Evaluation of OSHA’s SST Program Using a Randomized Controlled Trial Design and a Regression Discontinuity Design

Prepared for:
APPAM 39th Annual Fall Research Conference
November 2, 2017

Authors:
George Cave, Ph.D. (George.Cave@summitllc.us)
Ed Dieterle, Ed.D. (Ed.Dieterle@summitllc.us)
Balint Peto
Laura Hoesly
David Kretch
TABLE OF CONTENTS

Introduction .............................................................................................................................................1
   Two Designs .........................................................................................................................................1
Outline of the Rest of the Paper ...........................................................................................................2

1.0 Special Econometric and Operational Issues for OSHA SST Evaluation ........................................3
   1.1 Possible Regression to the Mean ....................................................................................................3
   1.2 Possible Dilution of the Treatment Contrast by Other OSHA Inspection Activities ..................3
   1.3 A “Two-Year Rule” Constrains Timing of Surprise Programmed Inspections ............................4
   1.4 Worksite Injury-Illness Data and SST Prioritization Thresholds for Enforcement Actions ..........4
   1.5 Attribution of Impact Estimates to Possible Mediators ...............................................................5
   1.6 Unrepresentativeness of OSHA SST Worksites of all Worksites ..............................................6
   1.7 Publication Bias ............................................................................................................................6

2.0 Technical Evaluation Design Planning ..........................................................................................7
   2.1 Counterfactuals and a Hierarchy of Evidence about Program Impacts ........................................7
      2.1.1 ITT Impacts versus TOT Impacts ...........................................................................................9
      2.1.2 ATE Impacts versus LATE Impacts .......................................................................................9
   2.2 Randomized Controlled Trials .....................................................................................................10
      2.2.1 Feasibility Issues ..................................................................................................................10
      2.2.2 Possibilities for Selection Bias .............................................................................................11
      2.2.3 Overall Credibility of Findings ............................................................................................11
      2.2.4 Dealing with Possible Regression to the Mean: A Special Concern for SST Evaluations ........11
      2.2.5 One-Way Analysis of Covariance for Calculating RCT Impacts ..........................................11
   2.3 Regression Discontinuity Designs .................................................................................................14
   2.4 Multiple Thresholds for Enforcement Actions and Multiple RDD Running Variables ............15

3.0 Tradeoffs between RCTs and MDI-equivalent RDDs ..................................................................16
   3.1 Ex Ante Practical Tradeoffs between Generic RCTs and Generic RDDs .....................................16
   3.2 Ex Ante Tradeoffs between the OSHA SST Evaluation RCT and the OSHA SST Evaluation RDD ...18

4.0 Ex Post Tradeoffs between the OSHA SST Evaluation RCT and the OSHA SST Evaluation RDD ....19

5.0 Lessons from DOL-CEO’s 2016 OSHA SST Evaluation, and Possible Implications for Future
   Evaluation Strategies ..........................................................................................................................22

References ..............................................................................................................................................24
INTRODUCTION

In this paper, we examine the strengths and limitations in evaluation design and credibility of findings when using Randomized Controlled Trial (RCT) and Regression Discontinuity Designs (RDD). We use the U.S. Department of Labor’s Occupational Safety and Health Administration (OSHA), Site-Specific Targeting (SST) program evaluation (Peto et al. 2016) as a case study for exploring design and credibility issues and tradeoffs, and for lessons that might inform future evaluations.¹ Thus, this paper emphasizes SST Evaluation methodological issues and insights rather than the numeric results of the evaluation.

The OSHA SST impact evaluation used both an RCT and an RDD to evaluate the same program. Unlike many other RCT/RDD studies, this was not a reanalysis, but work under contract and under deadline to answer a client’s research questions. Also, unlike other studies in this Association for Public Policy Analysis and Management session, this paper does not present a Within-Study Comparison: Although the RCT and the RDD used data from the same program, the data were three years apart, mainly for an unexpected reason beyond our control, which limits the closeness of such a comparison.

SST was a planned inspection program managed by OSHA from 2004-2014. SST carried out enforcement actions to improve workplace health and safety where injury and illness rates were high. Enforcement actions included:

1. High-Rate Letters (HRLs) sent to workplaces to warn managers about their high injury-illness rates, with names and addresses on the HRL list published on the DOL website, as possible leads or business intelligence for competitors, safety equipment salespersons, union organizers, and personal injury lawyers;² and
2. Programmed inspections, during which OSHA inspectors scrutinize worksites for compliance with OSHA safety and health rules, standards, and regulations. Inspections potentially can result in citations, penalties, fines, hazard abatement orders, and bad publicity.

The SST evaluation study was designed to measure whether 2011 SST Directive enforcement actions affected: (1) regulatory compliance, measured by the probability of being cited for a violation during a follow-up programmed inspection; and (2) health and safety, measured by follow-up injury and illness measures. The study measured both the direct impacts of receiving letters and inspections and the indirect impacts of being assigned by OSHA to these enforcement actions. Indirect impacts allow detecting potential deterrent effects that fear of inspections may have had on regulatory compliance and health and safety.

Two Designs

DOL CEO’s initial contractor applied an RCT design to assess causal impacts of enforcement actions on regulatory compliance and health and safety. Implementation challenges, including cancellation of the

---

¹ Summit Consulting, LLC (Summit), under contract to DOL’s Chief Evaluation Office (CEO), completed OSHA’s SST program evaluation in 2016. IMPAQ International designed the original RCT impact evaluation in 2010. Summit first received evaluation materials from IMPAQ in late 2013 and submitted a Final Report to DOL-CEO in May 2016. DOL makes the report available online: https://www.dol.gov/asp/evaluation/completed-studies/SST_Evaluation_Final_Report.pdf.

² An example of a high-rate letter can be found in Appendix A, PDF page 64 at https://www.dol.gov/asp/evaluation/completed-studies/SST_Evaluation_Final_Report.pdf.
annual OSHA Data Initiative (ODI) workplace survey to measure follow-up outcomes in 2013-2014, diminished the quality of the data and the ex post statistical power of the RCT, limiting its ability to detect impacts of policy-relevant size.

Existing injury-illness-incidence inspection thresholds, final ODI survey data, available inspections data, and the study team’s matching algorithm for merging relevant OSHA datasets allowed us to implement an RDD evaluating the 2008 SST Directive with sufficient power at an acceptable additional cost. Worksites above relevant thresholds in 2006 injuries and illnesses from the 2007 ODI became treatment sites; those under the thresholds became comparison sites. This design allowed attributing observed differences in outcomes between groups on either side of the threshold to (1) OSHA’s inclusion of worksites on treatment site lists and, under some econometric assumptions, to (2) actual receipt of a specific treatment.

Although neither the RCT nor the RDD found statistically significant impacts, the project methodologically advanced how program evaluators can apply RDD to existing administrative data to measure impacts. Using a secondary study design reinforced the weak but underpowered RCT impact findings, helped build knowledge about appropriate applications of quasi-experimental designs, and revealed some strengths and some weaknesses of both approaches.

Outline of the Rest of the Paper

The next section of the paper, Special Econometric and Operational Issues for OSHA SST Evaluation, describes special econometric and operational aspects of OSHA SST that any of the effective evaluation design must accommodate in establishing valid counterfactuals and developing theories of change about how, when, and where the OSHA SST program might cause differences in post-program outcomes between program worksites and comparison group worksites. In particular, regression to the mean is always an evaluation danger when lagged values of the ultimate evaluation outcome are used to prioritize elements of a population for program treatment. Also, SST programmed inspections were not the only inspections OSHA carried out during evaluation time periods. SST control, comparison, and treatment group members could be inspected under National Emphasis and other OSHA programmed inspection programs, and unprogrammed inspections could occur in response to a workplace incident, complaint, or referral. For the SST Evaluation RDD, multiple baseline injury and illness thresholds could be addressed in several different ways (Wong, Steiner, and Cook 2013). We chose the centering approach.

Using an ANCOVA framework (e.g., Huitema 2011), the second section of the paper, Technical Evaluation Design Planning, describes the fundamental evaluation problem of selection bias and the statistical properties evaluation methods must possess to avoid selection bias convincingly. It also describes two kinds of impact estimates either random assignment or RDD might yield: intention-to-treat (ITT) impacts and treatment-on-treated (TOT) impacts. Another important dichotomy is average treatment effects (ATEs) that compare whole-sample group means, and local average treatment effects (LATEs), which compare group means only for sample points that are “close” to each other. Minimum detectable impact formulas are derived for use in assessing comparative statistical power of alternative RCT and less efficient RDD impact methods.
The third section, *Tradeoffs between RCTs and MDI-equivalent RDDs*, describes *ex ante* tradeoffs between generic RCTs and generic RDDs, along with *ex ante* tradeoffs between RCTs and RDDs in the specific context of SST. For example, RCTs must be run prospectively, while RDDs can use archived administrative and survey data available immediately, allowing funders to get evaluation results more expeditiously, with negligible risk that data might prove unavailable, and requiring little change in ordinary business processes at an agency.

In *Ex Post Tradeoffs between the OSHA SST Evaluation RCT and the OSHA SST Evaluation RDD*, the fourth section of the paper, we present the same comparison for SST from an ex post perspective, including a summary of ex post tradeoffs between the two specific SST evaluation designs whose results are presented in the SST Evaluation Final Report.

In the concluding section of the paper, *Lessons from DOL-CEO’s 2016 OSHA SST Evaluation, and Possible Implications for Future Evaluation Strategies*, we summarize important lessons from the OSHA SST Evaluation for future evaluation designers that consider applying RCT, RDD, or both designs simultaneously.

### 1.0 Special Econometric and Operational Issues for OSHA SST Evaluation

We identify six salient econometric and operational issues associated with implementation of RCT and RDD that are especially relevant for evaluating SST. The most important issue is possible regression to the mean, which results when the running variable is a lagged outcome.

Another key issue is possible dilution of the treatment contrast by other OSHA inspection activities. An issue for any RDD work arises from worksite injury-illness data and SST prioritization thresholds for enforcement actions, resulting in multiple assignment frontiers. We used Wong, Steiner, and Cook’s (2013) centering approach to pool different enforcement streams and maximize sample sizes, which we describe in detail later in the paper. An ethical issue is possible future contribution to *publication bias* had we abandoned RCT follow-up and analysis entirely when its planned follow-up survey was canceled. Below, we unpack each issue.

#### 1.1 Possible Regression to the Mean

Ruser (1995) and others have pointed out that simple before-after comparisons of injuries and illnesses at worksites could easily yield spurious estimates of the impacts of OSHA inspections. A worksite’s environmental hazards may remain the same from year to year, but random chance sometimes might put a worksite over an enforcement threshold. A return to a worksite’s long-term expected injury-illness rate might be measured as a spurious impact. Balance at baseline on the running variable is crucial for internal validity of measured impacts.

#### 1.2 Possible Dilution of the Treatment Contrast by Other OSHA Inspection Activities

Non-SST inspections might dilute the treatment-counterfactual contrast. OSHA conducts two classes of inspections of worksites to enforce compliance with OSHA requirements: unprogrammed and programmed inspections. Unprogrammed inspections are reactive, quick responses to complaints, referrals, and incidents, while programmed inspections are proactive measures aimed at specific
industries or worksites that have experienced high rates of injuries or illnesses in a prior year.\textsuperscript{3} In FY2014, OSHA conducted 19,222 programmed inspections and 16,941 unprogrammed inspections, or 36,163 inspections in all. If random assignment were to stop control group worksites’ access to SST Evaluation inspections completely, they still might be inspected under another surprise programmed inspection initiative or in response to complaints, referrals, and incidents. Such inspections among control group sites would move the RCT treatment contrast away from 100 percent to zero percent inspections toward something weaker.

1.3 A “Two-Year Rule” Constrains Timing of Surprise Programmed Inspections

Since OSHA does not have the resources to inspect even the least safe workplaces even as infrequently as once a year, programmed inspections use the element of surprise to make fear of sudden inspection a tool for greater influence on workplace managers’ safety cultures.\textsuperscript{4} But litigious employers have gotten courts to constrain OSHA’s ability to carry out surprise inspections. OSHA must show probable cause to believe there are unsafe conditions in a workplace to get a search warrant from an administrative law judge. Once a workplace has undergone a programmed inspection, OSHA will not schedule a programmed reinspection for at least two years, to allow probable cause to build up again.

1.4 Worksite Injury-Illness Data and SST Prioritization Thresholds for Enforcement Actions

From 1996 to 2011, OSHA fielded an annual survey, the ODI (OSHA Data Initiative), where employers self-reported a measure of the incidence and duration of nonfatal injuries and illnesses that caused their employees to lose days of work or be restricted or transferred at work during the past year.\textsuperscript{5} The 2011 ODI, for example, gathered information about worksite injuries and illnesses during 2010.

Until 2014, SST was one of OSHA’s largest programmed inspection vehicles. SST focused enforcement actions on worksites with highest measured injury or illness in an industry during a prior year, as measured by the ODI.\textsuperscript{6} OSHA set thresholds to determine which worksites would be subject to each SST action. Thresholds were set using standardized measures that rose with both higher incidence and longer duration of injury and illness, and fell with more employees and more overtime at a worksite. These measures are termed industry-specific DART (i.e., days away, restricted, or transferred) and DAFWII (i.e., days away from work due to injury and illness) per 100 employees working 40 hours per week 50 weeks a year.

SST inspections were comprehensive examinations of all potentially hazardous areas of a worksite. Such inspections could take as little as three hours or up to four weeks, depending on the size of the worksite and the complexity of the inspection. With millions of covered workplaces to regulate, but resources for fewer than 20,000 surprise programmed inspections a year, OSHA has been forced to prioritize its use of scarce resources, saving the costliest enforcement actions for the highest-priority worksites. To extend the reach of the SST program beyond actual inspections, OSHA used high-rate letters to inform

\textsuperscript{3} See https://www.osha.gov/dep/2014_enforcement_summary.html
\textsuperscript{5} A copy of the final 2011 ODI survey instrument is in Appendix A at https://www.dol.gov/asp/evaluation/completed-studies/SST_Evaluation_Final_Report.pdf.
\textsuperscript{6} For example, the 2011 SST Program Directive prioritized inspections from fall 2011 through the end of 2012. Priorities were based on results from the 2010 ODI, which ascertained self-reported injuries and illnesses during 2009.
employers that they had high rates of injury or illness compared to their industry competitors and to suggest ways to improve the safety of their worksites. Because such letters cost very little, OSHA can afford to send them to somewhat lower-priority worksites as well as to higher-priority worksites. 2011 SST Directive thresholds for enforcement actions were levels of injury and illness from the 2010 ODI survey, which gathered 2009 employer self-reports.

Based on these thresholds, OSHA classified worksites into four groups as follows:

- Worksites with injury/illness rates at or above a first tier of injury/illness rate thresholds were made part of the Primary List of worksites to be subject to SST inspection.
- Worksites with rates above a second tier of thresholds but below Primary List thresholds joined the Secondary List and were to be inspected after all the Primary List worksites were inspected, given resource availability.
- Worksites above a third and lowest threshold, including all worksites in the Primary and Secondary Lists, were sent high-rate letters.
- Worksites below the lowest threshold did not receive a high-rate letter and were not on the SST inspection list but still potentially could receive an inspection for another reason.

Table 1 presents OSHA SST Directive enforcement groups and actions by prioritization scheme.

<table>
<thead>
<tr>
<th>Group</th>
<th>Enforcement Action</th>
<th>Probability of Inspection</th>
</tr>
</thead>
<tbody>
<tr>
<td>#1 Primary List</td>
<td>Receives Letter? Yes</td>
<td>On SST Inspection List? Yes</td>
</tr>
<tr>
<td>#2 Secondary List</td>
<td>Receives Letter? Yes</td>
<td>On SST Inspection List? Yes</td>
</tr>
<tr>
<td>#3 Letter List</td>
<td>Receives Letter? Yes</td>
<td>On SST Inspection List? No</td>
</tr>
<tr>
<td>#4 No letter group</td>
<td>Receives Letter? No</td>
<td>On SST Inspection List? No</td>
</tr>
</tbody>
</table>

1.5 Attribution of Impact Estimates to Possible Mediators

As Weiss (1998) pointed out repeatedly in various contexts, the most rigorous and credible evaluation methods provide no rigorous answers to many questions of likely high interest to stakeholders. An example of such a question might be, “Do programmed inspections that last longer have larger impacts on follow-up injuries and illnesses? Administrative data on inspections at treatment-group sites easily could allow calculation of the median duration of an inspection, using zero as the duration at worksites where inspection did not happen. A binary endogenous subgroup indicator could be constructed for all members of the SST RCT Treatment Group. But it would be very difficult to construct a credible indicator of the same construct for treatment group worksite counterparts in the control group. Pearl (2012) surveys and adds to a mediator attribution approach that was proposed 30 years ago. Building on Rubin (2005), Page (2012), and others have proposed promising principal stratification approaches to
mediation that have become very controversial. One of the generic weaknesses of most current empirical evaluation research may be that it generally ignores the crucial issue of precisely what led to impacts, if interventions have multiple elements.

1.6 Unrepresentativeness of OSHA SST Worksites of all Worksites
The OSHA SST evaluation excludes “State Plan” states and small worksites. OSHA’s 2011 SST Directive was applicable only to 29 States and the District of Columbia. The OSHA SST Evaluation excluded all worksites in 21 State Plan states and two State Plan territories, which OSHA has authorized to run their own workplace safety regulation programs.

Three of these State Plan states—Indiana, Iowa, and Tennessee—participated in both SST 2011 and the 2010 ODI Survey, but were excluded from the OSHA SST Evaluation RCT and RDD because they were not obligated to follow the 2011 SST Directive.

Some State Plan jurisdictions may have weaker regulations and enforcement actions than federal OSHA jurisdictions, and some, such as California, may have much stronger regulations and enforcement actions (Levine, Toffel, and Johnson 2012). Thus, OSHA SST Evaluation results may understate or overstate ATE or LATE results from an evaluation that included all US jurisdictions.

SST Directives also exclude worksites with fewer than 40 employees. Some economists, perhaps influenced by the pioneering work of Oi (1962) on fixed costs of production, have argued that OSHA regulations affect safety decisions of small employers much more than safety decisions of large employers. To the extent that investment in safer equipment is a fixed cost, it increases the average cost of production more for smaller employers than for larger employers. Larger workplaces that have been in existence for many years before 2011 already might have sunk investment in optimal levels of safety equipment, and would not have changed their safety behavior much in response to enforcement actions under the 2011 SST Directive. Thus, OSHA SST Evaluation results may understate ATE or LATE results from an evaluation that included all US jurisdictions.

1.7 Publication Bias
Because large multi-year RCTs, particularly RCTs for evaluation of human service programs, are risky and expensive, they sometimes produce unexpected results, or produce inconclusive results because of unexpected operational problems. Stakeholders invest in RCTs with the hypothesis that the treatment will reveal positive and significant impacts. So, it is wise to always carry random-assignment evaluations past the baseline sample intake stage through follow-up data collection and a final report, even if it becomes apparent that ex-post follow-up sample sizes will be inadequate for conclusive findings.

The OSHA SST Evaluation RCT was particularly vulnerable to operational surprises because it was the first attempt to mount a large RCT to address the impact of OSHA inspections, a nationwide complex activity with many centers of management power and authority. Especially if this first rigorous evaluation attempt is not the last, it could have been a disservice to evaluation research not to continue

---

with RCT data gathering, analysis, and reporting even after it became apparent impact results would be inconclusive.

Suppression of unfavorable or inconclusive RCT results is widely considered unethical, and there is a large and longstanding publication bias literature on the phenomenon. For more information on publication bias, see Begg and Berlin (1988) for a thorough survey of the earlier literature. For more recent discussions on publication bias, see Dwan et al. (2013) and Franco, Malhotra, and Simonovits (2014).

2.0 TECHNICAL EVALUATION DESIGN PLANNING

There are many methodological approaches to evaluation. Why was an RCT chosen as the planned approach to evaluate SST? And, why was an RDD added four years later? The impact of OSHA planned inspections has a longstanding non-experimental literature, and only the most rigorous evaluation methods available could help resolve remaining uncertainty about the effects of OSHA enforcement actions.

2.1 Counterfactuals and a Hierarchy of Evidence about Program Impacts

Whenever a social program is evaluated, a number of key post-program outcomes are observed for the program group: for example, mandatorily reportable deaths, injuries, or illnesses.

The question that must be answered to evaluate the program is: what would these outcomes have been had these sample members not experienced the program treatment? Since it is not possible to observe a second set of outcomes for program group sites, a counterfactual set of outcomes must be obtained in some other way, for comparison with observed outcomes for the program group. The credibility of an evaluation depends on the quality of its comparison outcomes; some counterfactuals are more credible than others. This paper considers two different ways of creating counterfactual outcomes for program evaluation, and assesses their overall believability, advantages, disadvantages, and tradeoffs in the context of evaluating SST.

Many potential counterfactuals have severe drawbacks. Pre-program outcomes for the program group remain static even though outcomes might have improved or worsened over time even without the program. In particular, regression to the mean after a prior year of unusually good or bad luck with random fluctuations in health and safety outcomes might present the appearance of negative or positive spurious impacts. This is a fundamental problem for evaluating any program that prioritizes a finite collection of worksites for inspection based on prior-year outcomes.

Thus, most program evaluations get counterfactual outcomes for the same period from comparison groups who did not experience the program. For various evaluation designs, such comparison groups have included those who applied for the program but never entered it (Rohe 1995), those who applied for a number of different programs at the same time and were offered another program instead.

---

Evaluation of OSHA’s SST Program Using a Randomized Controlled Trial Design and a Regression Discontinuity Design

(Popkin, Rosenbaum, and Meaden 1993), and, from non-program locations (Slavin 1992) or anonymous national databases (LaLonde 1986), people or other units of analysis whose characteristics are statistically most similar to those of program group members. In most cases, it is quite likely that each of these kinds of comparison groups will not generate appropriate counterfactual outcomes for evaluation purposes, because these comparison groups will differ systematically from the program groups they were intended to match.

Such designs may confound program impacts with background characteristics, such as lagged workplace injury and illness measures. Such systematic biases in evaluation results caused by pre-program differences between program and comparison groups are called selection biases because they result from program selection by individuals or participant selection—often *creaming* or *reverse creaming*—by program intake staff. A reliable way to eliminate such biases is for neutral researchers to gain control of prospective access to the program and use a statistically random process to try to control who may experience the program and who becomes part of a control group without program access.

As shown in the next section, *Randomized Controlled Trials*, when many sufficient program and control group members are randomly assigned, and when there is substantial compliance with the research protocol (i.e., the program group members experience the program and the control group members do not), evaluation results generally will be consistent. We may expect that on average, the control group outcomes will be precisely what we would see if we could go back in time and take away the program from the program group members. The most credible ATE results usually stem from evaluations which randomly assigned individuals to program or control group status and had high take-up and low contamination.

Random assignment of individual units of analysis usually is at the top of the hierarchy of evaluation designs, because it most thoroughly removes possible selection biases from outcome comparisons between those chosen for treatment and those chosen for comparison (e.g., Cave and Quint 1990, Finn and Achilles 1990). However, it is often quite difficult to implement individual-level random assignment for program evaluations. Because it must be carried out prospectively over often several years, researchers must have a substantial amount of cooperation from program staff, enough cooperation to put into researchers’ hands decisions about who may enter the program and who may not, and to monitor and ensure compliance with hundreds or thousands of these random decisions. In the empirical literature on regulating workplace hazards, sometimes the unit of analysis is the individual worker, using data from Workers Compensation agencies, while the treatment unit is the workplace, in clustered designs (Haviland et al. 2012).^9

Sometimes, randomization of any kind is infeasible, but there may be grounds for interpreting what unfolds during the normal allocation of resources to sample members as a natural experiment (Popkin, Rosenbaum, and Meaden 1993); for building an evaluation design on systematic nonrandom allocation of program resources (Berk et al. 2010); or for using dropouts from program recruitment as a

---

^9 For more information on this topic, see DOL’s Evidence on the Effects of OSHA Activities [https://clear.dol.gov/synthesis-report/evidence-effects-osha-activities](https://clear.dol.gov/synthesis-report/evidence-effects-osha-activities)
comparison group (Rohe 1995). When other designs fail, purely statistical procedures must be used to construct comparison groups of people comparable to those in a program to be evaluated (Heckman and Robb 1986), or to project from a pre-program time series what would have happened without the program (McCrary 2008).

### 2.1.1 ITT Impacts versus TOT Impacts

Assigning plants or lab animals to treatments is essentially the same as ensuring that only the plants or animals who are assigned a treatment actually get that treatment, and that none of them assigned to a control group get a treatment. This is not so for human beings. One of the reasons RCTs may not be generalizable outside the sample of individuals recruited for an experiment is that measures often are taken to ensure high take-up rates, or proportions assigned to a treatment who actually get that treatment. But such measures may result in equally highly-motivated control group members’ seeking out similar services outside the control of the RCT, and actually getting them, if they are available.

If the control services are at least somewhat effective, then impact estimates may be diluted by control group contamination. The best way to screen recruits for high take-up rates yet not risk contaminating the control group is to recruit sample members who have few alternative programs available. For example, Career Beginnings, a program to encourage college going among the disadvantaged, recruited mainly B-/C+ students, rather than A+ students who often had many alternatives outside the Career Beginnings program (Cave and Quint 1990). The calculations just described are known as intent to treat (ITT) average treatment effect (ATE) impacts, or ITT-ATE.

The dilution problem has some potential cures but the cures may be worse than the problem. Under some strong econometric assumptions, the instrumental variable procedure due to Wald (1940) can be used to calculate TOT impacts on ultimate outcomes of interest, such as follow-up illnesses and injuries. Stock and Watson (2015) describe careful use of a two-step procedure we followed to calculate TOT impacts. For an RCT, the first step is an ITT impact regression for the instrument, here randomly assigned placement on a 2011 SST Directive inspection, letter list, or both. To avoid a weak instrument problem, which could result in spurious TOT impact estimates, Stock and Watson recommend an instrument relevance rule that the F-statistic for this first-stage regression reach a value of at least ten. The second stage is an ANCOVA regression, which uses the predicted value of the instrument as the key right-hand-side variable. For RDD, TOT impact calculations are very similar adaptations of RDD ANCOVA regression equations.

For this reason, the Food and Drug Administration (FDA) regulations prohibit drug, device, or procedure trial results from going beyond the ITT framework. Arguably, ITT results may actually be more policy-relevant than TOT results, since outside interventionists never have complete power to force human beings to undergo any treatment or program, even incarcerated human beings.

### 2.1.2 ATE Impacts versus LATE Impacts

Average Treatment Effects (ATEs) are global comparisons of mean post-program outcomes for all members of baseline treatment groups with mean outcomes for all members of their corresponding control groups. If it can be shown that two groups were balanced at baseline on potential confounders, such as lagged outcomes, then many skeptics might be convinced that ATE comparisons represent treatment effects rather than biases from imbalances in confounders. However, ATEs hide a great deal
of potential heterogeneity in impacts. There can be big and small losers as well as big and small winners from any given program. In cost-benefit analysis, ATEs may be quite useful for high-stakes zero-based budgeting, but not so useful for considering marginal increases or decreases in program scale. ATEs may be most useful for high stakes decisions to terminate, continue, or double program activities, but not so useful for marginal decisions about smaller increases or decreases in program resources.

In contrast, Local Average Treatment Effects (LATEs) are comparisons of mean post-program outcomes only for treatment group members and control group members whose values of confounders are closely matched. LATE impacts may be most useful for making marginal decisions, such as whether or not to add additional program staff, expand physical facilities, loosen entry criteria, tighten entry criteria, lengthen minimum lengths of stay, etc.

2.2 Randomized Controlled Trials
This sub-section explains why, when it is feasible, and ATE is the parameter of interest, random assignment might be considered preferable to other methods; how it removes possible biases in estimates of program effects; and the statistical properties of the most common econometric models used in random-assignment program evaluations.

2.2.1 Feasibility Issues
For any evaluation to be feasible, it must be possible to separate research sample members into groups who are reasonably sure to have very different levels of access to the program being evaluated or programs quite similar to it. The need to recruit more program applicants than will be admitted, in order to create a control group, is much less problematic when the program is oversubscribed, and random selection may be seen as a fair way to allocate scarce program slots. It must be reasonable to expect or possible to ensure substantial compliance with the research protocol that program group members experience the program and control group members do not.

For the OSHA SST RCT, recruiting a control group may have entailed lowering 2011 SST Directive thresholds slightly, to recruit Primary Inspection List controls from what otherwise may have been marginal Secondary List worksites, and to recruit Secondary List controls from what otherwise may have been Letter-Only worksites. Less efficient two-to-one splits between RCT treatment group sites and RCT control group sites may have resulted from attempts to minimize such disruption of usual procedures.

Table 2 presents the 2011 SST directive RCT randomized groups and sample sizes with the baseline samples that were recruited for the OSHA SST Evaluation RCT.

### Table 2. 2011 SST Directive RCT Randomized Groups and Baseline Sample Sizes

<table>
<thead>
<tr>
<th>Study</th>
<th>Study and Randomized Group</th>
<th>Programmed Inspection Date</th>
<th>Follow-Up Inspection Date</th>
<th>N at baseline</th>
<th>Number and % of sites treated</th>
</tr>
</thead>
<tbody>
<tr>
<td>Letter</td>
<td>Treatment</td>
<td>2011</td>
<td>2013</td>
<td>840</td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td>Control</td>
<td>2011</td>
<td>2013</td>
<td>415</td>
<td>N/A</td>
</tr>
<tr>
<td>Letter Plus Inspection</td>
<td>Treatment</td>
<td>2011</td>
<td>2014</td>
<td>840</td>
<td>616 (73.3%)</td>
</tr>
<tr>
<td></td>
<td>Control</td>
<td>2011</td>
<td>2014</td>
<td>425</td>
<td>29 (6.8%)</td>
</tr>
</tbody>
</table>
2.2.2 Possibilities for Selection Bias
Random assignment is often preferred for evaluation purposes because it is widely considered quite likely to remove systematic differences between program and control groups in both observed and unobserved factors correlated with key outcomes. For example, if the objective of a workplace safety regulation program is fewer injuries and illnesses at follow-up, randomly-assigned worksites are likely to be matched evenly on such factors as lagged outcomes.

Combined with the knowledge that ex ante the mechanism for assignment was random, an ex post demonstration of overwhelming evenness of distributions of pre-program characteristics for the two research groups tends to support the hypothesis that there is as well a close match on unobserved pre-program factors such as safety culture. The larger the sample the more convincing will be a demonstration of the absence of statistically significant systematic baseline differences between randomly assigned program and control groups. Thus, random assignment of enough sample members usually makes selection bias rather unlikely for this kind of design.

2.2.3 Overall Credibility of Findings
RCT ATE impact findings have their maximum degree of credibility when the following can be shown:

- **ex ante**, true randomization was used to assign sample members to experimental or control groups;
- **ex post**, the ability of measured baseline characteristics to explain assignment to the experimental group is nil;
- a sample of sufficient size was recruited at baseline; and
- no selection bias was introduced by poor data-gathering procedures (e.g., differential attrition at follow-up).

To the extent that these conditions are not met, the credibility of findings is reduced.

2.2.4 Dealing with Possible Regression to the Mean: A Special Concern for SST Evaluations
In evaluating SST, and other workplace safety programs whose priorities stem from prior safety outcome rankings, regression to the mean has been an alternative explanation for several nonexperimental evaluation results (Ruser 1995). Thus, it would be essential to have good measures of past health and safety outcomes for worksites in an SST RCT, and it would be important to be able to show treatment-control group balance at baseline on such measures. Thus, demonstrating baseline balance on the SST RCT sample on 2009 DART, as well as an F-statistic far from statistical significance for the joint hypothesis that all covariate coefficients are zero, would be important.

2.2.5 One-Way Analysis of Covariance for Calculating RCT Impacts
To measure the effect of assignment to a program on one continuous outcome (e.g., DART or DAFWII) with a completely randomized two group experiment of the kind just described, classical optimal design theory leads to a kind of linear modeling known as analysis of covariance or ANCOVA (Huitema 2011). For the program’s target population as a whole, equation 1 presents the basic ANCOVA impact model.
Equation 1

\[ Y_i = \beta_0 + \delta S_i + \sum_k \gamma_k z_{ki} + \epsilon_i \]

Where:
- \( Y_i \) is the outcome variable of interest (DART or DAFWII sometime after being placed on 2011 SST Directive enforcement action lists);
- \( S_i \) is a dichotomous indicator variable of the ith sample member’s random assignment to the treatment group (\( S_i = 1 \)) or the control group (\( S_i = 0 \)); and
- \( z_{ki} \) is baseline characteristic k for the ith worksite (e.g., number of employees, hours per employee, OSHA region, OSHA area office, industry).

The impact parameter of interest is \( \delta \), the coefficient of the treatment indicator variable. The value of \( \delta \) is interpreted as the ATE, that is, the average change in DART or DAFWII in a follow-up period after being assigned to an enforcement action list.

In several ways, this model improves upon a simple comparison of means for the program group and the control group. First, it corrects the estimate of the program impact for slight random imbalances in pre-program characteristics that may be strongly correlated with the outcome (e.g., pre-program illness and injury measures). Second, it is more economical in the sample size required to find a true program impact of a given size.

2.2.5.1 An ITT ANCOVA Sample Size Formula

To measure the effect of assignment to a program on one continuous outcome with a completely randomized two-group experiment, let

- \( \alpha \) = significance level (10%, two-tailed);
- \( \beta \) = statistical power (1-80%);
- \( c \) = the fraction of the sample which belongs to the control group;
- \( z(x) \) = the inverse of the cumulative standardized normal distribution function.
- \( Y_i \) = the continuous outcome variable;
- \( Z_i \) = a k-vector of covariates measured just before assignment to treatment or control; and
- \( S_i \) = a dummy variable which is zero for those assigned to control status and unity for those assigned to treatment.

In the fitted regression equation \( \hat{Y}_i = b_0 + S_i d + Z_i \gamma_{1:k} \), which explains fraction \( \hat{R}^2 \) of sample outcome variance, the second coefficient is interpreted as the sample impact of assignment to the treatment. Its expected value is the population effect of treatment, \( \delta_0 \).

Equation 2 presents an expression for its variance, derived in Cave (1987), which may be manipulated as outlined there to yield the sample size formula.

Equation 2

\[ n \geq \frac{4(z(1 - \alpha) + z(1 - \beta))^2 (1 - \hat{R}^2) \text{Var}(y)}{\delta_0^2 (1 - R^2_{\hat{S}_z})} \]

\( n \) is a positive integer for the total number of usable data points required for analysis, \( \text{Var}(y) \) is the sample variance of the outcome or average squared distance of each sample member from the overall mean outcome, and \( R^2_{\hat{S}_z} \) is the proportion of the variation in \( S_i \) explained by a regression of \( S_i \) on \( Z_i \) and a
constant. $R^2_{sz}$ has expected value zero if assignment to treatment truly is random; the multiplicative factor $(1/(1-R^2_{sz}))$ is a sample-size-increasing randomization design effect analogous to a survey design effect.\(^{10}\)

\(a\) is the required significance level, or desired probability of a false positive finding or Type I error, ordinarily assumed to be 10 percent on a two-tailed test; \(\beta\) is the probability of a Type II error or false negative finding, equal to unity less the statistical power. Ordinarily, following Cohen (1988), required power is taken to be 80 percent. \(z(x)\) is the inverse of the cumulative standardized normal distribution function (thus $z(1-0.05) = 1.64485$ and $z(1-0.20) = 0.84162$ are the usual values for the constants in Equation 2.

The multiplicative factor involving \(c\) is a sample split inflation factor which takes the value unity when \(c=0.5\). Table 2 shows that in the Letter Plus Inspection list, \(c=0.336\), so that $(1/(4c(1-c))) = 1.121$. Thus, reducing the control fraction from the most efficient 50 percent to about one-third required a 12.1 percent increase in OSHA SST RCT sample size to find any given effect size with given significance at given power.

The factor $(1 – R^2)$ in the numerator is a variance deflation factor reflecting the degree of success for baseline covariates in soaking up outcome variance.

Multiplying inequality in Equation 2 by $\frac{\delta_0^2}{n}$, and taking square roots of both sides, turns it into MDI formula. Whatever increased required sample size for the previous formula now decreases MDI, as shown in Equation 3.

**Equation 3**

$$\delta_0 \geq \frac{4 \left( z(1-a) + z(1-\beta) \right)^2 \left( 1 – R^2 \right) Var(y) }{n \left( 1 – R^2_{sz} \right) 1 - c (1-c) }$$

Random assignment ensures that the regression estimate for the program impact will be consistent (in other words, that its expected value as the sample size increases will be the true value) by making the error term independent of the other variables on the right-hand side of Equation 1.

- This expression shows that the ANCOVA approach reduces the required sample size (below what would be needed for a simple comparison of means) by a fraction, which is the gain in R-squared from adding the \(k\) baseline covariates. The better the list of covariates (in terms of how much of the variation in the outcome they explain), the fewer sample points are needed for the evaluation.
- Since the true effect of treatment appears in the denominator of (2) raised to the second power, the expression shows that if the size of the effect to be found is cut in half, the size of the sample required to find it increases fourfold.
- Because the \(z()\) function slopes upward, pushing the significance threshold below 0.10 or demanding greater statistical power than 80 percent increases the required sample size.

\(^{10}\) To ensure the internal validity of inferences about effects, it is important to test this hypothesis for every sample and subsample of complete data used in the analysis of a social experiment.
To use this ANCOVA sample size formula, estimates of the true effect, \( \text{Var}(y) \), and \( R^2 \) must be obtained or assumed beforehand. Sample variances and proportions of variance explained can come from prior studies of similar populations and treatments. The estimate of the true effect can come from prior studies, from the cost of the treatment, or from some notion of what size effect would be policy-relevant. To ensure an adequate chance of finding a statistically significant impact in an evaluation sample, randomization is not enough; sufficient replication of the randomized experiment is required as well, and expression (2) provides important guidelines about how much replication is enough.

### 2.3 Regression Discontinuity Designs

Shadish, Cook, and Campbell (2001) explain very clearly how, under some strong assumptions, the “sharp” regression discontinuity design (RDD) mimics the RCT:

Regression Discontinuity as a Complete Model of the Selection Process... if the selection process could be completely known and perfectly measured, then one could adjust for differences in selection to obtain an unbiased estimate of treatment effect. In theory, these conditions are met in both RD and the randomized experiment, and so both designs can be viewed as special (successful) cases of selection bias modeling... In both cases, the assignment mechanism can be perfectly measured and implemented—that is, the researcher records correctly whether the coin came up heads or tails or whether a person’s score is above or below the cutoff... That is the key that allows simple statistics such as analysis of covariance (ANCOVA) to yield unbiased effect estimates in RD (Overall & Woodward, 1977). In other quasi-experiments, selection into conditions is neither fully known nor perfectly measured. (p. 224)

Operationalizing this insight about sharp RDDs with a statistical model requires an additional assumption if all potential program entrants are to have their post-program outcomes measured correctly. Shadish, Cook, and Campbell (2001) go on:

It is important to know the overall functional form that relates the assignment and outcome variables (i.e., whether it is linear, curvilinear, cyclical, etc.). A polynomial model may be appropriate for describing that form (Trochim 1984), or some other transformation may be used either for the assignment or posttest variable (e.g., a log transformation). Such functions can adequately describe most relationships. But if the functional form is misspecified in the analysis, treatment effects will be estimated with bias. (p. 218)

With uncertainty about the precise functional form, slightly more complex RDD approaches still may yield consistent estimates of LATE impacts.

Near a strictly enforced cutpoint on a running variable used as the sole criterion for program entry, slight random fluctuations in the value of the running variable could move a recruit just above or just below the cutpoint. Thus, a sharp RDD is almost the same as random assignment within a certain bandwidth of the threshold for program entry eligibility.
2.3.1 “Fuzzy” RDDs

Sometimes cutpoints for program entry are used inconsistently. The cutpoint may not provide strict classification of a baseline sample into program eligibles and non-eligibles, but rather may merely change probabilities of program admission.

When there has been such loose operational control over program access, under alternative sets of econometric assumptions (Rosenbaum and Rubin 1983, Trochim 1984), it still may be possible to calculate consistent program impacts using a two-step procedure or two-equation model. The first step or equation entails modeling program admission so well that there are wide ranges of accurate predictions, no omitted predictors, and no correlations between predictors and residual variation in program access. The second step entails calculating the regression coefficient of predicted program access in explaining outcomes. Such a coefficient is a kind of TOT impact, subject to the same extra scrutiny as other TOT impacts.

Advantages of the regression-discontinuity approach include the following considerations:

- Both baseline and follow-up measures may be available immediately in archived data. Unlike an RCT, which may have to unfold prospectively over several years, an RDD can be performed retroactively.
- LATE impact estimates are consistent and reliable with adequate samples and careful field procedures.
- It sometimes happens that ongoing programs already use sharp cutoff scores for strict decisions about program entry, so that it is not necessary to introduce such procedures just for research purposes.
- Impact calculations are not complicated when the data pass appropriate specification tests.

Disadvantages include the following considerations:

- Small deviations from an assumed statistical relationship can lead to spurious impacts, often with biases in favor of the program.
- Results of careful specification tests may lead to the conclusion, after a great deal of field and data-gathering expense, that it is impossible to calculate impacts.
- When used to evaluate a new or ongoing program that has no history of controlling entry with a cutoff score, the regression discontinuity design can be as costly to implement prospectively as an individual random assignment design, and will have most of the disadvantages of such a design (including worries about treatment group take-up and control group contamination). Thus, in most prospective evaluation situations it might be wiser to choose a random assignment design rather than a pure regression discontinuity design, or a hybrid design such as the one employed by Moss, Yeaton, and Lloyd (2014).

2.4 Multiple Thresholds for Enforcement Actions and Multiple RDD Running Variables

With millions of covered workplaces to regulate, but resources for fewer than 20,000 inspections a year, OSHA has been forced to prioritize its use of scarce resources. Starting in 2005, OSHA SST Directives applied a combination of baseline DART and DAFWII to prioritize worksites for Primary and Secondary inspection lists. Since the 2005 SST Directive, a Secondary or Primary list threshold is specified by “a DART rate over \( X \) or a DAFWII rate over \( Y \) (only one of these criteria must be met)”. (Note that \( Y \) is
always smaller than or equal to $X$.) This means there are two running variables for assignment to primary and secondary inspection lists.

Wong, Steiner, and Cook (2013) and Reardon and Robinson (2012) list options for dealing with multiple running variables in RDDs, and we used the “centering approach” to deal with the multiple running variable issue. We selected this approach because of its straightforward implementation and interpretation.

Therefore, for 2006 baseline injury/illness data used as “running variable” for the 2008 SST Directive RDD, cohort Primary and Secondary list worksites, we generated one single RDD running variable using the two original running variables as follows:

First, for each worksite $i$, we centered DART and DAFWII rates to their respective cutoffs, that is $dart_i - dart_c$ and $dafwii_i - dafwiic$.

Second, we chose the maximum centered value $z_i = \max(dart_i - dart_c, dafwii_i - dafwiic)$ as the worksite’s sole assignment score. This ensures that the value of $z_i$ is larger than zero when either DART or DAFWII or both are over their respective cutoffs, and it is negative only when both are below the cutoffs.

The way we generated the centered injury-illness rates assumes equal variances of the DART and the DAFWII case rates.

Since 2011 SST Directive thresholds cannot have been known in 2009, when baseline data for prioritization were gathered, there could be no manipulation of the SST RDD running variable of the kind McCrary (2008) worried about. Nevertheless, we tested for it anyway, and did not find any.

Sample size and MDI formulas for the simplest sharp linear RDD analyses are the same as the formulas for RCT, except that the list of covariates must include the running variable, even though the threshold dummy for the running variable is the key impact variable of interest.

### 3.0 TRADEOFFS BETWEEN RCTs AND MDI-EQUIVALENT RDDs

This section examines the tradeoffs evaluation designers must consider when planning empirical impact work. We consider the issue in a broad context before zeroing in on practical tradeoffs in the context of planning the OSHA SST Evaluation.

#### 3.1 Ex Ante Practical Tradeoffs between Generic RCTs and Generic RDDs

Table 3, presented below, outlines very broad practical tradeoffs, at the planning stage, between generic RCT and generic RDD approaches. It is essential to understand that RCTs and RDDs provide estimates of different parameters. RCTs yield Average Treatment Effects (ATEs), differences in adjusted or unadjusted mean outcomes between all members of the program group and all members of the control group. RDDs yield Local Average Treatment Effects (LATEs) which come from only those sample members within a certain bandwidth of a cutpoint in a running variable for program eligibility. The efficiency with which RDDs use sample data is lower, because, unlike an RCT indicator variable, an RDD

---

11 Note that Wong, Steiner, and Cook (2013) recommend using the min. function, but the assignment is different here, the relationship between the two variables is an OR relationship, while in the Wong, Steiner, and Cook (2013) study it is an AND relationship. Also note that this calculation assumes equal variances of the DART and the DAFWII case rates.
indicator variable for program group membership can be expected to be correlated with other baseline variables.

**Table 3. Generic Ex Ante Methodology Choice**

<table>
<thead>
<tr>
<th>Feature</th>
<th>RCT</th>
<th>RDD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Impact parameter estimated</td>
<td>ATE</td>
<td>LATE</td>
</tr>
<tr>
<td>Planning and execution</td>
<td>Must be prospective</td>
<td>Prospective OR Retrospective</td>
</tr>
<tr>
<td>Minimum wait for results</td>
<td>For a rectangular sample, Length of sample intake Plus Length of follow-up; Plus Time for data preparation and analysis</td>
<td>Time for data preparation and analysis, if retrospective longitudinal data are available</td>
</tr>
<tr>
<td>Source of program process data on administration of the treatment</td>
<td>Customary administrative data, with some special RCT modifications</td>
<td>Customary survey and administrative data</td>
</tr>
<tr>
<td>Level of interference with normal agency processes</td>
<td>Varies</td>
<td>None</td>
</tr>
<tr>
<td>Timing of follow-up</td>
<td>Inflexible; Must be coordinated with time of sample entry</td>
<td>Flexible; Determined at same time baseline is specified</td>
</tr>
<tr>
<td>Vulnerability to unexpected data losses</td>
<td>High</td>
<td>Low</td>
</tr>
<tr>
<td>Validity of results</td>
<td>In general, high internal, low external</td>
<td>High internal, under certain circumstances; generalizable to population served by the program</td>
</tr>
<tr>
<td>Estimation efficiency, ceteris paribus</td>
<td>Highest</td>
<td>Reduced</td>
</tr>
</tbody>
</table>

RCTs must be carried out prospectively, with much cooperation from program staff. Researchers take control of program entry at a certain point in the flow of recruits from the target population into program slots. Data are gathered (1) at baseline, the point of random assignment, when program and control groups are chosen; (2) during program activities, at a minimum to determine the distribution of start and stop dates for program group members; and (3) at follow-up, after most program group members have exited the program and evaluators expect to measure program impacts.

RCTs generally take much more time than alternative RDDs for pre-existing programs with good archived administrative data. They also generally entail changing the normal funnel of program recruits from a target population through eligibility screening and into treatment slots. Just creating an RCT control group entails changing normal procedures at an agency whose services are being evaluated. In addition, to ensure high take-up rates, RCTs often screen more thoroughly than usual for interest in program services, or otherwise interfere with the normal flow of applicants from the target population.
into the program, diminishing their external validity. This feature often makes RCT results less generalizable than comparable RDD results.

Because RCTs must be prospective, the timing of follow-up for a rectangular RCT sample is inflexible. Using archived data for an RDD generally provides much more flexibility. However, if a program is fairly new and no archived data are available, then an RDD would have to be run prospectively, much like an RCT, and face the same kinds of possible operational difficulties as RCTs.

### 3.2 Ex Ante Tradeoffs between the OSHA SST Evaluation RCT and the OSHA SST Evaluation RDD

These tradeoffs took special forms in the OSHA SST Evaluation. Table 4 outlines how the RCT and RDD approaches would have differed had everything gone as planned.

**Table 4. Ex Ante SST Evaluation RCT/RDD Tradeoffs**

<table>
<thead>
<tr>
<th>Feature</th>
<th>RCT</th>
<th>RDD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wait for results</td>
<td>6 years</td>
<td>2 years</td>
</tr>
<tr>
<td>Deviations from usual administrative procedures needed</td>
<td>Withhold controls from online high-rate list; Execute follow-up inspections on a tight schedule</td>
<td>None</td>
</tr>
<tr>
<td>Research questions addressed</td>
<td>ATE on regulatory compliance; ATE on DART</td>
<td>LATE on DART</td>
</tr>
<tr>
<td>Baseline period</td>
<td>2009</td>
<td>2006</td>
</tr>
<tr>
<td>When treatment administered</td>
<td>2011</td>
<td>2008</td>
</tr>
<tr>
<td>When follow-up inspections executed</td>
<td>2013-2014</td>
<td>2010-2011</td>
</tr>
<tr>
<td>Unit of analysis</td>
<td>Worksite</td>
<td>Worksite</td>
</tr>
<tr>
<td>Number of worksites</td>
<td>2,520</td>
<td>7,045 (within bandwidths)</td>
</tr>
<tr>
<td>MDI (Minimum Detectable Impact) on DART if full baseline sample were available at follow-up</td>
<td>12.5% - 13.5% of the sample mean, depending on treatment</td>
<td>12.25% - 20.49% of the sample mean, depending on treatment</td>
</tr>
</tbody>
</table>

Looking first at the RCT column, the 2011 SST Directive Randomized Controlled Trial, begun in 2010, would take a total of six years to complete. Interference with usual agency processes for Letter Only and Letter Plus Inspection controls would consist of withholding their names and addresses from high-rate-letter mailings and from online lists of high-rate-letter recipients. To measure regulatory compliance, follow-up inspections for treatment group sites that had undergone programmed SST inspections in 2011 would have to be executed on tight schedules in 2013 and 2014. ODI Surveys of those same sites, along with their control group counterparts, would have to be fielded and completed during the same time periods. There were 2,520 worksites in the RCT sample, selected because their 2009 injuries and illnesses exceeded the thresholds set in Table 1.
When DOL/CEO learned that, for budgetary reasons, the ODI would be canceled as of 2013, plans for finishing the SST RCT were disrupted. It became impossible to gather survey follow-up data on injuries and illnesses to measure the effectiveness of the Letter-Only and Letter-Plus-Inspection treatments that had been administered in 2011. These surveys, which took place every year from 1996-2012, were intended to be the sources of OSHA SST RCT follow-up data for 2013 and 2014. Cancellation of the ODI survey necessitated cancellation of SST, which was based on thresholds from prior ODI Survey results (see Table 1). The 2012 ODI would be the final OSHA survey of employer-reported injuries and illnesses, providing data for 2011.

Since it likely could have been considered unethical to suppress an RCT whose full baseline sample already had been enrolled, it was important to do the best we could getting follow-up data without the survey and to analyze and report on inconclusive RCT results. But if we stopped there, DOL/CEO’s research questions about OSHA SST would go unanswered.

At a crucial Technical Working Group meeting in 2014, Summit suggested supplementing the OSHA SST RCT with an RDD, using archived OSHA ODI survey and administrative inspections data. The TWG was unanimous in enthusiasm for this possible solution to the predicament of answering DOL/CEO’s research questions without most of the RCT survey data they and CEO had counted on having.

It was decided to make the RDD as similar as possible to the RCT. If the RDD analysis was to use the latest available ODI data for follow-up, the post-program period would have to become 2010-2011 instead of 2013-2014. Most other decisions about the timing and structure of the alternative RDD followed from this decision.

4.0 EX POST TRADEOFFS BETWEEN THE OSHA SST EVALUATION RCT AND THE OSHA SST EVALUATION RDD

Table 5 elaborates the ways an RDD was used to supplement findings from a broken SST RCT. As the “Timing of Follow-up” box in Table 5 explains, 2011 was the final year of outcome data from the annual ODI survey. If we used 2010-2011 ODI data for the RDD, we would have to evaluate the 2008 SST Directive for maximum consistency with the timing of the RCT. That made 2006 baseline injuries and illnesses from the 2007 ODI the source of the RDD running variable. Thus, preparing a dataset for use in determining RDD impacts required merging ODI survey data for 2006, 2010, and 2011 with OSHA inspections data for 2008.
Table 5. SST-Specific Methodology Tradeoffs During Data Preparation and Analysis

<table>
<thead>
<tr>
<th>Feature</th>
<th>RCT</th>
<th>RDD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Planning and execution</td>
<td>Prospective</td>
<td>Retrospective</td>
</tr>
<tr>
<td>Wait for results</td>
<td>6 years</td>
<td>2 years</td>
</tr>
<tr>
<td>Source of program process data on administration of the treatment</td>
<td>Customary administrative data, with some special RCT modifications</td>
<td>Customary survey and administrative data</td>
</tr>
<tr>
<td>Timing of follow-up</td>
<td>ODI surveys that were to supply follow-up injury and illness outcomes were canceled. Many scheduled reinspections were not executed, often because of a two-year rule apparently forgotten in drawing up SST 2013 and 2014 reinspection lists.</td>
<td>Determined at same time baseline was specified. Termination of annual ODI Survey made 2010 and 2011 the final years of follow-up injury and illness data available. Using the same follow-up length as the RCT led to using the 2008 SST Directive for treatment data and the 2006 ODI for baseline and running variable data.</td>
</tr>
<tr>
<td>Other unexpected data losses</td>
<td>Online high-rate-letter list included RCT controls</td>
<td>Difficulty matching and merging files using faulty identifiers</td>
</tr>
<tr>
<td>Impact parameter estimated</td>
<td>ATE</td>
<td>LATE</td>
</tr>
<tr>
<td>Estimation efficiency, ceteris paribus</td>
<td>Highest</td>
<td>Reduced</td>
</tr>
</tbody>
</table>

Table 6 compares the most important features of the two evaluations reported on in the May 2016 OSHA SST Evaluation Final Report.
Table 6. *Ex Post* SST Evaluation RCT and RDD Features

<table>
<thead>
<tr>
<th>Feature</th>
<th>RCT</th>
<th>RDD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wait for results</td>
<td>6 years</td>
<td>2 years</td>
</tr>
<tr>
<td>Research questions addressed</td>
<td>ATE on regulatory compliance</td>
<td>LATE on DART</td>
</tr>
<tr>
<td></td>
<td>ATE on DART</td>
<td></td>
</tr>
<tr>
<td>Deviations from usual administrative procedures needed and executed</td>
<td>Did Not Withhold Most Controls from online high-rate list</td>
<td>None</td>
</tr>
<tr>
<td></td>
<td>Did Not Execute many scheduled follow-up inspections at all, because of two-year rule</td>
<td></td>
</tr>
<tr>
<td>Generalizability</td>
<td>None</td>
<td>To other SST Directives under the same constraints and circumstances</td>
</tr>
<tr>
<td>Baseline period</td>
<td>2009</td>
<td>2006</td>
</tr>
<tr>
<td>When treatment was administered</td>
<td>2011</td>
<td>2008</td>
</tr>
<tr>
<td>When follow-up inspections were executed</td>
<td>2013-2014</td>
<td>2010-2011</td>
</tr>
<tr>
<td>Source of follow-up injury/illness data</td>
<td>Forms 300A collected at follow-up inspections</td>
<td>2010-2011 ODI Surveys</td>
</tr>
<tr>
<td>Proportion of baseline letter plus inspection sample with follow-up injury/illness data</td>
<td>29 percent of controls, 53 percent of the treatment group</td>
<td>100 percent, by construction</td>
</tr>
<tr>
<td>Unit of analysis</td>
<td>Worksite</td>
<td>Worksite</td>
</tr>
<tr>
<td>Number of worksites with injury/illness followup</td>
<td>1,264</td>
<td>7,045 (within bandwidths)</td>
</tr>
<tr>
<td>Ex Post MDI (Minimum Detectable Impact) on DART</td>
<td>17.8% - 21.3% of the sample mean, depending on treatment</td>
<td>12.25% - 20.49% of the sample mean, depending on treatment‡</td>
</tr>
<tr>
<td>Impact results</td>
<td>Inconclusive. Small and statistically insignificant; but severely underpowered ex post</td>
<td>Small and statistically insignificant</td>
</tr>
</tbody>
</table>

‡ While the RDD for the Ex Post MDI on DART was efficient on the Letter group, it lost efficiency on Primary worksites because of an unexpectedly large RDD non-randomization design effect (see Equation 3).

The RCT took six years, while the RDD took only two years from its beginning in 2014. The RDD provided an LATE estimate of the impact of the 2008 SST Directive on injuries and illnesses in 2010 and 2011. The retrospective RDD did not interfere in any way with normal SST operations, but it could not provide impacts on regulatory compliance, because 2010-2011 follow-up reinspections had not been
coordinated with original SST inspections in 2008. RDD results could be generalized to SST Directives other than the 2008 Directive, provided they were carried out to the same extend under the same constraints and circumstances.

OSHA is a large national organization with many different centers of managerial power and authority. The OSHA SST RCT was the first attempt to carry out a random-assignment evaluation of OSHA inspections, a task that still may be necessary given the uncertainties of nonexperimental findings on the effects of inspections by Ruser, Mendeloff, Gray, Smith, and others. The third row of Table 6 lists two additional implementation problems, in addition to the canceled follow-up survey, that befell the 2011 SST Directive RCT Evaluation.

A backup strategy for gathering SST RCT follow-up data without the ODI Survey was to use OSHA Form 300A data, where workplace managers are required to log injuries and illnesses, for submission to OSHA during any programmed inspection. Thus, only RCT worksites that would actually undergo reinspection would supply any follow-up data on injuries and illnesses. But, because of the two-year rule, many reinspections researchers had scheduled never actually took place. Ex post RCT follow-up sample sizes thus would fall far short of baseline sample sizes. Table 6 reveals that just 29 percent of baseline control sites are represented in the RCT follow-up, along with 53 percent of the program group. These rates suggest differential attrition, for which we tested and found. We reweighted the follow-up sample so that it represented the composition of the full baseline sample under the missing-at-random assumption, and found small and statistically insignificant RCT impacts of the 2011 SST Directive on both regulatory compliance and injury and illness outcomes.

Analysis of the RDD yielded similar results. The ex post MDIs for both the RCTs and the RDD were above the impact estimates, so that both analyses are consistent with both insufficient statistical power as well as with zero impacts of inspections. Thus the 2016 DOL/CEO OSHA SST Evaluation does not change the conclusions of its research synthesis on the effects of OSHA inspections.

5. Lessons from DOL-CEO’s 2016 OSHA SST Evaluation, and Possible Implications for Future Evaluation Strategies

Although neither the SST Evaluation RCT nor the SST Evaluation RDD found statistically significant impacts, the project advanced methodologically how program evaluators can apply RDD to existing administrative data to measure impacts. Using a secondary study design reinforced the weak but ex post underpowered RCT impact findings, helped build knowledge about appropriate applications of quasi-experimental designs, and revealed some strengths and some weaknesses of both approaches.

Because they must be run prospectively, RCTs are riskier, more time-consuming, and likely much more expensive than MDI-equivalent RDDs, but can provide the most credible ATE evaluation results possible when they are feasible, enroll adequate sample sizes, minimize noncompliance, and are run well. But it is debatable whether ATEs or LATEs are the parameters of greatest value to policymakers in some contexts, such as whether to expand or contract program scale at margins of prioritization.

---

12 See DOL’s research synthesis on the effects of OSHA inspections https://clear.dol.gov/sites/default/files/CLEAR_OSHAEvidenceEffectsResearchSynthesis_3282014.pdf
13 Ibid.
Well-run RCTs require buy-in, accountability, and regular early feedback from program staff (e.g., OSHA SST Directive writers, inspectors, webmasters, and archivists), and need regular exploration of partial data from the earliest program sites. A superb Technical Working Group with the best experts on OSHA inspections could not guarantee the 2011 SST Directive RCT evaluation operational success. In a large national organization such as OSHA, RCTs pose tremendous management, communication, and accountability challenges. Many of the OSHA staff whose cooperation would have been needed for better operational success of the OSHA SST RCT never may even have known that an experiment was underway.

In sharp contrast, RDDs can be run retrospectively, using administrative and survey data that already have been archived. For example, using the final two years of OSHA’s annual ODI Survey provided injury and illness outcomes for a quickly-mounted RDD impact analysis of the 2008 SST Directive. RDDs can achieve adequate statistical power at much less expense than MDI-equivalent RDDs, even though they are less efficient in using data. But RDDs generally yield LATE estimates, which may be less valuable to policymakers than ATEs in some contexts, such as zero-based budgeting.

Designers of future RCTs might consider planning from the beginning to maximize use of archived administrative data in addition to any special baseline, program process, and follow-up survey data they may gather. As revealed in our study, OSHA’s archived administrative data and archived 1996-2011 ODI Survey data were rich enough to allow developing a backup RDD evaluation plan from scratch when most RCT follow-up data were lost to a survey budget cut.

Random assignment is no panacea for evaluation, and must be carried out prospectively over a number of years. Many things can go horribly wrong at great expense to funders and the validity of the study. It gives precise answers to ATE questions that may be of little interest to stakeholders. If the most crucial decisions to be made require heterogeneous impact estimates, endogenous subgroup impacts, or attribution of impacts to various possible mediators, random assignment may provide no better answers than other evaluation methodologies. There is no single best evaluation methodology for every situation.

If LATEs are the parameters of greatest interest, perhaps an evaluation approach which combines quantitative prioritization with true random assignment within bandwidths around thresholds for program element eligibility would be a first-best evaluation strategy. Program staff may be more amenable to giving researchers control over program admittance decisions when first a kind of triage has taken place. Program staff (e.g., OSHA SST Directive authors) who uses quantitative prioritization may want strongly to serve highest-priority population members, not waste expensive resources on low-priority population members, but have no strong opinion about those in between. Campbell (1969) and Boruch (2016), it should be noted, suggest embedding true random assignment in a band around an RDD cutpoint, on equity grounds. Moss, Yeaton, and Lloyd (2014) is a recent application of the strategy.

Results from an ex post severely underpowered OSHA SST RCT and an RDD for an earlier year were similarly inconclusive. If both RCTs and RDDs are credible methodologies for evaluating OSHA activities, the LATEs from RDD evaluations might be more valuable for stakeholders than the ATEs from RCTs, if dispensing with workplace health and safety inspections entirely is off the table. The idea of determining LATEs at various cascading thresholds of a numeric priority could be quite attractive for helping policymakers decide whether and where to deploy additional resources for various levels of program resource use. Implemented correctly, with a large enough sample, cascading multiple thresholds for the
same running variable could be a fairly rigorous tool for getting at heterogeneity in program impacts due to differential eligibility for program elements of various perceived strengths.

REFERENCES


