.

Similarly, both sub-samples have fairly similar breakdowns for racial and Hispanic identity categories. The public administrator sub-sample is slightly less diverse than the general population sample, but slightly more diverse than the BLS benchmarks.

Finally, Ten Item Personality Inventory scores are fairly similar across the two sub-samples. If there is sectoral sorting between the private and public sectors, or between employment categories, it does not seem to be highly correlated with personality differences.

Table 18: Demographic breakdown for full sample and sub-samples

	Full Sample	General Population		Public Administrator	
	Actual	Actual	Target	Actual	Target
Educational attainment (%)					
Less than high school	2.97	4.22		1.71	
High school graduate	17.40	22.23		12.54	
Some college	33.60	40.05		27.12	
College graduate	26.30	22.86		29.76	
Masters or doctorate degree	19.49	10.44		28.59	
None of the above	0.24	0.19		0.29	
Organization type (%)					
Government	24.01	3.35		44.78	
Nonprofit	30.88	6.65		55.22	
Private and other	45.11	90.00			
Organizational decision-making authority (%)					
Final or significant	68.93	38.01		100.00	
Minimal or no	31.07	61.99			
Evaluation experience (%)					
No or not sure	64.72	78.69		50.68	
Yes	35.28	21.31		49.32	
Ideology (%)					
Very conservative	19.49	18.59		20.39	
Not very conservative	15.28	16.31		14.24	
Closer to conservative	7.76	7.18		8.34	
Neither	19.81	22.43		17.17	
Closer to liberal	7.45	7.04		7.85	
Not very liberal	11.61	11.60		11.61	

Very liberal	18.61	16.84		20.39	
Political affiliation (%)					
Independent	29.71	30.78		28.63	
Republican	28.30	28.93		27.66	
Democrat	37.15	35.39		38.93	
Prefer not to answer	4.84	4.90		4.78	
Age (%)					
18-24	11.73	12.18	12	11.27	
25-34	22.00	18.64	18	25.37	
35-44	22.14	16.31	16	28.00	
45-54	16.84	15.92	16	17.76	
55-64	14.48	16.80	17	12.15	
65+	12.82	20.15	21	5.46	
Gender (%)					
Female or another	51.48	51.41	51	51.56	46
Male	48.52	48.59	49	48.44	54
Race (%)					
White	70.80	70.68	75	70.93	73
Black	15.23	14.22	13	16.24	18
Another race	13.97	15.10	12	12.83	9
Hispanic ethnicity (%)					
Not hispanic	83.67	82.43	84	84.93	87
Hispanic	16.33	17.57	16	15.07	13
Employment status (%)					
Employed	69.39	38.93		100.00	
Self-employed	5.28	10.53			

Another employment status	25.33	50.53	
Ten Item Personality Inventory (Mean, SD)			
Extraversion	3.90	3.79	4.01
	1.42	1.40	1.43
Agreeableness	5.18	5.18	5.19
	1.18	1.17	1.18
Conscientiousness	5.51	5.45	5.57
	1.25	1.27	1.23
Stability	4.86	4.76	4.95
	1.40	1.44	1.35
Openness	5.13	5.03	5.23
	1.20	1.22	1.17

Figures 8 through 10 show the results of balance checks by policy scenario element. These graphs represent coefficients from a regression model that predicts scenario element assignment using our suite of demographic covariates. There are a few statistically significant coefficients in each model, but they are not consistent across the scenario element models. For example, some of the Ten Item Personality Inventory coefficients are statistically significant in the business-as-usual condition model and the scaling path model, but not in the intervention type model.

Figure 8: Balance Checks for Business-as-Usual Policy Element

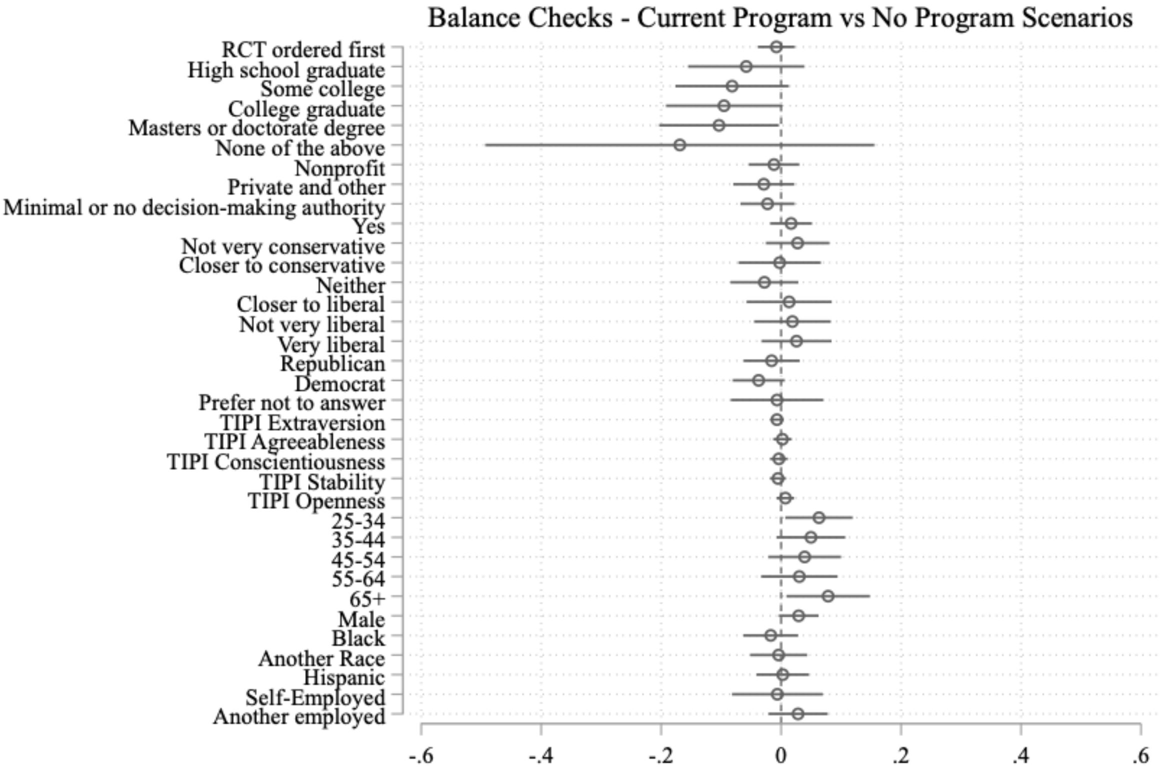


Figure 9: Balance Checks for Intervention Policy Element

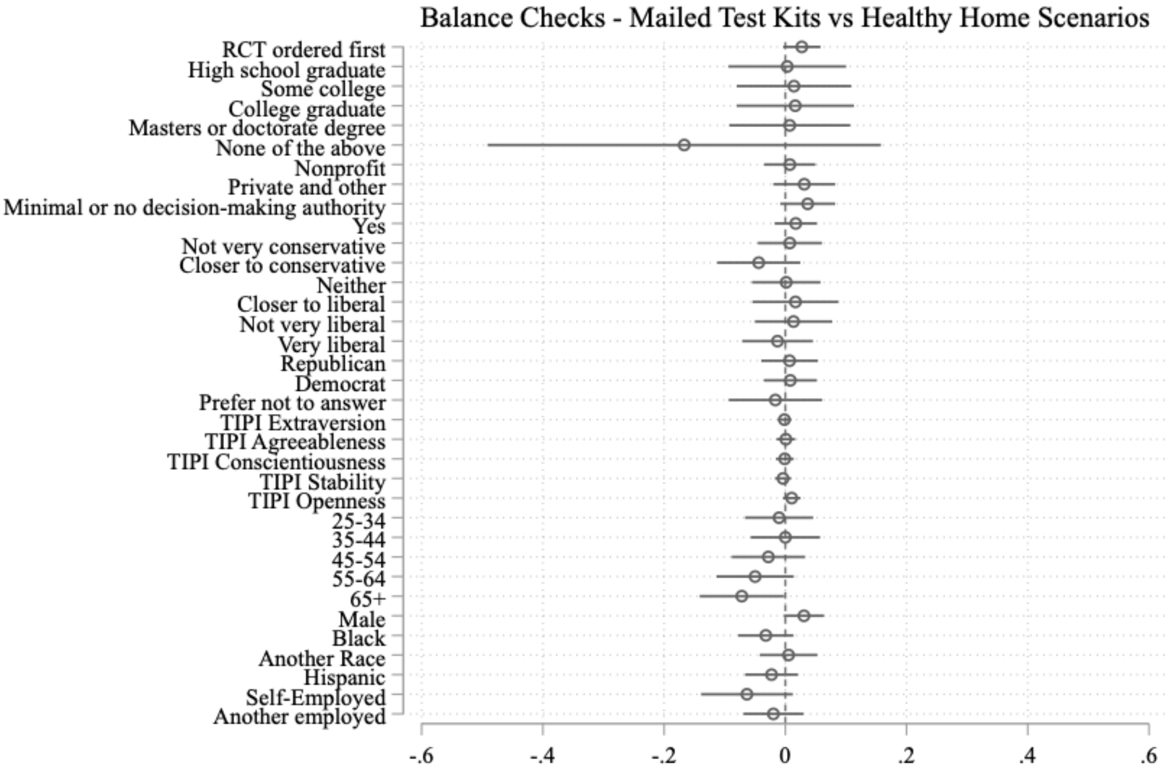
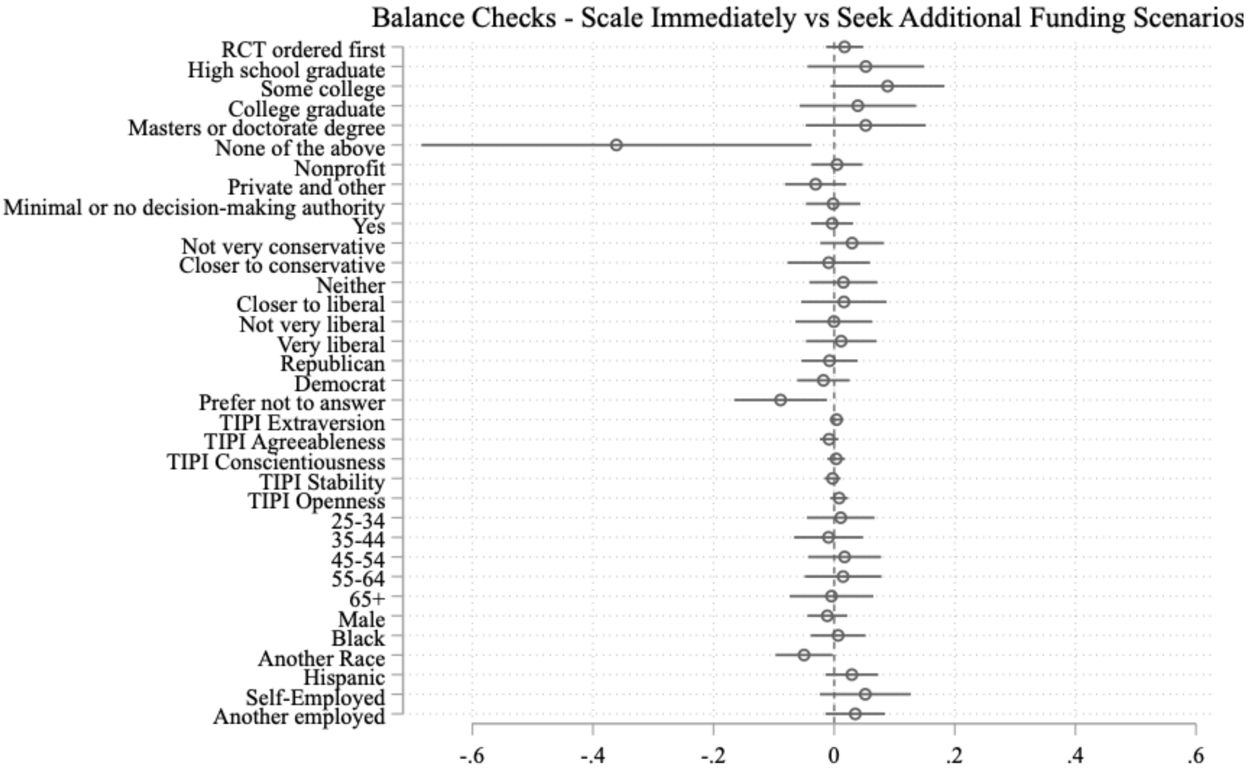


Figure 10: Balance Checks for Scaling Path Policy Element



Descriptive Results

Perhaps unsurprisingly, I find that most survey participants select the first evaluation option presented to them. Table 19 shows that 57% of the respondents choose “Option A”, versus 43% who choose “Option B.”

Table 19: Breakdown of survey participants selecting option A or B

How many choose "Option A"?

	Count	Percent
Option A	2328	56.64
Option B	1782	43.36

When looking at the overall rate for evaluation choice in Table 20, I find that 62% of respondents chose the quasi-experimental option, compared to 38% who chose the RCT.

Table 20: Breakdown of survey participants selecting the RCT or QE

What is the overall rate for choosing the RCT?

	Count	Percent
Chose the QE	2540	61.80
Chose the RCT	1570	38.20

I then look at the proportion who chose the quasi-experiment versus the RCT by scenario, shown in Table 21. Here I find that the overall trend continues, with an overall pattern of a preference for the quasi-experiment between 60-64%, and only 36-40% choosing the RCT. The Pearson chi-square statistic testing the relationship between scenario and evaluation choice is 6.07 and not statistically significant. Figure 11 graphs the proportion of respondents selecting the RCT by scenario.

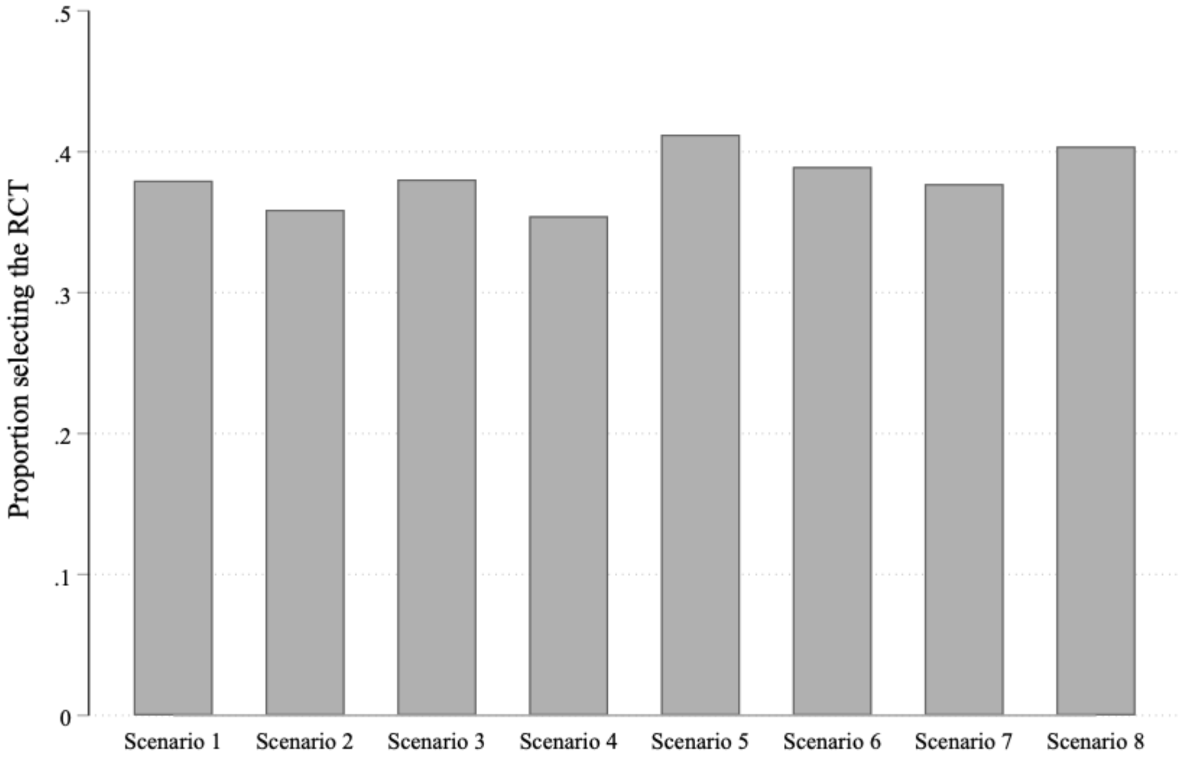
Table 21: Breakdown of survey participants evaluation selection by scenario

What is the rate of choosing the RCT by scenario?

	Chose QE	Chose RCT
Scenario 1	296	181
	62.05	37.95
Scenario 2	302	169
	64.12	35.88
Scenario 3	334	205
	61.97	38.03
Scenario 4	350	192

	64.58	35.42
Scenario 5	301	211
	58.79	41.21
Scenario 6	295	188
	61.08	38.92
Scenario 7	337	204
	62.29	37.71
Scenario 8	325	220
	59.63	40.37
Pearson Chi2	6.0738	
P-value	0.531	

Figure 11: Proportion of survey participants selecting the RCT by scenario



I then further examine the relationship between the order in which the evaluation option was presented, and the rate at which the RCT was selected. Table 22 shows the breakdown in respondents choosing the quasi-experiment and the RCT by survey arm. In Survey Arms 1 through 8, the RCT was presented first as “Option A.” While participants still tend to select the QE at higher rates than the RCT, we see a lot more variation. In Survey Arm 5, a greater number of respondents actually chose the RCT over the QE (51% to 49%). In Survey Arms 9 through 16, the RCT was presented second as “Option B.” In these survey arms, respondents greatly favored the quasi-experimental option. The Pearson chi-square statistic testing the relationship between survey arm and evaluation choice is 87.94 and statistically significant. Figure 12 shows this graphically, with each bar representing a survey arm.

Table 22: Breakdown of survey participants evaluation selection by survey arm

What is the rate of choosing RCT by survey arm?

	Chose QE	Chose RCT
Survey Arm 1	125	109
	53.42	46.58
Survey Arm 2	150	96
	60.98	39.02
Survey Arm 3	163	123
	56.99	43.01
Survey Arm 4	158	111
	58.74	41.26
Survey Arm 5	120	126
	48.78	51.22
Survey Arm 6	116	103
	52.97	47.03
Survey Arm 7	150	121
	55.35	44.65
Survey Arm 8	146	127
	53.48	46.52
Survey Arm 9	171	72
	70.37	29.63
Survey Arm 10	152	73
	67.56	32.44
Survey Arm 11	171	82

	67.59	32.41
Survey Arm 12	192	81
	70.33	29.67
Survey Arm 13	181	85
	68.05	31.95
Survey Arm 14	179	85
	67.80	32.20
Survey Arm 15	187	83
	69.26	30.74
Survey Arm 16	179	93
	65.81	34.19
Pearson chi2	87.9448	
P-value	<0.000	

Figure 12: Proportion of survey participants selecting the RCT by survey arm

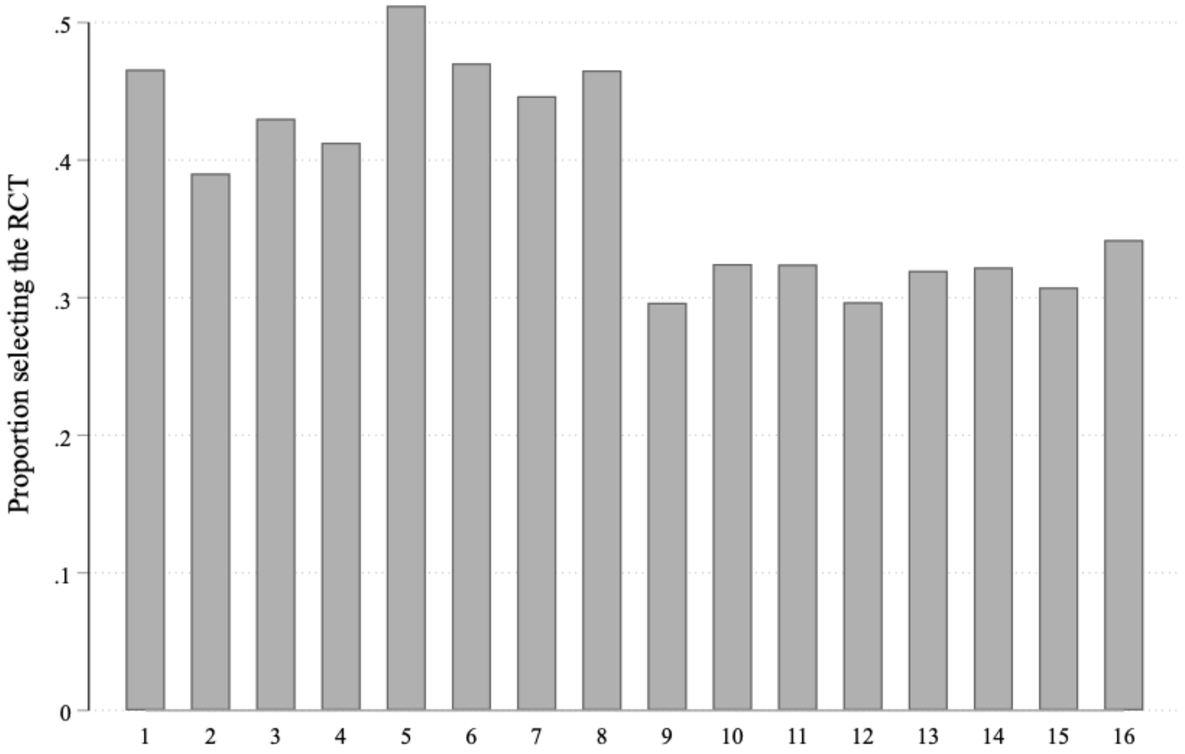


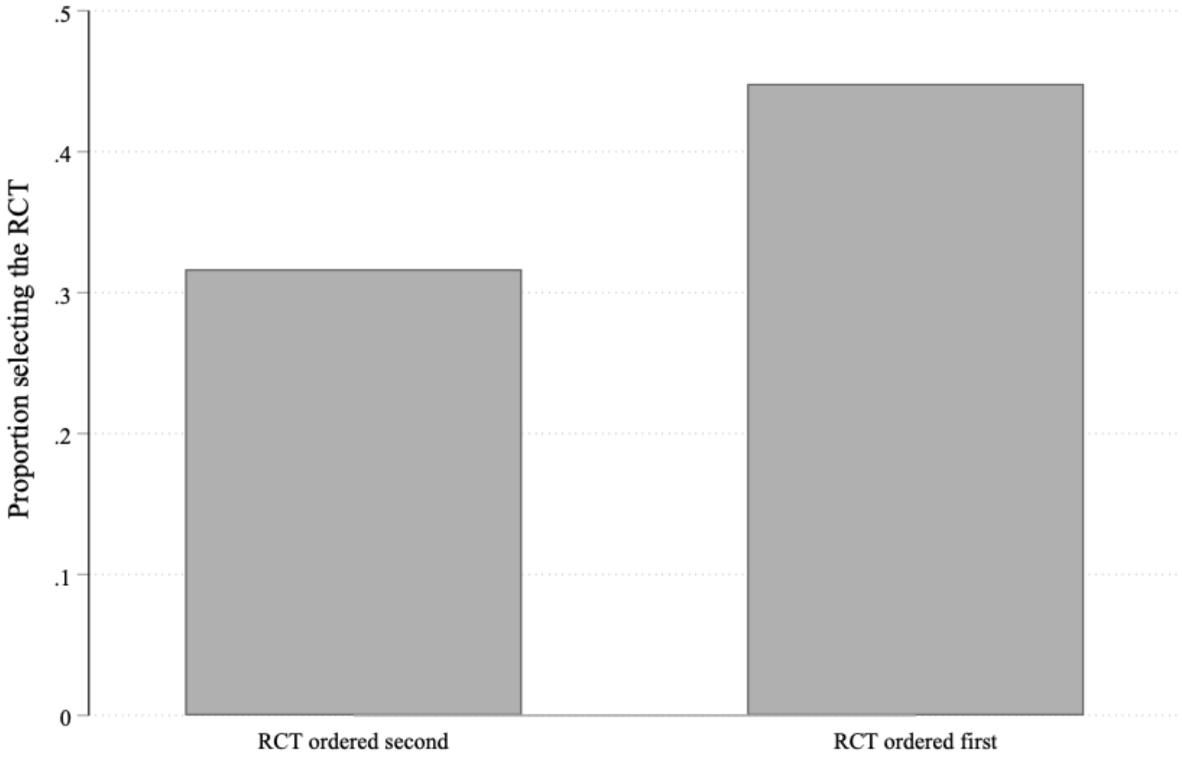
Table 23 tests this relationship more explicitly, showing the breakdown in respondents choosing the quasi-experiment versus the RCT by survey variant. In this table, Variant 1 represents the average across survey arms 1 through 8, while Variant 2 represents the average across survey arms 9 through 16. The Pearson chi-square statistic testing the relationship between survey variant and evaluation choice is 75.36 and statistically significant. Figure 10 displays the proportion of respondents choosing the RCT by survey variant. I conclude that there is strong relationship between the ordering of the evaluation options and the likelihood that the first option is selected. There are two potential explanations for this large effect. First, that while I did successfully screen out some inattentive respondents, there are still a large number who are not taking a large degree of care when responding. So while they may not respond to an item incorrectly, they may also not care to fully engage with the scenario and decision. Second, it may be that many people felt ambivalent between the two evaluation options, for many reasons. Rather than answering randomly, they select the first available option. Both explanations would be consistent with ordering effects.

Table 23: Breakdown of survey participants evaluation selection by survey variant

What is the rate of choosing RCT by survey variant?

	Chose QE	Chose RCT
Variant 1 (RCT ordered first)	1128	916
	55.19	44.81
Variant 2 (RCT ordered second)	1412	654
	68.34	31.66
Pearson chi2	75.361	
P-value	<0.000	

Figure 13: Proportion of survey participants selecting the RCT by survey variant



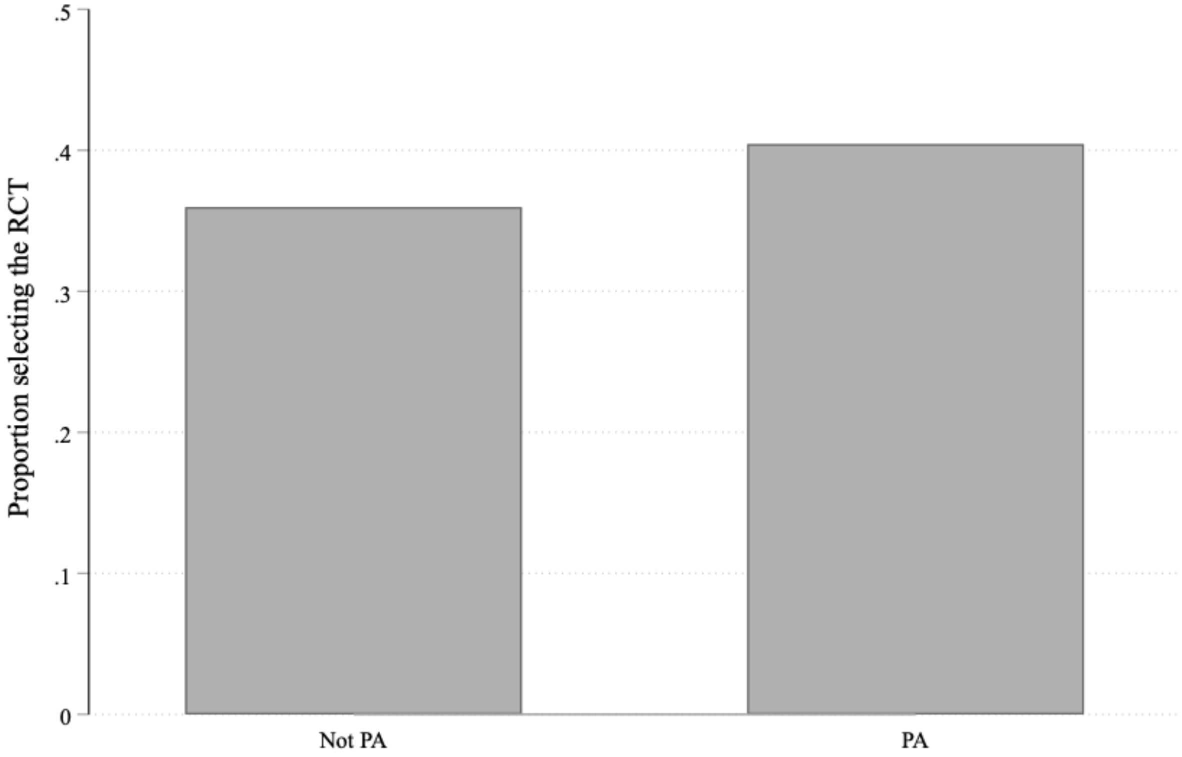
Finally, I examine the descriptive differences in the number of participants selecting the RCT by sub-sample. Table 24 shows this breakdown by the general population and public administrator samples. Both sub-samples follow the overall trend of preferring the quasi-experiment to the RCT. However, I find that the public administrators seem to have a slightly weaker preference for the quasi-experiment compared to their general population peers. About 40% of the public administrators selected the RCT as their preferred evaluation option, compared to 36% of the general population respondents. The Pearson chi-square statistic testing the relationship between sub-sample and evaluation choice is 8.69 and statistically significant. Figure 14 displays the proportion of respondents choosing the RCT by sub-sample.

Table 24: Breakdown of evaluation selection by survey participant sub-sample

What is the rate of choosing RCT by sub-sample?

	Chose QE	Chose RCT
General population	1319	741
	64.03	35.97
Public administrator	1221	829
	59.56	40.44
Pearson chi2	8.6893	
P-value	0.003	

Figure 14: Proportion of evaluation selection by survey participant sub-sample



Primary Results

FULL SAMPLE

I next examine results for our primary hypothesis that respondents will select the RCT less in the policy scenario variants where the control or comparison group is more disadvantaged compared to the treatment or intervention groups. The three variables of interest were 1) the business-as-usual condition, or what non-treated individuals would receive, 2) the intensity of the intervention being tested, and 3) the intervention scaling path. I hypothesized that respondents would prefer the quasi-experiment to the RCT in the scenarios where 1) the non-treated individuals would not have access to any lead screening program, 2) when the intervention was the Healthy Home visits, and 3) when Cityville would need to seek additional funding in order to scale the program.

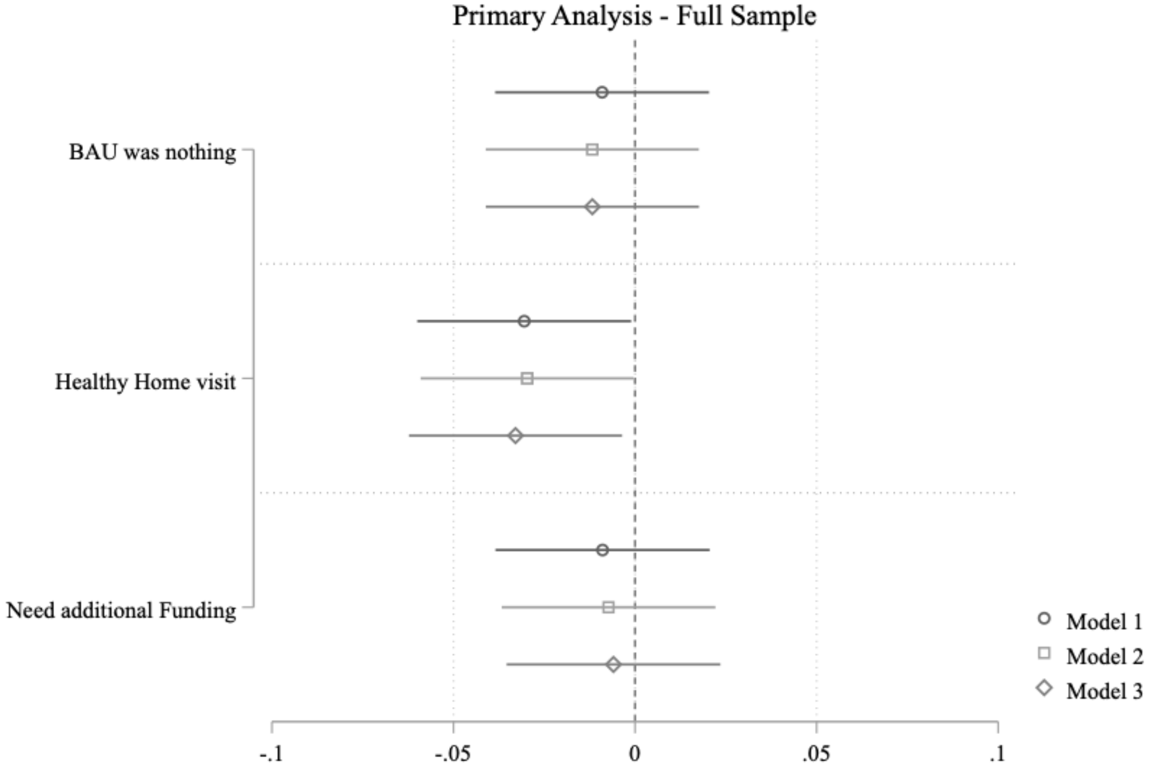
Table 25 shows the results from my primary model on the full sample. While all estimated coefficients are directionally consistent with my hypotheses, the "No program" and "Additional funding" scenarios are not statistically significantly different from zero. The "Healthy Home" scenarios are statistically significant in Models 1 through 3. However, no model's statistical significance is robust to the multiple comparison adjustment (p-values of less than .016). Figure 15 graphs the coefficients from the three models for each of the scenario elements.

Table 25: Primary analysis results for full sample

	Full sample		
	Model 1	Model 2	Model 3
No program	-0.009	-0.012	-0.012
	0.0150	0.0150	0.0150
	0.546	0.431	0.433
Healthy Home visit	-0.031	-0.030	-0.033
	0.0150	0.0150	0.0150
	0.042	0.047	0.028
Additional funding	-0.009	-0.007	-0.006
	0.0150	0.0150	0.0150
	0.552	0.627	0.692
N	4,110	4,110	4,110
Reference mean		0.317	
Covariates A		Y	Y
Covariates B			Y

Note 1: Coefficient estimates are in the first row, standard errors are in the second, and p-values are in the third row
Note 2: Reference mean controls for survey variant

Figure 15: Primary analysis results for full sample



SUB-GROUP ANALYSIS

Next, I examine the same analysis by sub-sample. Table 26 shows the results from this regression analysis. None of the scenario element indicators are statistically significantly different from zero for the general population sub-sample.

However, the Healthy Home visit indicator is statistically significant for the public administrator subsamples in Models 1 through 3. Model 3 is robust to the multiple comparison adjustments with a p-value of less than .016. The Model 3 estimates that on average, public administrators choose the RCT 5.5 percentage points less when scenario intervention is the Healthy Home visits compared to the free mailed test kits. Using the reference mean for the public administrator sample, controlling for survey variant, this indicates a relative increase in

RCT aversion of around 16 percent. Figures 16 and 17 graph the coefficients from the three models for each of the scenario elements for the two sub-samples.

Table 26: Primary analysis results for sub-samples

	General Population			**Public Administrator**		
	Model 1	Model 2	Model 3	Model 1	Model 2	Model 3
No program	-0.016	-0.016	-0.014	-0.003	-0.008	-0.012
	0.0210	0.0210	0.0211	0.0214	0.0214	0.0214
	0.454	0.446	0.498	0.875	0.722	0.572
Healthy Home visit	-0.010	-0.009	-0.013	-0.051	-0.051	-0.055
	0.0210	0.0210	0.0211	0.0215	0.0214	0.0215
	0.641	0.666	0.548	0.018	0.017	0.010
Additional funding	0.004	0.004	0.006	-0.023	-0.020	-0.019
	0.0210	0.0210	0.0211	0.0215	0.0216	0.0216
	0.854	0.854	0.787	0.288	0.353	0.372
N	2,060	2,060	2,060	2,050	2,050	2,050
Reference mean		0.299			0.333	
Covariates A		Y	Y		Y	Y
Covariates B			Y			Y

Note 1: Coefficient estimates are in the first row, standard errors are in the second, and p-values are in the third row
Note 2: Reference mean controls for survey variant

Figure 16: Primary analysis results for general population sub-sample

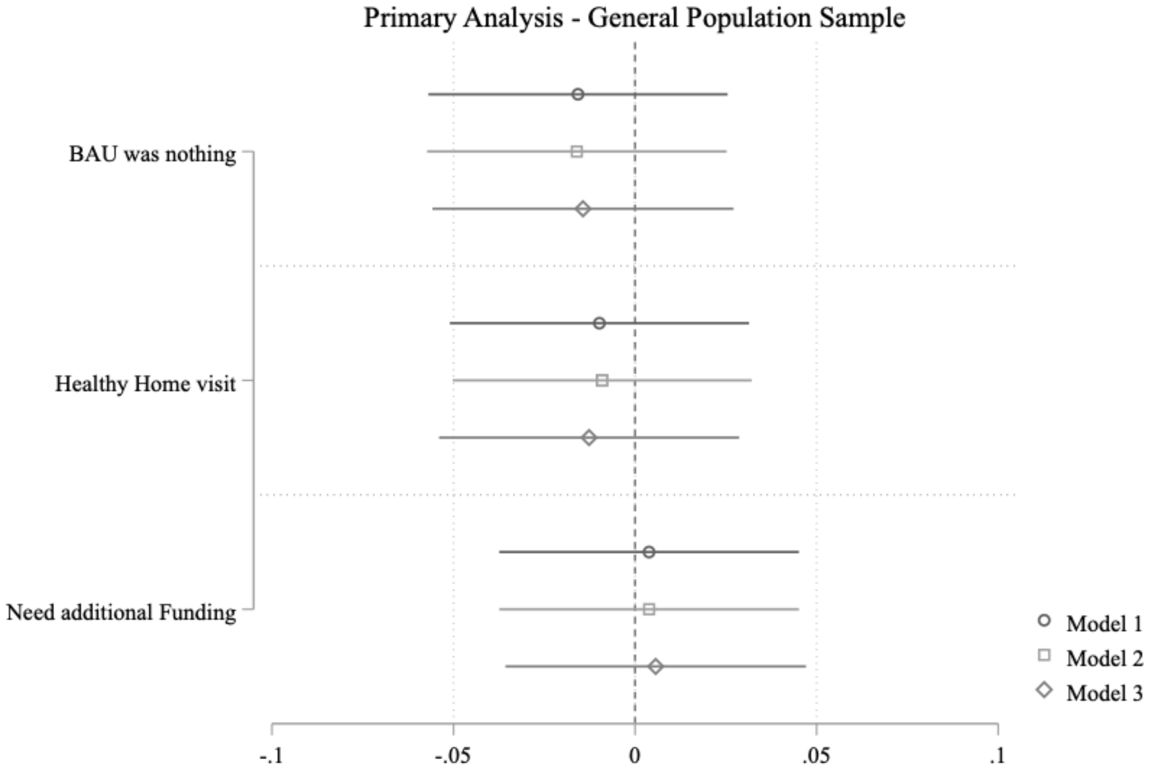
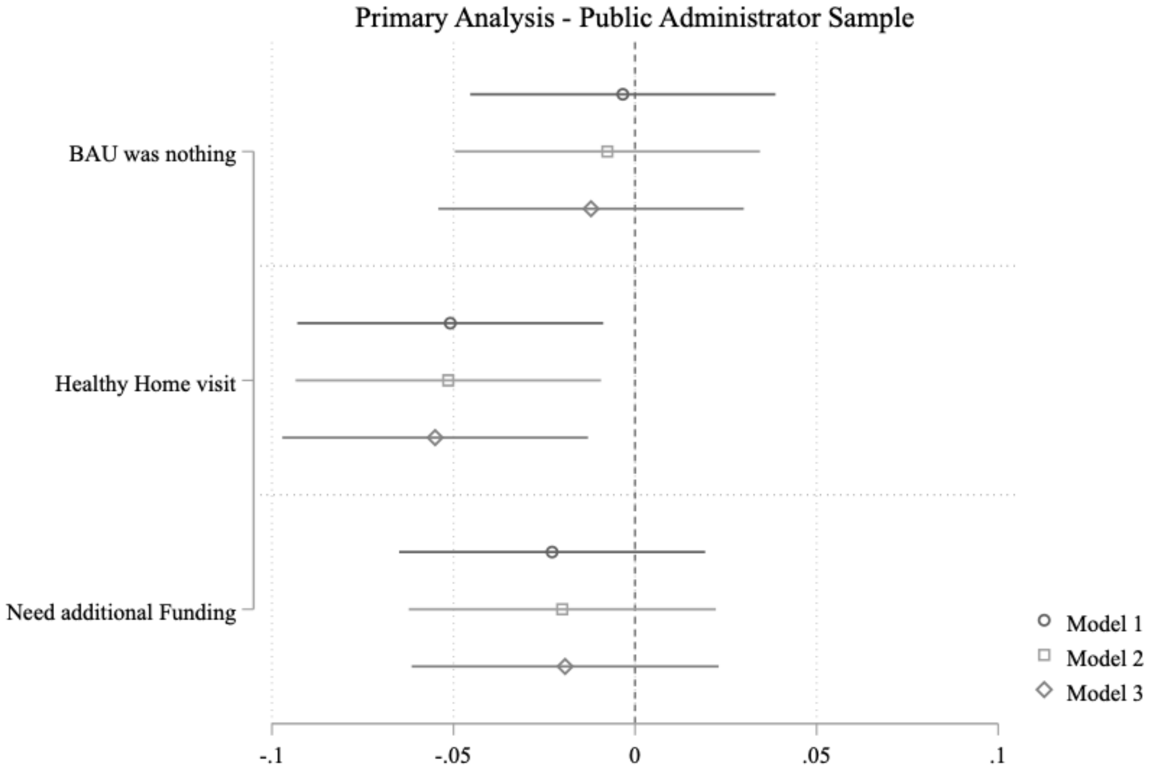


Figure 17: Primary analysis results for public administrator sample



SENSITIVITY CHECKS

I then check the sensitivity of these results to model specification by running the same analysis as a logistic regression model rather than linear probability model. Table 27 shows these results for our “Model 3” specification using both Group A and Group B covariates. These results are consistent with the linear probability model. Only the Healthy Home visit policy element is statistically significant for the full sample and public administrator sub-sample. Table 28 shows the estimated marginal effects from the logistic regression for ease of comparison with the linear probability model. The estimates are remarkably similar to the primary analysis, to within a tenth of a percentage point.

As an additional check for data quality, I also run the primary analysis again, dropping participants who spend less than 60 seconds on the survey, and then less than 90 seconds. The overall pattern of results is very similar, presented in Table 29, and the coefficients are very close to the original Model 3 specification.

Given both of these checks, I feel confident that the overall pattern of results is consistent and stable. While the “No program” and “Additional Funding” policy scenario elements are not statistically significant for either of my samples of interest, “Healthy Home visit” is consistently negative and statistically significant for Public Administrators.

Table 27: Sensitivity analysis, logistic regression

Logistic regression results

	Full Sample	General Population	Public Administrator
No program	-0.053	-0.066	-0.053
	0.0658	0.0946	0.0930
	0.417	0.482	0.571
Healthy Home visit	-0.146	-0.057	-0.240
	0.0659	0.0944	0.0933
	0.026	0.543	0.010
Additional funding	-0.026	0.026	-0.083
	0.0660	0.0945	0.0935
	0.692	0.784	0.372
N	4,110	2,060	2,050
Covariates A	Y	Y	Y
Covariates B	Y	Y	Y

Note 1: Coefficient estimates are in the first row, standard errors are in the second, and p-values are in the third row

Table 28: Sensitivity analysis, logistic regression average marginal effects

Average marginal effects

	Full Sample	General Population	Public Administrator
No program	-0.012	-0.015	-0.012
	0.0149	0.0209	0.0212
	0.417	0.482	0.571
Healthy Home visit	-0.033	-0.013	-0.055
	0.0149	0.0209	0.0212
	0.026	0.543	0.010
Additional funding	-0.006	0.006	-0.019
	0.0149	0.0209	0.0214

0.692	0.784	0.372

Note 1: Coefficient estimates are in the first row, standard errors are in the second, and p-values are in the third row

Table 29: Sensitivity analysis, trimming analytical sample by survey completion time

	Full Sample		General Population		Public Administrator	
Dropping respondents who took	<60 seconds	<90 seconds	<60 seconds	<90 seconds	<60 seconds	<90 seconds
No program	-0.011	-0.013	-0.013	-0.013	-0.012	-0.016
	0.0150	0.0156	0.0212	0.0219	0.0215	0.0224
	0.471	0.390	0.554	0.547	0.574	0.4640
Healthy Home visit	-0.034	-0.035	-0.013	-0.016	-0.057	-0.056
	0.0150	0.0156	0.0212	0.0218	0.0215	0.0225
	0.024	0.024	0.552	0.476	0.009	0.0130
Additional funding	-0.007	-0.010	0.004	0.000	-0.019	-0.020
	0.0151	0.0156	0.0212	0.0218	0.0216	0.0225
	0.655	0.522	0.861	0.993	0.386	0.3820
N	4,082	3,818	2,043	1,924	2,039	1,894
Covariates A	Y	Y	Y	Y	Y	Y
Covariates B	Y	Y	Y	Y	Y	Y

Note 1: Coefficient estimates are in the first row, standard errors are in the second, and p-values are in the third row

Exploratory Results

I next turn to exploring the difference in RCT aversion between public administrators and the general public respondents. One potential explanation is that public administrators are systematically different than their non-public administrator counterparts across demographic and socio-economic characteristics, personality, or experience. To understand if this is the case, I first test for whether public administrators are statistically different than the general public respondents. To do this, I run chi-square tests for each of the covariates I have in my primary models, comparing public administrators to non-public administrators. I find that public administrators are statistically different in seven categories: educational attainment, level of organizational authority, prior evaluation experience, ideology, age, race, and Hispanic ethnicity.

I then explore whether any of these variables can close the gap between public administrators and the general public in predicting whether they will prefer the RCT as an evaluation option. These results are shown in Table 30. As before, public administrators are 4.5 percentage points more likely to choose the RCT than their general public counterparts, statistically significant at a p-value of .003. I then run seven separate regressions, with only a single covariate at a time in addition to the public administrator indicator to see which demographic variables may be most correlated with the public administrator indicator. Ideology, race, and Hispanic ethnicity cannot explain the gap, as the coefficient for public administrator essentially remains unchanged. Educational attainment, evaluation experience, and age can partially explain the difference, shrinking the gap between public administrators and the general public to between 3.3 and 3.9 percentage points, with the coefficient remaining statistically significant at conventional levels of less than .05. However, level of organizational authority can completely explain the difference, shrinking the gap to 1.2 percentage points and eliminating the statistical significance.

Table 30: Exploratory analysis, public adminstrator correlates for selecting the RCT

	Public Administrator		
	Coefficient estimate	Standard error	P-value
No covariates	0.045	0.0151	0.0030
Educational Attainment	0.039	0.0158	0.0140
Organizational Authority	0.012	0.0204	0.5460
Evaluation Experience	0.036	0.0158	0.0210
Ideology	0.043	0.0152	0.0050

Age	0.033	0.0156	0.0340
Race	0.046	0.0152	0.0030
Hispanic ethnicity	0.044	0.0152	0.0030
N		4,110	

To explore this further, I then run a regression to understand how level of organizational authority is related to the likelihood of selecting the RCT. These results are shown in Table 31. I use four levels of organizational authority: no input, minimal decision-making authority, significant decision-making authority, and final decision-making authority. Final decision-makers are my reference category, and pick the RCT as their preferred evaluation option 41.3 percent of the time. There are no statistically significant differences between final decision makers and those with significant or minimal decision-making authority or input. There is a statistically significant difference between final decision makers and those with no input. Those with no input are 8 percentage points less likely to choose the RCT, an increase of 20 percent (0.08/0.413) in RCT aversion.

Table 31: Exploratory analysis, correlation between organizational authority and RCT selection

	Organizational Authority		
	Coefficient estimate	Standard error	P-value
Significant decision-making authority	-0.020	0.0186	0.2870
Minimal decision-making authority or input	-0.050	0.0315	0.1160
No input	-0.080	0.0212	0.0000
N		4,110	
Final decision-making authority (Constant)		0.413	

Discussion

Taken together, the experimental results provide compelling evidence of RCT aversion. Regardless of the subsample, survey participants preferred the quasi-experimental evaluation option to the RCT. Slightly less than 38% of the sample selected the RCT, on average. The degree of observed RCT aversion is very dependent on how the evaluation options are ordered. If the RCT is presented first, the gap between the quasi-experiment and the RCT is much smaller. If the RCT is ordered second, the difference is much larger. While it is difficult to distinguish whether the degree of primacy bias is driven more by inattention or indifference, the overall results remain the same even when I drop participants who complete the survey very quickly.

Interestingly, public administrators were less RCT averse than the general population participants. The non-public administrator participants selected the RCT about 35% of the time on average. Public administrators, on the other hand, chose the RCT about 40% of the time on average. In other words, public administrators were about 14 percent less RCT averse compared to their general public peers. The gap between public administrators and non-public administrators can be explained by differences in educational attainment, prior evaluation experience, age, and decision-making authority. Regardless of type of organization, employees with no input in decision-making are 20 percent more likely to be RCT averse compared to final decision makers. This provides support for my hypothesis that the level of RCT aversion will be influenced by greater exposure to evaluation.

I find mixed evidence for our hypothesis that differential treatment between intervention and comparison groups contributes to RCT aversion. I do not find evidence that the general population's evaluation option preferences were sensitive to the policy scenario element. Similarly, public administrators did not seem to be influenced by the business-as-usual or scaling path elements of the policy scenarios. However, public administrators were sensitive to the magnitude or intensity of the intervention being tested. This suggests that the intervention may be the most salient factor when making a distributive justice choice or an allocation, and that people may be more focused on the relative differences between groups, rather than absolute. Public administrators were 16 percent more RCT averse in scenario conditions where the intervention being trialed was more intense (the Healthy Home visits) compared to a less intense intervention (the mailed test kits.)

It is important to note two things about the generalizability of these results. First, this experiment was designed to test only RCT aversion in a US policy context. All of the participants were current residents of the US, and the policy scenario was written to apply to US cities and local government. While theory suggests that RCT aversion is likely to exist in other countries, the specific policy factors may significantly affect the results. For example, the role of transparency and corruption could heighten RCT aversion among public adminstrators while making it a more attractive option to the general public.

Second, this experiment tested whether participants preferred the RCT or the quasi-experiment, in the context of an impact evaluation being required. In other words, the results are conditional on there being an evaluation at all, and that evaluation needing to measure the intervention's impact. Evaluation aversion and impact evaluation aversion may look very different than the RCT aversion found in this experiment. In other words, it is possible that public administrators are more evaluation and impact evaluation averse than the general public, but when forced to choose between two impact evaluations, they prefer the RCT slightly more than the general public. While the general public may favor quasi-experiments to RCTs, they may prefer process evaluations even more. This experiment can only speak to a small branch of the RCT decision-making tree.

Another limitation of the current experiment is that I cannot be sure exactly what is driving the RCT aversion for both groups. However, because of how we constructed the scenarios, I know that at least some part of RCT aversion is related to the structure of the evaluation itself, rather than simply cost, timeline, or complexity.

This "structural" aversion may be related to the assignment mechanism or the assignment unit. If related to the assignment mechanism, it may be that people prefer evaluation designs that prioritize need or merit in the distribution of a fixed good, rather than randomization. In the case of this study, it may be that participants preferred the quasi-experiment because it prioritized people who may need the intervention most -- those living in older homes. Participants may have seen the quasi-experiment as achieving two aims in one: checking the box on an impact evaluation while also meeting the policy goal of greatest reduced risk of lead exposure.

If it is related to the assignment unit, it may be that people prefer evaluations that treat similar people in a similar way. In our scenarios, people may have preferred to treat everyone living in the same neighborhood equally, where they either all got access to the intervention or did not. In other words, people may be prioritizing equality and need allocation principles in the distribution of the lead interventions over contribution and efficiency. But how people may choose between need and equity as the primary distributive principle is not clear. And the distributive justice principles may apply very differently in different policy domains or for different issues. For example, shifting to a compliance-based intervention may prioritize equality or efficiency over contribution and need.

Future studies could explore these potential mechanisms. This could look like proposing different versions of the RCT or quasi-experimental designs. For example, alternative trial designs could target neighborhoods with the highest concentration of older homes first, and then randomize within those. Alternative quasi-experiment designs could propose a regression discontinuity with some threshold, or sorting property owners alphabetically and splitting the list in half. Ideally, we would test evaluation options in which the RCT and quasi-experimental designs are well matched on accommodating other distributive principles (such as need or merit) and on the level of aggregation of the assignment unity (such as individual or a cluster.)

Overall, the experimental results lead me to make two primary conclusions for sources of public administrator RCT aversion. The first is that public administrators are correctly assessing that RCTs are not widely accepted by the public. While they themselves may be slightly more open to RCTs on average, the public's distaste is likely to be a significant factor in making the quasi-experiment a preferred evaluation option.

Another source of RCT aversion among public administrators is related to the intervention itself. While we can't be certain about the mechanism, it seems likely that the more "valuable" or costly the intervention is perceived to be, the more public administrators may be risk averse in how they will control access to that limited good. Taken together with the knowledge that the public may dislike RCTs to begin with, it seems reasonable to assume that public administrators may be extra sensitive to potential blowback of using an RCT to study a highly-desirable intervention.

Future studies should explore this further along several avenues. The first is to attempt to replicate these results in different policy areas for different interventions. This experiment only tested a public health intervention in the particular domain of lead remediation in the US. Theoretically, we think these results would be consistent in other policy areas and contexts as well. However, it is possible that people, general public member and public administrator alike, may be more sensitive to RCTs in the high-stakes areas of public health or education, and less sensitive in areas like permitting and tax administration. The experiment could also be replicated in policy contexts with greater or lower levels of corruption to see whether this changes the results.

A second avenue to explore would be to test a wider range of intervention magnitudes. In this experiment, we were only able to test two levels: the mailed lead test kits and the healthy home visits. However, it would be interesting to widen the range of intervention. For example, a less intense intervention may have been to simply send a reminder mailer to homes prompting them to apply for free testing. A more intense intervention may have been a grant program that provided families thousands of dollars in aid to help them remediate lead paint in their homes. We would want to observe a pattern of RCT aversion increasing as the intensity of the intervention increased. If we did see this pattern replicate across a wider range, it would be additional evidence that "intervention type" is a factor in public administrator RCT aversion.

We could also explore intervention variation by manipulating the perceived effectiveness of the potential interventions. This experiment only studied two lead interventions which were unlikely to have large downsides. At worse, they may be an ineffective use of public funds and cause some minor disruption, hassle, or stress for residents who may not be at high risk for lead contamination. If the intervention may be more risky, with large upsides and downsides, it may lead people to be more in favor of the RCT as compared to the quasi-experiment. This type of intervention is a closer approximation of the clinical RCT case, where treatments can be very beneficial on average but may also have significant side effects and result in very poor outcomes for a subset of individuals. However, public administrator risk aversion may lead to avoiding these types of interventions altogether, and favoring only interventions where downside risk is low.

A third avenue for extension would be to provide different signals about the public to the public administrators, and observe how this may change their answers about preferred evaluation options. For example, one could vary how “visible” or notable the new intervention would be to the public, perhaps by telling the public administrators that it was going to be written about in the local paper. Alternatively, one could explicitly vary prior public acceptance of RCTs, perhaps by telling the public administrators that Cityville had run an RCT of similar scale previously, and it was or was not popular. If these signals prompt public administrators to select the RCT at higher or lower rates, it would confirm that “fear of public response” is a strong potential mechanism for public administrator RCT aversion.

A final avenue for additional exploration would be to test interventions specifically targeted at countering RCT aversion. One path may be to address potential overconfidence in the intervention. Providing participants additional information about how much uncertainty exists about the cost/benefit ratio of the intervention, the potential backfires for certain populations, or alternative interventions or uses of that money may cause people to rate the RCT as being more worthwhile. One hypothesis is that people may value the RCT when the tradeoffs are more explicit. For example, would people be less RCT averse if the study were designed to test the effectiveness of the mailed lead test kits compared to the Healthy Home visits?

An alternative path would be to address understanding and decision-making about research design. People may be undervaluing the benefits of the RCT because they are not correctly calibrated on evaluation risks and validity. A potential intervention may be examples of what makes a good counterfactual, or providing information about the potential confounders to the RCT and quasi-experimental designs.

While I believe this survey experiment has quite strong internal validity, it also has several limitations. The first is we should make limited generalizations from the point estimates. The participant sample was not a representative random sample drawn from the full population, and we know that context is likely to matter quite a bit in driving preferences. Additionally, this survey experiment was quite narrow in domain. I was only able to test two different interventions in a single policy area, with only two different evaluation designs. In reality, there is a whole universe of intervention options across policy areas with various combinations of evaluation designs. It is very early to say whether these results will replicate in other scenarios.

However, the survey experiment's strength is in the lack of constraints and noisy environmental factors that make it very difficult to systematically tease out why public administrators are averse to RCTs. Given that we have found evidence of significant RCT aversion in this context, where the stakes are not real, it seems very likely to be evidence that RCT aversion also exists in the wild.

Chapter 4. Conclusion

This dissertation has examined whether public administrators are averse to randomized controlled trials, and why. More specifically, it explored whether public administrators are reluctant or opposed to conducting RCTs for the evaluation of social welfare programs in situations where they would help to answer an important and useful policy question and are feasible to conduct. While RCTs are widely used to evaluate clinical interventions and practices, they are used far less often in social policy.

I define RCT aversion as a systematic preference against choosing a randomized controlled trial, in the cases where an RCT would be both practical and useful for answering a policy question of interest. It must be the result of a conscious choice to select alternatives over the randomized controlled trial. I focus this study on public administrators specifically, as they are important stakeholders in evaluation gatekeeping: they play a critical role in deciding whether an RCT proceeds or is abandoned.

Chapter 2 explored the public administrator's decision-making process as they contemplate whether to engage in program evaluation, whether to select an impact evaluation, and whether that impact evaluation should be an RCT. I adopt a rational choice model and assume that public administrators will select the evaluation option that best maximizes their utility. However, their decision-making process is likely to be more of a satisficing rather than optimizing approach. They are also likely to consider collective welfare alongside their individual best interests, and uncertainty is likely to significantly influence the public administrator to prefer options that minimized downside risks.

I then outlined nine factors that likely contribute to RCT aversion. Overall, there are very few clear incentives for public administrators to conduct evaluations, let alone impact evaluations and RCTs. At the same time, there are several clear disincentives. Risk aversion makes this dynamic worse, as the disincentives are likely to be weighed more than equivalent incentives. In other words, to favor the RCT, the public administrator must believe that the benefits far outweigh the potential downsides. This chapter's main contribution is to outline a potential research agenda to better understand the public administrator's evaluation decision-making process. Advancing this research agenda will help clarify which of the outlined factors are most influential, and their net effect.

Chapter 3 presents an empirical test of the last of these nine factors. I argue that one potential source of RCT aversion lies in the challenge it presents to distributive justice norms. The randomized controlled trial requires the public administrator to explicitly deny or withhold a social welfare intervention to an equivalent group of people on the basis of an arbitrary assignment mechanism. Other impact evaluation methods do not force the public administrator to make a similar active choice. For example, a quasi-experiment may use an arbitrary implementation date or eligibility threshold to construct an estimate of counterfactual outcomes. While these alternative allocation mechanisms may not be any fairer than randomization, they do not force the public administrator to make an explicit decision to deviate from the implementation plan. Therefore, public administrators are not likely to view quasi-experiments as a similar challenge to ethical norms around distributive justice.

Using a survey experiment with a nationally-representative sample, I investigate several factors that may influence evaluation choices and perceptions of distributive justice. Specifically, in the context of a city-run lead abatement program, I present participants with an evaluation scenario and ask them to decide between an RCT or a quasi-experimental (QE) approach. I use a vignette factorial survey methodology to answer three research questions. First, do people, on average, prefer a quasi-experiment to the RCT? Second, do features of the policy environment that create a greater perceived difference in treatment between groups contribute to more RCT aversion? And third, do preferences for the RCT differ by certain characteristics? Specifically, do public administrators have different evaluation preferences compared to their non-public administrator peers?

I find that the majority of people demonstrate a strong preference for the quasi-experiment to the RCT. Public administrators are RCT averse on average, but less so compared to their general public peers. Specifically, they are about 13 percent more likely than a group of non-public administrators to prefer an RCT to a quasi-experiment. This higher likelihood to select the RCT evaluation option can be partially explained by greater educational attainment, greater prior experience with program evaluation, and greater decision-making authority. Finally, public administrators are likely to be more RCT averse when the intervention is perceived to be very different, and potentially superior, to the status-quo option available to members of the control group. People who are not public administrators are not sensitive to features of the policy environment.

Overall, the experimental results lead me to make two primary conclusions for potential sources of public administrator RCT aversion. The first is that public administrators are correctly assessing that RCTs are not widely accepted by the public, or at least not in the domain of social policy. While they may be more open to RCTs on average, the public's objection is likely to be a significant factor in making quasi-experimental options the preferred approach. The second source of RCT aversion is related to the intervention itself. While we cannot be certain about the precise mechanism, it seems likely that the more intervention is perceived to be valuable, costly, or beneficial, the more public administrators are to be risk-averse in how they will allocate or control access to that limited good. Taken together with the finding that the public dislikes RCTs in general, public administrators are likely to be extra sensitive to political opposition when using an RCT to study a highly-desirable intervention. The chapter concludes by outlining several avenues for future exploration of public administrator RCT aversion.

The survey experiment suffers from several limitations, namely that it is only able to test a very specific policy scenario, and this scenario does not carry the weight and complexity of real-world decision-making. However, the strength of the design lies in the lack of constraints and noisy contextual factors that would make it very difficult to tease out individual dimensions of public administrator RCT aversion. Given that I have found evidence of significant RCT aversion in this context, where the stakes are not real, it seems very likely that RCT aversion exists in the wild.

These results have several larger implications for social policy researchers, evidence-based policy advocates, and public policy and administration educators. First, researchers engaging with policy practitioners and public administrators with the aim of evaluating public programs should know that the RCT and policy experiment may not be the default preference for evaluation methodology. Researchers may be able to have more constructive conversations by starting from a place of understanding that there are likely a host of reasonable explanations for hesitancy around the use of RCTs. If we want public administrators and decision-makers to engage with us, we should be prepared to speak to these factors, and ideally, find ways to alleviate them. This may be through reducing costs, providing professional incentives and rewards, or helping to structure an assessment about whether the RCT violates ethical or moral norms compared to the status quo approach. Future investigations could test interventions to support researchers in having these conversations and enable better academic-practitioner collaboration.

Second, advocates for evidence-based policy likely need to think about additional ways to reduce the downsides of evaluation more broadly, and RCTs specifically, and maximize the upsides. The Evidence Act is a step in the right direction, by making data sharing easier and setting explicit performance expectations around evaluation agendas, and appointing a specific role to encourage and manage evaluations. However, it is still too early to know how much of an impact the legislation will have, and the Act only applies at the federal level. Risk aversion will cause a strong bias towards the status quo. Because the risks will have a greater weight than equivalent gains, more is likely needed to shift the public administrator's utility calculation in favor of more evaluation, impact evaluations, and RCTs. Further research can hopefully help to unpack which of the nine influencing factors are likely the most determinant and in what contexts. Teasing this out is likely to be quite complicated, however, as a change in one factor may affect another. For example, mandating evaluation may crowd out intrinsic motivation. Understanding these net effects, and designing policies or interventions to incentivize evidence generation, will not be simple.

Finally, the results have implications for public policy and administration educators. In particular, educational attainment and prior evaluation experience are influential predictors in decreased RCT aversion. Given the high degree of RCT aversion among the general public, this raises the potential impact of incorporating program evaluation basics and how to assess evidence into core curriculums, at either the high school or college level. For those focused on preparing future public administrators, this study's results suggest a need to focus not just on the mechanics of evaluation, but also on the implementation of evaluation in the real world. Conducting impact evaluations requires nearly as many stakeholder management skills as it does technical or methodological capability. Public administration students should be asked to discuss and carefully deliberate on the ethical considerations surrounding program implementation and evaluation, alongside a consideration of its practical mechanics. Just because the RCT makes ethical tradeoffs around the distribution of fixed resources more salient does not mean there isn't a default, if hidden, ethical choice reflected in other evaluation approaches or when there isn't even an evaluation at all.

Appendices

Appendix 1: Screener Questions

[age] What is your age? Please enter a numeric response only.

- Number from …
- Exclude if <18

Response items in red are excluded

[gender] What is your gender?

- Male
- Female

[region] What is your region?

- Midwest
- Northeast
- South
- West

[race] What is your race?

- White
- Black or African American
- Native American or Alaska Native
- Asian (Select Asian Group)
    - Asian Indian
    - Chinese
    - Filipino
    - Japanese
    - Korean
    - Vietnamese
    - Other
- Pacific Islander (Select Pacific Islander Group)
    - Native Hawaiian
    - Guamanian
    - Samoan
    - Other Pacific Islander
- Some other race
- Prefer not to answer

Response items in red are excluded

[hispanic] Are you of Hispanic, Latino, or Spanish origin?

- No, not of Hispanic, Latino, or Spanish origin
- Yes, Mexican, Mexican American, Chicano
- Yes, Cuban

- Yes Puerto Rican
- Yes, another Hispanic, Latino, or Spanish origin [Select Country of Choice]
- Prefer not to answer

Response items in red are excluded

[employmentStatus] What is your current employment status?

- Employed full-time
- Employed part-time
- Self-employed full-time
- Self-employed part-time
- Active military
- Inactive military/veteran
- Temporarily unemployed
- Full-time homemaker
- Retired
- Student
- Disabled
- Prefer not to answer

Response items in blue are required for entry into the “Public Administrator” sample

[organizationType] Which of the following best describes your organization?

- A for-profit, business-to-business company
- A for-profit, business-to-consumer company
- A for-profit company that is both business-to-business and business-to-consumer
- A non-profit organization
- A governmental organization
- None of the above

Response items in blue are required for entry into the “Public Administrator” sample

[org Authority] What level of decision-making authority do you have on how resources are used within your department, office, or organization?

- Final decision-making authority (individually or as part of a group)
- Significant decision-making or influence (individually or as part of a group)
- Minimal decision-making or influence
- No input
- Prefer not to say

Response items in blue are required for entry into the “Public Administrator” sample

Appendix 2: Survey Text

Page 1

Viewed by all participants

Welcome and thank you for participating in this survey.

Task: In this survey, we are going to ask you some multiple choice questions about the evaluation of public programs.

Duration: The survey should take between 6 to 8 minutes to complete and requires your attention, so please only participate if you can dedicate this time.

Please note that you cannot go back to previous pages.

Page 2

Viewed by all participants

For our research, careful attention to survey questions is critical! To show that you are paying attention, please select "I have a question."

- I understand
- I do not understand
- I have a question

Page 3

Viewed by participants who answered first attention check incorrectly

You didn’t select the correct answer to our last question. Your attention to the survey questions is very important for our research, so we’d like to give you another chance to respond. To show that you are paying attention, please select “I have a question.”

- I understand
- I do not understand
- I have a question

Page 5

Viewed by participants who answered second attention check incorrectly

You have answered our questions incorrectly. We can only accept surveys from people who are paying close attention, so we have ended this survey early. Please click ‘Next’ to return to your panel website.

Page 6

Viewed by all participants who passed the attention check

Cityville wants to use federal funds to improve resident health.

Cityville housing is old, with most houses built between 1900 and 1950. Older homes are at higher risk for various environmental health hazards. These include lead, radon, and mold.

Lead is difficult to detect because it is odorless and invisible. It is also dangerous to children. It can cause:

- Damage to the brain and nervous system
- Slowed growth and development
- Learning and behavior problems
- Hearing and speech problems

Lead exposure is preventable. Residents need to first identify lead sources, and then either control or remove them.

Governments can help by providing access to education, testing, and resources. If implemented well, these lead programs work to improve resident health. They can also result in higher academic achievement and reduced crime.

Page 7 - Scenario 1

Viewed by participants assigned to scenario 1 who passed the attention check

While Cityville has a lead screening program, it's not having much impact. The current program requires residents to submit an application for free testing, and less than 1 percent of Cityville households have applied.

So Cityville wants to try a new service, called a "healthy home assessment." Community health workers will visit resident homes and conduct free screening for health hazards, including lead. If the worker identifies an issue, they will also contain or fix it immediately. Cityville hopes that this new program will reduce the number of children exposed to lead. They predict that it will cost an average of $125 per household.

The federal funding requires Cityville to measure the program's impact. The City Manager wants to know whether the program worked to:

- Improve resident access to quality housing
- Improve resident health outcomes

The City Manager's staff have developed two evaluation options. They are expected to cost the same and offer the new service to the same number of residents.

Option A	**Option B**
● Start with all addresses from tax records. ● **Randomly pick** one half to a treatment group and the other to a control group. ○ The treatment group will be offered the new healthy home assessment.	● Start with all addresses from tax records. ● **Select neighborhoods** with the greatest number of older homes. ○ Selected neighborhoods will be offered the new healthy home assessment.

○ The control group won’t be offered the new healthy home assessment. ● Continue to provide everyone access to the current lead abatement program: Any resident can apply for free testing. ● After one year, compare health outcomes for both groups.	○ All other Cityville neighborhoods will be the comparison group. They won’t be offered the new healthy home assessment. ● Continue to provide everyone access to the current lead abatement program: Any resident can apply for free testing. ● After one year, compare health outcomes for the different neighborhoods.

If the healthy home assessments are successful, Cityville will offer them to everyone.

Which evaluation option would you choose?
- Option A
- Option B

Page 7 - Scenario 2

Viewed by participants assigned to scenario 2 who passed the attention check

Cityville doesn’t currently have a lead screening program.

So Cityville wants to try a new service, called a "healthy home assessment." Community health workers will visit resident homes and conduct free screening for health hazards, including lead. If the worker identifies an issue, they will also contain or fix it immediately. Cityville hopes that this new program will reduce the number of children exposed to lead. They predict that it will cost an average of $125 per household.

The federal funding requires Cityville to measure the program's impact. The City Manager wants to know whether the program worked to:
- Improve resident access to quality housing
- Improve resident health outcomes

The City Manager's staff have developed two evaluation options. They are expected to cost the same and offer the new service to the same number of residents.

Option A	**Option B**
● Start with all addresses from tax records. ● **Randomly pick** one half to a treatment group and the other to a control group. ○ The treatment group will be offered the new healthy home assessment. ○ The control group won’t be offered the new healthy home assessment.	● Start with all addresses from tax records. ● **Select neighborhoods** with the greatest number of older homes. ○ Selected neighborhoods will be offered the new healthy home assessment. ○ All other Cityville neighborhoods will be the comparison group. They won’t

● After one year, compare health outcomes for both groups.	be offered the new healthy home assessment. ● After one year, compare health outcomes for the different neighborhoods.

If the healthy home assessments are successful, Cityville will offer them to everyone.

Which evaluation option would you choose?

- Option A
- Option B

Page 7 - Scenario 3

Viewed by participants assigned to scenario 3 who passed the attention check

While Cityville has a lead screening program, it's not having much impact. The current program requires residents to submit an application for free testing, and less than 1 percent of Cityville households have applied.

So Cityville wants to try a new service, called a "healthy home assessment." Community health workers will visit resident homes and conduct free screening for health hazards, including lead. If the worker identifies an issue, they will also contain or fix it immediately. Cityville hopes that this new program will reduce the number of children exposed to lead. They predict that it will cost an average of $125 per household.

The federal funding requires Cityville to measure the program's impact. The City Manager wants to know whether the program worked to:

- Improve resident access to quality housing
- Improve resident health outcomes

The City Manager's staff have developed two evaluation options. They are expected to cost the same and offer the new service to the same number of residents.

Option A	**Option B**
● Start with all addresses from tax records. ● **Randomly pick** one half to a treatment group and the other to a control group. ○ The treatment group will be offered the new healthy home assessment. ○ The control group won't be offered the new healthy home assessment. ● Continue to provide everyone access to the current lead abatement program: Any resident can apply for free testing.	● Start with all addresses from tax records. ● **Select neighborhoods** with the greatest number of older homes. ○ Selected neighborhoods will be offered the new healthy home assessment. ○ All other Cityville neighborhoods will be the comparison group. They won't be offered the new healthy home assessment.

● After one year, compare health outcomes for both groups.	● Continue to provide everyone access to the current lead abatement program: Any resident can apply for free testing. ● After one year, compare health outcomes for the different neighborhoods.
If the healthy home assessments are successful, Cityville will apply for more funding to keep the program running. Which evaluation option would you choose? ● Option A ● Option B	

Page 7 - Scenario 4

Viewed by participants assigned to scenario 4 who passed the attention check

Cityville doesn't currently have a lead screening program.

So Cityville wants to try a new service, called a "healthy home assessment." Community health workers will visit resident homes and conduct free screening for health hazards, including lead. If the worker identifies an issue, they will also contain or fix it immediately. Cityville hopes that this new program will reduce the number of children exposed to lead. They predict that it will cost an average of $125 per household.

The federal funding requires Cityville to measure the program's impact. The City Manager wants to know whether the program worked to:

- Improve resident access to quality housing
- Improve resident health outcomes

The City Manager's staff have developed two evaluation options. They are expected to cost the same and offer the new service to the same number of residents.

Option A	**Option B**
● Start with all addresses from tax records. ● **Randomly pick** one half to a treatment group and the other to a control group. ○ The treatment group will be offered the new healthy home assessment. ○ The control group won't be offered the new healthy home assessment. ● After one year, compare health outcomes for both groups.	● Start with all addresses from tax records. ● **Select neighborhoods** with the greatest number of older homes. ○ Selected neighborhoods will be offered the new healthy home assessment. ○ All other Cityville neighborhoods will be the comparison group. They won't be offered the new healthy home assessment. ● After one year, compare health outcomes for the different neighborhoods.

If the healthy home assessments are successful, Cityville will apply for more funding to keep the program running.

Which evaluation option would you choose?

- Option A
- Option B

Page 7 - Scenario 5

Viewed by participants assigned to scenario 5 who passed the attention check

While Cityville has a lead screening program, it's not having much impact. The current program requires residents to submit an application for free testing, and less than 1 percent of Cityville households have applied.

So Cityville wants to try mailing free test kits to residents. Families would receive two instant lead paint testing kits. These kits allow families to test surfaces and detect whether there is lead in seconds. Cityville hopes that this new program will reduce the number of children exposed to lead. They predict that it will cost an average of $25 per household.

The federal funding requires Cityville to measure the program's impact. The City Manager wants to know whether the program worked to:

- Improve resident access to quality housing
- Improve resident health outcomes

The City Manager's staff have developed two evaluation options. They are expected to cost the same and offer the new service to the same number of residents.

Option A

- Start with all addresses from tax records.
- **Randomly pick** one half to a treatment group and the other to a control group.
    - The treatment group will be mailed the free test kits.
    - The control group won't be mailed the free test kits.
- Continue to provide everyone access to the current lead abatement program: Any resident can apply for free testing.
- After one year, compare health outcomes for both groups.

Option B

- Start with all addresses from tax records.
- **Select neighborhoods** with the greatest number of older homes.
    - Selected neighborhoods will be mailed the free test kits.
    - All other Cityville neighborhoods will be the comparison group. They won't be mailed the free test kits.
- Continue to provide everyone access to the current lead abatement program: Any resident can apply for free testing.
- After one year, compare health outcomes for the different neighborhoods.

If the two free test kits are successful, Cityville will mail them to everyone.

Which evaluation option would you choose?

- Option A
- Option B

Page 7 - Scenario 6

Viewed by participants assigned to scenario 6 who passed the attention check

Cityville doesn't currently have a lead screening program.

So Cityville wants to try mailing free test kits to residents. Families would receive two instant lead paint testing kits. These kits allow families to test surfaces and detect whether there is lead in seconds. Cityville hopes that this new program will reduce the number of children exposed to lead. They predict that it will cost an average of $25 per household.

The federal funding requires Cityville to measure the program's impact. The City Manager wants to know whether the program worked to:

- Improve resident access to quality housing
- Improve resident health outcomes

The City Manager's staff have developed two evaluation options. They are expected to cost the same and offer the new service to the same number of residents.

Option A

- Start with all addresses from tax records.
- **Randomly pick** one half to a treatment group and the other to a control group.
  - The treatment group will be mailed the free test kits.
  - The control group won't be mailed the free test kits.
- After one year, compare health outcomes for both groups.

Option B

- Start with all addresses from tax records.
- **Select neighborhoods** with the greatest number of older homes.
  - Selected neighborhoods will be mailed the free test kits.
  - All other Cityville neighborhoods will be the comparison group. They won't be mailed the free test kits.
- After one year, compare health outcomes for the different neighborhoods.

If the two free test kits are successful, Cityville will mail them to everyone.

Which evaluation option would you choose?

- Option A
- Option B

Page 7 - Scenario 7

Viewed by participants assigned to scenario 7 who passed the attention check

While Cityville has a lead screening program, it's not having much impact. The current program requires residents to submit an application for free testing, and less than 1 percent of Cityville households have applied.

So Cityville wants to try mailing free test kits to residents. Families would receive two instant lead paint testing kits. These kits allow families to test surfaces and detect whether there is lead in seconds. Cityville hopes that this new program will reduce the number of children exposed to lead. They predict that it will cost an average of $25 per household.

The federal funding requires Cityville to measure the program's impact. The City Manager wants to know whether the program worked to:

- Improve resident access to quality housing
- Improve resident health outcomes

The City Manager's staff have developed two evaluation options. They are expected to cost the same and offer the new service to the same number of residents.

Option A

- Start with all addresses from tax records.
- **Randomly pick** one half to a treatment group and the other to a control group.
  - The treatment group will be mailed the free test kits.
  - The control group won't be mailed the free test kits.
- Continue to provide everyone access to the current lead abatement program: Any resident can apply for free testing.
- After one year, compare health outcomes for both groups.

Option B

- Start with all addresses from tax records.
- **Select neighborhoods** with the greatest number of older homes.
  - Selected neighborhoods will be mailed the free test kits.
  - All other Cityville neighborhoods will be the comparison group. They won't be mailed the free test kits.
- Continue to provide everyone access to the current lead abatement program: Any resident can apply for free testing.
- After one year, compare health outcomes for the different neighborhoods.

If the two free test kits are successful, Cityville will apply for more funding to keep the program running.

Which evaluation option would you choose?

- Option A
- Option B

Page 7 - Scenario 8

Viewed by participants assigned to scenario 8 who passed the attention check

Cityville doesn't currently have a lead screening program.

So Cityville wants to try mailing free test kits to residents. Families would receive two instant lead paint testing kits. These kits allow families to test surfaces and detect whether there is lead in seconds.

Cityville hopes that this new program will reduce the number of children exposed to lead. They predict that it will cost an average of $25 per household.

The federal funding requires Cityville to measure the program's impact. The City Manager wants to know whether the program worked to:

- Improve resident access to quality housing
- Improve resident health outcomes

The City Manager's staff have developed two evaluation options. They are expected to cost the same and offer the new service to the same number of residents.

Option A

- Start with all addresses from tax records.
- **Randomly pick** one half to a treatment group and the other to a control group.
  - The treatment group will be mailed the free test kits.
  - The control group won't be mailed the free test kits.
- After one year, compare health outcomes for both groups.

Option B

- Start with all addresses from tax records.
- **Select neighborhoods** with the greatest number of older homes.
  - Selected neighborhoods will be mailed the free test kits.
  - All other Cityville neighborhoods will be the comparison group. They won't be mailed the free test kits.
- After one year, compare health outcomes for the different neighborhoods.

If the two free test kits are successful, Cityville will apply for more funding to keep the program running.

Which evaluation option would you choose?

- Option A
- Option B

Page 8

Viewed by all participants who passed the attention check

[ideology] When it comes to politics, would you describe yourself as liberal, conservative, or neither liberal nor conservative?

- Liberal
- Conservative
- Neither liberal nor conservative

[If answers "Liberal"] Would you call yourself very liberal or not very liberal?

- Very liberal
- Not very liberal

[If answers "Conservative"] Would you call yourself very conservative or not very conservative?

- Very conservative
- Not very conservative

[If answers "Neither"] Do you think of yourself as closer to liberals, or conservatives, or neither of these?

- Closer to liberals
- Closer to conservatives
- Neither of these

[politicalAffiliation] In politics today, do you consider yourself a Democrat, Republican, or Independent?

- Democrat
- Republican
- Independent
- Prefer not to answer

[personality] Here are a number of personality traits that may or may not apply to you. You should rate the extent to which the pair of traits applies to you, even if one characteristic applies more strongly than the other.

- Extraverted, enthusiastic
- Critical, quarrelsome
- Dependable, self-disciplined
- Anxious, easily upset
- Open to new experiences, complex
- Reserved, quiet
- Sympathetic, warm
- Disorganized, careless
- Calm, emotionally stable
- Conventional, uncreative

Item responses matrix 1-7

- Disagree strongly
- Disagree moderately
- Disagree a little
- Neither agree nor disagree
- Agree a little
- Agree moderately
- Agree Strongly

[eduAttainment] What is the highest level of education you have completed?

- 3rd Grade or less
- Middle School - Grades 4 - 8
- Completed some high school
- High school graduate
- Other post high school vocational training
- Completed some college, but no degree
- Associate Degree
- College Degree (such as B.A., B.S.)
- Completed some graduate, but no degree
- Masters degree
- Doctorate degree
- None of the above

[industry] Which of the following best describes your primary industry?

- Accommodation and Food services
- Administration and Support services
- Agriculture, Forestry, Fishing, and Hunting
- Arts, Entertainment, and Recreation
- Construction
- Educational services
- Finance and Insurance
- Government
- Health care and Social assistance
- Information
- Management of companies and enterprises
- Manufacturing
- Mining, Quarrying, and Oil and gas extraction
- Other services
- Professional, Scientific and Technical services
- Real estate and Rental and leasing
- Retail trade
- Transportation and Warehousing
- Utilities
- Wholesale trade

[evalExperience] Have you ever participated in an effort to evaluate a program, policy change, or new practice? For example, helping to design the investigation, implement it, or analyze the results.

- Yes
- No
- I’m not sure

Page 9

Viewed by all participants who passed the attention check

This is the end of the survey. Thanks for participating!
If you have any feedback, please enter it in the box below.

[free text box]

Please click ‘submit’ below to return back to the panel website.

References

Abraham, K., & Haskins, R. (2017). *The Promise of Evidence-Based Policymaking: Report of the Commission on Evidence-Based Policymaking*. Commission on Evidence-Based Policymaking. https://www2.census.gov/adrm/fesac/2017-12-15/Abraham-CEP-final-report.pdf

Akalis, S. A. (2008). A new spin on losses looming larger than gains: asymmetric implicit associations from slot machine experience. *Journal of Behavioral Decision Making*, *21*(4), 378–398.

Alexander, C. S., & Becker, H. J. (1978). The Use of Vignettes in Survey Research. *Public Opinion Quarterly*, *42*(1), 93–104.

Alvarez, M. R., Atkeson, L. R., Levin, I., & Li, Y. (2019). Paying Attention to Inattentive Survey Respondents. *Political Analysis*, *27*(2), 145–162.

Anduiza, E., & Galais, C. (2016). Answering Without Reading: IMCs and Strong Satisficing in Online Surveys. *International Journal of Public Opinion Research*, *29*(3), 497–519.

Arno, V. H., Koen, A., & Bart, M. (2020). Differentiated Distributive Justice Preferences? Configurations of Preferences for Equality, Equity and Need in Three Welfare Domains. *Social Justice Research; New York*, *33*(3), 257–283.

Aronow, P. M., Kalla, J., Orr, L., & Ternovski, J. (2020). *Evidence of Rising Rates of Inattentiveness on Lucid in 2020*. https://doi.org/10.31235/osf.io/8sbe4

Arrow, K. (2004). Is Bounded Rationality Unboundedly Rational? Some Ruminations. In M. Augier & J. G. March (Eds.), *Models of a man: Essays in memory of Herbert A. Simon* (pp. 47–56). MIT Press.

Baker, R., Brick, J. M., Bates, N. A., Battaglia, M., Couper, M. P., Dever, J. A., Gile, K. J., & Tourangeau, R. (2013). Summary Report of the AAPOR Task Force on Non-probability Sampling. *Journal of Survey Statistics and Methodology*, *1*(2), 90–143.

Baldassarri, D., & Abascal, M. (2017). Field experiments across the social sciences. *Annual Review of Sociology*, *43*(1), 41–73.

Baron, R. A., & Ensley, M. D. (2006). Opportunity recognition as the detection of meaningful patterns:

Evidence from comparisons of novice and experienced entrepreneurs. *Management Science*, *52*(9), 1331–1344.

Bellante, D., & Link, A. N. (1981). Are public sector workers more risk averse than private sector workers? *Industrial & Labor Relations Review*, *34*(3), 408–412.

Bénabou, R., & Tirole, J. (2003). Intrinsic and Extrinsic Motivation. *The Review of Economic Studies*, *70*(3), 489–520.

Berinsky, A. J., Margolis, M. F., & Sances, M. W. (2014). Separating the shirkers from the workers? Making sure respondents pay attention on self-administered surveys. *American Journal of Political Science*, *58*(3), 739–753.

Bicchieri, C. (2005). *The Grammar of Society: The Nature and Dynamics of Social Norms*. Cambridge University Press.

Bicchieri, C., & Muldoon, R. (2011, March 1). *Social norms*. Stanford Encyclopedia of Philosophy Archive. https://stanford.library.sydney.edu.au/archives/sum2016/entries/social-norms/

Bloomberg Philanthropies. (2017). *What Works Cities Certification*. Bloomberg Philanthropies What Works Cities. https://whatworkscities.bloomberg.org/certification/

Bordalo, P., Gennaioli, N., & Shleifer, A. (2013). Salience and Consumer Choice. *The Journal of Political Economy*, *121*(5), 803–843.

Boruvka, E., & Perry, J. L. (2020). Understanding evolving public motivational practices: An institutional analysis. *Governance* , *33*(3), 565–584.

Brick, J. M. (2014). Explorations in non-probability sampling using the web. *Proceedings of the Conference on beyond Traditional Survey Taking: Adapting to a Changing World*, 1–6.

Buskens, V. (2015). Rational Choice Theory in Sociology. In J. D. Wright (Ed.), *International Encyclopedia of the Social & Behavioral Sciences (Second Edition)* (pp. 901–906). Elsevier.

Caspi, A., Roberts, B. W., & Shiner, R. L. (2005). Personality development: stability and change. *Annual Review of Psychology*, *56*, 453–484.

Cialdini, R. B., Kallgren, C. A., & Reno, R. R. (1991). A Focus Theory of Normative Conduct: A

Theoretical Refinement and Reevaluation of the Role of Norms in Human Behavior. In M. P. Zanna (Ed.), *Advances in Experimental Social Psychology* (Vol. 24, pp. 201–234). Academic Press.

Coleman, J. S. (1988). Social Capital in the Creation of Human Capital. *The American Journal of Sociology*, *94*, S95–S120.

Coleman, J. S. (1990). The emergence of norms. In M. Hechter, K.-D. Opp, & R. Wippler (Eds.), *Social Institutions: Their Emergence, Maintenance and Effects* (pp. 35–60). Routledge.

Cook, P. J., & Ludwig, J. (2006). Aiming for evidence-based gun policy. *Journal of Policy Analysis and Management*, *25*(3), 691–735.

Cook, T. D., & Payne, M. R. (2002). Objecting to the Objections to Using Random Assignment in Educational Research. In F. Mosteller & R. Boruch (Eds.), *Evidence Matters* (pp. 150–178). Brookings Institution Press.

Coppock, A., & McClellan, O. A. (2019). Validating the demographic, political, psychological, and experimental results obtained from a new source of online survey respondents. *Research & Politics*, *6*. https://doi.org/10.1177/2053168018822174

Costa, P. T., & McCrae, R. R. (1992). The five-factor model of personality and its relevance to personality disorders. *Journal of Personality Disorders*, *6*(1992), 343–359.

Crewson, P. E. (1997). Public-Service Motivation: Building Empirical Evidence of Incidence and Effect. *Journal of Public Administration Research and Theory*, *7*(4), 499–518.

Deaton, A., & Cartwright, N. (2018). Understanding and misunderstanding randomized controlled trials. *Social Science & Medicine*, *210*, 2–21.

Deci, E. L. (1976). The hidden costs of rewards. *Organizational Dynamics*, *4*(3), 61–72.

Deci, E. L., Koestner, R., & Ryan, R. M. (1999). A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psychological Bulletin*, *125*(6), 627–668; discussion 692–700.

DellaVigna, S., & Linos, E. (2022). RCTs to scale: Comprehensive evidence from two Nudge Units. *Econometrica: Journal of the Econometric Society*, *90*(1), 81–116.

Di Domenico, S. I., & Ryan, R. M. (2017). The Emerging Neuroscience of Intrinsic Motivation: A New Frontier in Self-Determination Research. *Frontiers in Human Neuroscience*, *11*, 145.

Djulbegovic, B., Lacevic, M., Cantor, A., Fields, K. K., Bennett, C. L., Adams, J. R., Kuderer, N. M., & Lyman, G. H. (2000). The uncertainty principle and industry-sponsored research. *The Lancet*, *356*(9230), 635–638.

Dong, H.-K. D. (2017). Individual Risk Preference and Sector Choice: Are Risk-Averse Individuals More Likely to Choose Careers in the Public Sector? *Administration & Society*, *49*(8), 1121–1142.

Eisenhardt, K. M. (1989). Agency Theory: An Assessment and Review. *Academy of Management Review. Academy of Management*, *14*(1), 57–74.

Elster, J. (1993). *Local Justice: How Institutions Allocate Scarce Goods and Necessary Burdens*. Russell Sage Foundation.

Evans, J. R., & Mathur, A. (2018). The value of online surveys: a look back and a look ahead. *Internet Research*, *28*(4), 854–887.

Fehr, & Schmidt. (1999). A theory of fairness, competition, and cooperation. *The Quarterly Journal of Economics*. https://academic.oup.com/qje/article-abstract/114/3/817/1848113

Ferreira, J. A., & Zwinderman, A. H. (2006). On the Benjamini–Hochberg method. *Annals of Statistics*, *34*(4), 1827–1849.

Fong, E. A., & Tosi, H. L. (2007). Effort, Performance, and Conscientiousness: An Agency Theory Perspective. *Journal of Management*, *33*(2), 161–179.

Franco, A., Malhotra, N., & Simonovits, G. (2014). Publication bias in the social sciences: *Science*, *345*(6203), 1502–1505.

Frederick, S., Loewenstein, G., & O'Donoghue, T. (2002). Time Discounting and Time Preference: A Critical Review. *Journal of Economic Literature*, *40*(2), 351–401.

Gächter, S., Johnson, E. J., & Herrmann, A. (2022). Individual-level loss aversion in riskless and risky choices. *Theory and Decision*, *92*(3), 599–624.

Gigerenzer, G. (2008). *Rationality for Mortals: How People Cope with Uncertainty*. Oxford University

Press.

Gigerenzer, G. (2010). Moral satisficing: rethinking moral behavior as bounded rationality. *Topics in Cognitive Science*, *2*(3), 528–554.

Gigerenzer, G., & Selten, R. (Eds.). (2001). *Bounded Rationality: The Adaptive Toolbox*. MIT Press.

Gosling, S. D., Rentfrow, P. J., & Swann, W. B. (2003). A very brief measure of the Big-Five personality domains. *Journal of Research in Personality*, *37*(6), 504–528.

Gould, E. (2009). Childhood lead poisoning: conservative estimates of the social and economic benefits of lead hazard control. *Environmental Health Perspectives*, *117*(7), 1162–1167.

Granovetter, M. (2002). Economic action and social structure: The problem of embeddedness. *Journal of Economic Sociology*, *3*(3), 44–58.

Gueron, J. M. (2002). The Politics of Random Assignment: Implementing Studies and Affecting Policy. In F. Mosteller & R. Boruch (Eds.), *Evidence Matters: Randomized Trials in Education Research* (pp. 15–49). Brookings Institution Press.

Guesnerie, R., & Roberts, K. (1984). Effective Policy Tools and Quantity Controls. *Econometrica: Journal of the Econometric Society*, *52*(1), 59–86.

Hainmueller, J., Hangartner, D., & Yamamoto, T. (2015). Validating vignette and conjoint survey experiments against real-world behavior. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(8), 2395–2400.

Hainmueller, J., Hopkins, D. J., & Yamamoto, T. (2014). Causal Inference in Conjoint Analysis: Understanding Multidimensional Choices via Stated Preference Experiments. *Political Analysis: An Annual Publication of the Methodology Section of the American Political Science Association*, *22*(1), 1–30.

Hart, N., & Davis, S. (2017, November 30). *FACT SHEET: Foundations for Evidence-Based Policymaking Act*. Bipartisan Policy Center. https://bipartisanpolicy.org/blog/fact-sheet-foundations-for-evidence-based-policymaking-act/

Haskins, R. (2018). Evidence-Based Policy: The Movement, the Goals, the Issues, the Promise. *The*

Annals of the American Academy of Political and Social Science, *678*(1), 8–37.

Head, B. W. (2015). Policy Analysis: Evidence Based Policy-Making. In J. D. Wright (Ed.), *International Encyclopedia of the Social & Behavioral Sciences (Second Edition)* (pp. 281–287). Elsevier.

Heath, C., & Tversky, A. (1991). Preference and Belief: Ambiguity and Competence in Choice under Uncertainty. *Journal of Risk and Uncertainty*, *4*(1), 5–28.

Heckman, J. J., & Smith, J. A. (1995). Assessing the Case for Social Experiments. *The Journal of Economic Perspectives: A Journal of the American Economic Association*, *9*(2), 85–110.

Heck, P. R., Chabris, C. F., Watts, D. J., & Meyer, M. N. (2020). Objecting to experiments even while approving of the policies or treatments they compare. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(32), 18948–18950.

Heinrich, C. J. (2007). Evidence-Based Policy and Performance Management: Challenges and Prospects in Two Parallel Movements. *American Review of Public Administration*, *37*(3), 255–277.

Hill, I., Hawkes, C., Harrington, M., Bajaj, R., Black, W., Fasciano, N., Howell, E., Kapustka, H., & Lutzky, A. W. (2003). *Congressionally Mandated Evaluation of the State Children's Health Insurance Program: Final Cross-Cutting Report on the Findings from Ten State Site Visits* (No. 8782-072). Mathematica Policy Research, Inc.; The Urban Institute. https://www.aspe.hhs.gov/sites/default/files/migrated_legacy_files//138126/Final%20Cross-Cutting.pdf

Holt, S. B. (2018). For those who care: The effect of public service motivation on sector selection. *Public Administration Review*, *78*(3), 457–471.

Jachimowicz, J. M., Duncan, S., Weber, E. U., & Johnson, E. J. (2019). When and why defaults influence decisions: a meta-analysis of default effects. *Behavioural Public Policy*, 1–28.

Johnson, E. J., & Goldstein, D. (2003). Do Defaults Save Lives? *Science*, *302*(5649), 1338–1339.

Johnson, E. J., Hassin, R., Baker, T., Bajger, A. T., & Treuer, G. (2013). Can consumers make affordable care affordable? The value of choice architecture. *PloS One*, *8*(12), e81521.

Kahneman, D. (2003). Maps of bounded rationality: Psychology for behavioral economics. *The American*

Economic Review, *93*(5), 1449–1475.

Kahneman, D., Knetsch, J. L., & Thaler, R. H. (1991). Anomalies: The Endowment Effect, Loss Aversion, and Status Quo Bias'(1991). *Journal of Economic Perspectives*, *5*(1), 193–206.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decisions under risk. *Econometrica*, *47*(2), 263–292.

Lamm, H., & Schwinger, T. (1980). Norms Concerning Distributive Justice: Are Needs Taken Into Consideration in Allocation Decisions? *Social Psychology Quarterly*, *43*(4), 425–429.

LeBoeuf, R. A., & Shafir, E. B. (2005). Decision Making. In K. J. Holyoak (Ed.), *The Cambridge handbook of thinking and reasoning* (Vol. 858, pp. 243–265). Cambridge University Press.

Madrian, B. C., & Shea, D. F. (2001). The Power of Suggestion: Inertia in 401(k) Participation and Savings Behavior. *The Quarterly Journal of Economics*, *116*(4), 1149–1187.

Massey, J., & Straussman, J. D. (1985). Another Look at the Mandate Issue: Are Conditions-of-Aid Really so Burdensome? *Public Administration Review*, *45*(2), 292–300.

Maynard, R. A. (2006). Presidential address: Evidence-based decision making: What will it take for the decision makers to care? *Journal of Policy Analysis and Management*, *25*(2), 249–265.

McCrae, R. R., & Costa, P. T., Jr. (2008). The five-factor theory of personality. *Handbook of Personality: Theory and Research., 3rd Ed.*, *3*, 159–181.

Mehmood, S., Naseer, S., & Chen, D. L. (2021). *Training policymakers in econometrics*. Technical report. Working Paper. https://users.nber.org/~dlchen/papers/Training_Policy_Makers_in_Econometrics.pdf

Meyer, M. N., Heck, P. R., Holtzman, G. S., Anderson, S. M., Cai, W., Watts, D. J., & Chabris, C. F. (2019). Objecting to experiments that compare two unobjectionable policies or treatments. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(22), 10723–10728.

Miller, D. (1991). *Principles of Social Justice*. Harvard University Press.

Mislavsky, R., Dietvorst, B., & Simonsohn, U. (2019). Critical Condition: People Don't Dislike a Corporate Experiment More Than They Dislike Its Worst Condition. *Marketing Science*.

https://doi.org/10.1287/mksc.2019.1166

O'Donoghue, T., & Rabin, M. (1999). Doing It Now or Later. *The American Economic Review*, *89*(1), 103–124.

Office of the Assistant Secretary for Planning and Evaluation. (n.d.). *Implementing the Foundations for Evidence-Based Policymaking Act at the U.S. Department of Health & Human Services*. ASPE, U.S. Department of Health and Human Services. Retrieved June 24, 2022, from https://aspe.hhs.gov/topics/data/evidence-act-0

Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology*, *45*(4), 867–872.

Orr, L. L., Olsen, R. B., Bell, S. H., Schmid, I., Shivji, A., & Stuart, E. A. (2019). Using the results from rigorous multisite evaluations to inform local policy decisions. *Journal of Policy Analysis and Management*, *38*(4), 978–1003.

Perrin, A., & Atske, S. (2021, April 2). *7% of Americans don't use the internet. Who are they?* Pew Research Center. https://www.pewresearch.org/fact-tank/2021/04/02/7-of-americans-dont-use-the-internet-who-are-they/

Perry, J. L., & Hondeghem, A. (2008). Building Theory and Empirical Evidence about Public Service Motivation. *International Public Management Journal*, *11*(1), 3–12.

Perry, J. L., & Wise, L. R. (1990). The Motivational Bases of Public Service. *Public Administration Review*, *50*(3), 367–373.

Prokop, C., & Tepe, M. (2020). Do Future Bureaucrats Punish More? The Effect of PSM and Studying Public Administration on Contributions and Punishment in a Public Goods Game. *International Public Management Journal*, *23*(1), 84–112.

Quattrone, G. A., & Tversky, A. (1988). Contrasting Rational and Psychological Analyses of Political Choice. *The American Political Science Review*, *82*(3), 719–736.

Results for America. (2019, September 3). *Achieving the Promise of the Evidence Act*. Evidence Act Resource Hub. https://results4america.org/evidence-act-resources/

Results for America. (2020a, July 12). *Healthy home environment assessments*. Economic Mobility Catalog. https://catalog.results4america.org/program/healthy-home-environment-assessments?issueArea=2228

Results for America. (2020b, July 12). *Lead paint abatement programs*. Economic Mobility Catalog. https://catalog.results4america.org/program/lead-paint-abatement-programs?issueArea=2228

Ritz, A., Neumann, O., & Vandenabeele, W. (2016). Motivation in the public sector. *The Routledge Handbook of Global Public Policy and Administration*, 346–359.

Rossi, P. H., & Nock, S. L. (1982). *Measuring Social Judgments: The Factorial Survey Approach*. SAGE Publications.

Ross, S. (1973). *The Economic Theory of Agency: The Principal's Problem*. https://www.semanticscholar.org/paper/fcc14e6e3ca7f9b7f6cbe09b4e05712905527e47

Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*, *55*(1), 68–78.

Ryan, R. M., & Deci, E. L. (2017). *Self-determination theory: Basic psychological needs in motivation, development, and wellness*. The Guilford Press.

Sanderson, I. (2002). Evaluation, Policy Learning and Evidence-Based Policy Making. *Public Administration*, *80*(1), 1–22.

Scott, J. T., & Bornstein, B. H. (2009). What's Fair in Foul Weather and Fair? Distributive Justice across Different Allocation Contexts and Goods. *The Journal of Politics*, *71*(3), 831–846.

Scott, J. T., Matland, R. E., Michelbach, P. A., & Bornstein, B. H. (2001). Just Deserts: An Experimental Study of Distributive Justice Norms. *American Journal of Political Science*, *45*(4), 749–767.

Shafir, E., & LeBoeuf, R. A. (2002). Rationality. *Annual Review of Psychology*, *53*, 491–517.

Shankar, M. (2014, July 30). *How Low-cost Randomized Controlled Trials Can Drive Effective Social Spending*. The White House, President Barack Obama: Blog. https://obamawhitehouse.archives.gov/blog/2014/07/30/how-low-cost-randomized-controlled-trials-can-drive-effective-social-spending

Sharot, T. (2011). The optimism bias. *Current Biology*, *21*(23), R941–R945.

Sharot, T., Korn, C. W., & Dolan, R. J. (2011). How unrealistic optimism is maintained in the face of reality. *Nature Neuroscience*, *14*(11), 1475–1479.

Shulock, N. (1999). The Paradox of Policy Analysis: If It Is Not Used, Why Do We Produce So Much of It? *Journal of Policy Analysis and Management*, *18*(2), 226–244.

Simon, H. A. (1955). A Behavioral Model of Rational Choice. *The Quarterly Journal of Economics*, *69*(1), 99–118.

Simon, H. A. (1990). Bounded Rationality. In J. Eatwell, M. Milgate, & P. Newman (Eds.), *Utility and Probability* (pp. 15–18). Palgrave Macmillan UK.

Simon, H. A. (1991). Bounded Rationality and Organizational Learning. *Organization Science*, *2*(1), 125–134.

Smith, A. (2005). *An Inquiry into the Nature and Causes of the Wealth of Nations* (J. Manis (Ed.)). The Electronic Classics Series.

Thaler, R. (1980). Toward a positive theory of consumer choice. *Journal of Economic Behavior & Organization*, *1*(1), 39–60.

Thissen, D., Steinberg, L., & Kuang, D. (2002). Quick and Easy Implementation of the Benjamini-Hochberg Procedure for Controlling the False Positive Rate in Multiple Comparisons. *Journal of Educational and Behavioral Statistics: A Quarterly Publication Sponsored by the American Educational Research Association and the American Statistical Association*, *27*(1), 77–83.

Torugsa, N. (ann), & Arundel, A. (2017). Rethinking the effect of risk aversion on the benefits of service innovations in public administration agencies. *Research Policy*, *46*(5), 900–910.

Tversky, A., & Kahneman, D. (1991). Loss aversion in riskless choice: A reference-dependent model. *The Quarterly Journal of Economics*, *106*(4), 1039–1061.

U.S. Bureau of Labor Statistics. (2021, July 30). *2021 Annual Averages - Employed persons by detailed industry, sex, race, and Hispanic or Latino ethnicity*. Labor Force Statistics from the Current Population Survey. https://www.bls.gov/cps/cpsaat18.htm

Vandenabeele, W. (2007). Toward a public administration theory of public service motivation. *Public Management Review*, *9*(4), 545–556.

Vandenabeele, W., Ritz, A., & Neumann, O. (2018). Public Service Motivation: State of the Art and Conceptual Cleanup. In E. Ongaro & S. Van Thiel (Eds.), *The Palgrave Handbook of Public Administration and Management in Europe* (pp. 261–278). Palgrave Macmillan UK.

Vis, B. (2011). Prospect theory and political decision making. *Political Studies Review*, *9*(3), 334–343.

von Neumann, J., & Morgenstern, O. (2007). *Theory of Games and Economic Behavior*. Princeton University Press.

Wallander, L. (2009). 25 years of factorial surveys in sociology: A review. *Social Science Research*, *38*(3), 505–520.

Walzer, M. (2008). *Spheres Of Justice: A Defense Of Pluralism And Equality*. Basic Books.

Wang, T.-M., van Witteloostuijn, A., & Heine, F. (2020). A Moral Theory of Public Service Motivation. *Frontiers in Psychology*, *11*, 517763.

Wang, X. H., & Yang, B. Z. (2001). Fixed and Sunk Costs Revisited. *The Journal of Economic Education*, *32*(2), 178–185.

Ward, M. K., & Meade, A. W. (2018). Applying social psychology to prevent careless responding during online surveys. *Applied Psychology = Psychologie Appliquee*, *67*(2), 231–263.

Weinstein, N. D. (1980). Unrealistic optimism about future life events. *Journal of Personality and Social Psychology*, *39*(5), 806–820.

White Junod, S. (n.d.). *FDA and clinical drug trials: A short history*. U.S. Food and Drug Administration. Retrieved May 28, 2022, from https://www.fda.gov/media/110437/download

Williamson, O. E. (1979). Transaction-cost economics: The governance of contractual relations. *The Journal of Law & Economics*, *22*(2), 233–261.

Williamson, O. E. (1981). The Economics of Organization: The Transaction Cost Approach. *The American Journal of Sociology*, *87*(3), 548–577.

Wright, B. E. (2001). Public-Sector Work Motivation: A Review of the Current Literature and a Revised

Conceptual Model. *Journal of Public Administration Research and Theory*, *11*(4), 559–586.

Wright, B. E., & Christensen, R. K. (2010). Public Service Motivation: A Test of the Job Attraction–Selection–Attrition Model. *International Public Management Journal*, *13*(2), 155–176.

Yang, K., & Banamah, A. (2014). Quota Sampling as an Alternative to Probability Sampling? An Experimental Study. *Sociological Research Online*, *19*(1), 56–66.

Zey, M. A. (2015). Rational Choice and Organization Theory. In J. D. Wright (Ed.), *International Encyclopedia of the Social & Behavioral Sciences (Second Edition)* (pp. 892–895). Elsevier.

Vita

EMILY CARDON

Education

Ph.D. Public Administration, Maxwell School, Syracuse University — Aug 2022
Fields of Specialization: Social Policy, Public Finance
Dissertation: *Public Administrator Aversion to Randomized Controlled Trials*

MPA, Maxwell School, Syracuse University — Jun 2013

B.A. International Studies, Boston College — May 2008
Phi Beta Kappa, *Magna cum laude*

Experience

The Behavioral Insights Team, Brooklyn, NY
Principal Advisor, Head of Research — Sep 2019 – present
Senior Research Advisor — Mar 2018 – Aug 2019
Advisor — Feb 2017 – Mar 2018

Center for Policy Research, Syracuse University
Graduate Assistant — Sep 2013 – Jan 2017

Campbell Public Affairs Institute, Syracuse University
Graduate Assistant — Sep 2012 – May 2013

Office of Congressman James A. Himes, Washington, DC
Legislative Assistant — Jan 2011 – Jun 2012
Legislative Correspondent — Aug 2009 – Dec 2010
Staff Assistant — Jan 2009 – Aug 2009

Peer-Reviewed Publications

Lopoo, L.M, **Cardon, E.B.,** Raissian, K.M. (2018). Health insurance and human capital: Evidence from the Affordable Care Act's dependent coverage mandate. *Journal of Health Politics, Policy, and Law,* 43(6): 917–939.

Working Papers

Cardon, E.B. & Lopoo, L.M. RCT Aversion Among Public Administrators: A Survey Experiment

Teaching Experience

Syracuse University (All Masters level)

PAIA Stata Training Group (Instructor)	2016, 2017
Quantitative Analysis: Program Evaluation (Teaching Assistant)	2016, 2017
Introduction to Statistics (Teaching Assistant)	2015, 2016
Public Budgeting (Teaching Assistant)	2014, 2015, 2016

Health Economics and Policy (Teaching Assistant) 2015

Honors and Awards

Larry D. Schroeder Award, Syracuse University 2015

Summer Research Award, Syracuse University 2015, 2016