WHD Compliance Strategies: Directions for Future Research

June 2020

Sarah Dolfin, Nan Maxwell, and Ankita Patnaik

Submitted to:

U.S. Department of Labor
Chief Evaluation Office
200 Constitution Avenue, NW
Washington, DC 20210
Project Officer: Jessica Lohmann
Contract Number: DOLQ129633249

Submitted by:

Mathematica
1100 1st Street, NE, 12th Floor
Washington, DC 20002-4221
Telephone: (202) 484-9220
Facsimile: (202) 863-1763
Project Director: Sarah Dolfin
Reference Number: 50519

DISCLAIMER

This report was prepared for the U.S. Department of Labor (DOL) Chief Evaluation Office by Mathematica, under contract number DOLQ129633249. The views expressed are those of the authors and should not be attributed to DOL. Mention of trade names, commercial products, or organizations does not imply endorsement of same by the U.S. government.

ACKNOWLEDGMENTS

The authors acknowledge several people who contributed to and supported this research. Karen Livingston in the Wage and Hour Division (WHD) and Deborah Martierrez, Sam Rowe, and Jessica Lohmann in the Chief Evaluation Office provided guidance and support throughout the project. Before she returned to WHD, Libby Hendrix worked closely with us as a consultant to conceptualize how strategies might be evaluated.

Members of the Technical Working Group—Wayne Gray, David Levine, and Jay Shimshack—were joined in discussions by WHD staff—Karen Livingston, Brandon Brown, Michael Kravitz, Naixa Franquiz, Sara Johnson, and Dan Weeks—that resulted in a dynamic analysis of the difficulties in evaluating WHD strategies and initiatives and established the need for this feasibility study. Matthew Johnson provided very helpful feedback on the draft report.

From Mathematica, Jill Berk reviewed the report and provided comments that greatly enhanced its quality. Sharon Clark and Allison Pinckney helped prepare it, and Kerry Kern and Effie Metropoulos provided editorial assistance.

This page has been left blank for double-sided copying.

Contents

Tables

Figures

ABSTRACT

Monitoring and evaluation helps guide decision making by the Wage and Hour Division (WHD) of the U.S. Department of Labor by building an understanding of the strategies it adopts to encourage compliance with laws and regulations. This report identifies five factors underlying monitoring and impact evaluation efforts that can produce information about the effectiveness of a strategy and whether a lack of expected outcomes stems from ineffectiveness or from difficulties in implementation. Monitoring requires (1) documented and supported core activities, (2) measurable outputs and outcomes, and (3) available, appropriate data. An impact evaluation also requires (4) mature implementation and (5) internal validity. The report builds an understanding of these factors by developing a framework to illustrate how WHD might apply them, discussing the opportunities and challenges WHD faces in implementing them, and describing steps to consider to ensure potential future monitoring and evaluation yields useful information. The report does not identify or build a design for any specific monitoring or evaluation activity. Instead, it provides a theoretical framework and important considerations that can assist in ensuring that strategies are implemented in a way that makes them better suited to monitoring and evaluation. The options, ideas, and illustrations discussed are not intended for use as-is. Should WHD decide to conduct an evaluation, a specific design would be necessary to address the specific circumstances of that evaluation.

This page has been left blank for double-sided copying.

EXECUTIVE SUMMARY

The Wage and Hour Division (WHD) of the U.S. Department of Labor (DOL) protects and enhances the welfare of the nation's workforce by promoting and achieving compliance with labor standards. The statutes that WHD enforces give core protections to at least 143 million workers in more than 9.8 million establishments throughout the United States and its territories.[1] Data and research inform WHD compliance strategies and help the agency monitor and evaluate how effectively it uses those strategies.

DOL's Chief Evaluation Office contracted with Mathematica to support WHD in continuing to build evidence on the effectiveness of opportunities for evaluation of WHD's compliance strategies. Part of that effort involved assessing the opportunities and challenges of conducting an impact evaluation that could gauge the effectiveness of a WHD compliance strategy, called the "directions for future research" study. The study is the subject of this report.

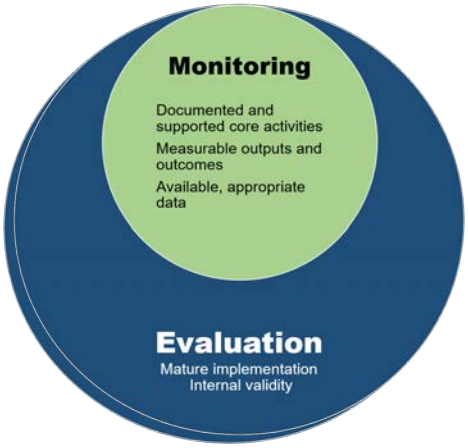
The directions for future research study is a resource for WHD and potentially other regulatory agencies considering how to evaluate the impacts of their activities, based on an up-to-date scan of literature and guidance from a wide range of technical experts. This report also highlights general good research practices, which do not reflect on what WHD has done or may be doing; WHD is already conducting many of these activities. Rather the report's focus is on building a broad theoretical framework and identifying important considerations that can assist in ensuring that strategies are implemented in a way that makes them better suited to monitoring and evaluation. It is important to note, however, that the options, ideas, and illustrations discussed here are not intended for use as-is. Should WHD decide to conduct an evaluation, a specific design would be necessary to address the specific circumstances of that evaluation.

Monitoring and evaluation processes help guide WHD's decision making by building an understanding of the strategies that it adopts. We define monitoring as ongoing tracking of a strategy's components—including activities conducted during implementation and outcomes observed afterwards—that provides information on the progress or delay of a strategy's achieving expected outcomes. We define evaluation as impact evaluation, building on the knowledge gleaned from monitoring and establishing whether and how much a strategy caused outcomes to change. To develop these deeper causal insights, evaluation has more requirements than monitoring. WHD may be able to use these tools as a starting point when developing monitoring tools or an evaluation of specific strategies when possible. These tools may not only help assess the value of strategies, but may also help identify whether a strategy's lack of expected outcomes is due to challenges in execution (implementation failure) or the strategy's ineffectiveness (theory failure).

[1] See https://www.dol.gov/agencies/whd/workers.

A critical component for monitoring and evaluation of a strategy is a well-articulated theory of change (TOC; see Weiss 1995). A theory of change explains a process of change by showing assumptions about causal steps that lead from program activities to outcomes and is frequently represented by a logic model. It ensures that all stakeholders—from administrators and policy advisers to district and regional directors and community outreach staff—share an understanding of how a strategy is expected to unfold to achieve expected outcomes. In addition, the four components of a strategy's TOC—inputs, activities, outputs, and outcomes—make monitoring and evaluation relatively straightforward.[2] A clear sense of what is needed to increase compliance makes it easier to identify what needs to be measured, both to monitor the strategy's performance and to determine if it is achieving its desired outcomes. This clarity in measurement will reveal the data that are needed for monitoring and performance, and these data needs can be compared to the data that are available.

Figure ES.1. Monitoring and evaluation



There are five factors to consider when designing a strong monitoring and evaluation process for a strategy and assessing whether the strategy is suitable for monitoring and evaluation.[3] The first three apply to both monitoring and evaluation, and the last two apply only to evaluation. Figure ES.1 summarizes the factors.

1. **Documented and supported core activities**. For both monitoring and evaluation to provide an assessment of a strategy's potential, the strategy must have (1) well-defined activities with key strategy components that are consistently implemented across employers, and (2) stakeholders and staff who support the monitoring and evaluation effort.
2. **Measurable outcomes**. It is impossible to assess the value of a strategy without examining its outcomes. To see how outcomes might follow from or be caused by a strategy's activities and inputs, they must be able to be quantified and captured (that is, observable) during the period of monitoring and evaluation.
3. **Available, appropriate data**. The most informative monitoring and evaluation efforts rely on comprehensive data that capture all the elements in the TOC at the appropriate unit of

[2] Inputs are the human, financial, and physical resources needed to perform the activities (core components) that are critical to a strategy's success. Activities are the things done through a strategy to bring about change. Outputs are direct, tangible products of a strategy's activities. Outcomes are the measurable changes in employer behavior that are expected to occur as a result of the strategy. See Tatian (2016).

[3] The factors were developed based on the literature on implementation science and impact evaluation design, including Tatian (2016), Fixsen et al. (2005), Fixsen and Blase (1993), Clearinghouse for Labor Evaluation and Research (n.d.), and What Works Clearinghouse (n.d.).

analysis. This means the data must include appropriate measures of all four components of the TOC (long-term, intermediate, and short-term outcomes; outputs; activities; and inputs) as well as contextual or environmental factors. Importantly, data must be at the appropriate unit of observation. If a compliance strategy is expected to affect individual employers' behavior, the data must be at the employer level (that is, employer- or establishment-level data). If it is to affect all employers in a geographic area (say, a metropolitan area), the data need to be at that level. Data should also be of high quality, containing reliable and complete information.

To learn whether a strategy caused any changes that are observed in outcomes, WHD may consider selecting a strategy whose characteristics allow it to be the subject of a rigorous evaluation and ensure that the evaluation's design allows it to provide accurate, actionable information that can be used to improve its compliance activities. WHD could consider two additional factors to ensure this.

4. **Implementation maturity**. Evaluation methods can be used to answer a range of questions related to implementation and outcomes. Understanding the maturity of a program or strategy's implementation will determine the appropriate evaluation methods. During the early stages of implementation, formative evaluation questions are more focused on refining the program design and determining effective approaches to implementation. After the implementation of the strategy is "mature," an impact or causal evaluation can provide a valid assessment of whether a strategy can increase compliance. That is, WHD has created any required infrastructure and integrated the strategy's inputs and activities into regular WHD routines, and outputs and outcomes follow the activities. Importantly, maturity does not require a long time to achieve, nor must there be perfect consistency in delivery across locations.

5. **Internal validity**. An evaluation with internal validity can clearly separate the outcomes determined by the strategy from other factors that may have impacted them by using a carefully constructed counterfactual condition (what would have happened if the strategy had not been implemented). Two types of evaluation designs could have internal validity: a randomized controlled trial (RCT) and a quasi-experimental design (QED). RCTs are usually considered the gold standard in design, whereas results of QEDs might not be as rigorous. Evaluation designs that can provide causal evidence often have greater needs for data than monitoring efforts do (for example, because they might need data from multiple points in time or for a suitable comparison group).

   Existing agency processes and resources, as well as the features of strategies, confront agencies with both potential opportunities and challenges in building and implementing monitoring and evaluation processes that will provide timely, insightful information to allow the agency to assess whether and how its strategies are improving compliance with the laws and regulations it enforces. In Table ES.1, we summarize general potential opportunities and challenges for designing and supporting monitoring and evaluation that address each of the five factors. The table also presents steps that could be considered to ensure potential future monitoring and evaluation yields useful information.

Table ES.1. Considerations for successful monitoring and evaluation of implemented strategies

Factors for successful monitoring and evaluation	Potential opportunities	Potential challenges
Successful monitoring and evaluation requires:		
1. **Documented and supported core activities**. Strategy has well-defined activities, with key components implemented consistently across employers, stakeholders, and staff who support monitoring and evaluation.	a. Internal coordination efforts offer opportunities to define and gain agreement on the TOC and to consistently implement core activities. b. Existing documentation and guidance can provide a foundation for a well-articulated TOC.	a. Complex strategies, or those that have extensive data collection needs, are often those in most need of stakeholder agreement on the TOC, and often require more resources to gain needed agreement.
2. **Measurable outputs and outcomes**. Outputs and outcomes can be quantified and observed during the period of monitoring and evaluation.	a. Violation and performance measures can help structure measurable outputs and outcomes.	a. Stakeholders may disagree on which measures to examine.
3. **Available, appropriate data**. Data capture the components of the theory of change and context comprehensively, are at the appropriate unit of observation, have high quality, and include appropriate outcomes and supplementary information.	a. Data quality assurance procedures could be formalized, including by aligning performance standards with data quality. b. Administrative data might be modified to capture additional elements in the TOC and thereby become the basis for a monitoring and evaluation data collection system. c. Electronic metadata might be leveraged as an inexpensive source of data. d. Existing interactions with entities receiving strategies offer opportunities to collect data. e. Follow-up investigations could provide valuable information to enforcement agencies. f. Investment in an external sampling frame could strengthen monitoring and evaluation efforts.	a. Modifying existing data systems to collect additional data for monitoring and evaluation may be difficult. b. Spillovers, which are often outcomes of strategies, can be difficult to capture.
Successful evaluation also requires:		

Factors for successful monitoring and evaluation	Potential opportunities	Potential challenges
4. **Implementation maturity.** Implementation maturity has been determined through systematic monitoring of inputs, activities, and outputs.	a. Well-established strategies can serve as a starting point to develop evaluations. b. Agencies can use monitoring of strategies as a way to ensure TOC components are in place.	a. Stakeholders may want evidence on strategies that do not have all components in place; the results of evaluating such strategies might understate their potential value.
5. **Internal validity.** Evaluation can separate the outcomes determined by the strategy from other factors that may have impacted them because there are no confounding factors influencing both the outcome and the strategy, and there are data on a comparison group representing the counterfactual (what would have happened in the absence of the strategy).	a. Performance standards aligned with evaluation goals may create incentives that support evaluation. b. Agencies could build on existing processes to develop RCTs and other rigorous designs for evaluation. c. Geographic variation in strategies could be exploited in developing evaluation designs.	a. Spillover effects make it difficult to capture a counterfactual. b. Random assignment may not be feasible for evaluation of some strategies.

Note: This table is intended to be illustrative. Should WHD decide to conduct an evaluation, a specific design would be necessary to address the specific circumstances of that evaluation.

RCT = randomized control trial. TOC = theory of change.

The following steps outline activities that agencies such as WHD could consider to ensure potential future monitoring and evaluation yields useful information. Note that the steps are intended as a general resource reflecting good research practices and that WHD already engages in many of these activities.

A. Build data infrastructure suitable for monitoring and evaluation

Agencies could consider investing in several efforts that may ensure that all data for monitoring and evaluation are of high quality.

1. **Build capacity for ensuring data quality and conducting data analytics**. Potential approaches to consider include investing in training for staff who report and analyze data and developing procedures to verify data and ensure their quality.
2. **Consider ways that administrative data could be collected and used to support the internal validity of evaluations.** Enforcement agencies might consider conducting follow-up investigations to develop panel data on establishments and estimate changes in compliance. They could use external data to develop a sampling frame, from which they could select establishments to investigate using random sampling and estimate violation prevalence. Alternatively, agencies could establish a protocol of pro-actively linking establishments to external data before an investigation begins, which would facilitate efforts to validate both external and administrative data and enhance the value of the external data by improving its match rate to the administrative data.
3. **Develop and maintain data sets that could be linked to agency data systems**. For example, these data could come from primary data collection on strategy implementation activities and intermediate outcomes; or through acquiring external data that reflect, for example, a population of establishments and their characteristics. Building capacity around the ongoing maintenance and statistical modeling of these data sets could help evolve monitoring and evaluation activities and improve the quality of analysis possible.

B. Specify performance measures and data collection needs to support evaluation goals

Agency staff often play a key role in implementing strategies. Agency performance measures create strong incentives for them to do this work in particular ways, which could make it challenging for staff to support evaluation activities. For example, performance measures related to efficiency (output per labor hour) could discourage staff from spending additional time to engage in implementation activities required for an evaluation. Agencies could consider creating performance measures aligned with monitoring and evaluation goals to support staff and ensure the production of high-quality evidence. Such measures might reflect specific activities conducted for an evaluation or specific contributions to data quality (such as the percentage of cases that went through quality control review).

C. Develop a system for monitoring and evaluating strategies consistent with the evaluation design

It may be helpful to consider developing a system for monitoring and evaluation of strategies. The advantage of creating such a system is that monitoring and evaluation processes could become more efficient and consistent through repeated use, and that all strategies could have the opportunity for monitoring and evaluation. It would be important to build consensus within an agency and among stakeholders at each stage to strengthen the monitoring and evaluation designs and help reach agreement about how to interpret the findings. The monitoring and evaluation process could include the following steps based on a summary of the discussion of five factors for consideration:

1. **Build a detailed, evidence-driven TOC for each strategy.** To describe the components of the TOC—inputs, activities, outputs, and outcomes—in detail, data on how the strategy is actually implemented and the specific goals it pursues are important. These data could be collected in many ways, including through review of documents, observation of activities, and interviews with field staff who deliver the strategy and key personnel associated with the entities that receive the strategy.
2. **Monitor strategies to determine their maturity.** To gauge whether a strategy is being implemented as planned, it is important to identify and analyze data measuring the components of the TOC. The maturity of implementation and the extent to which the TOC has been implemented will determine the appropriate evaluation design. Both qualitative and quantitative data can be collected through these activities using a range of techniques. For example, interviews with staff and stakeholders, administrative data analysis, and research can all support learning and ongoing improvements during program implementation.
3. **Develop appropriate evaluation designs.** Integrating evaluation design with planning for implementation could offer the best chance for a successful evaluation. By planning an evaluation before implementation, agencies may be able to improve or strengthen the conditions that support success, such as gathering or enhancing documentation and data during implementation and constructing a counterfactual condition for evaluating the relative effectiveness of strategies.

This page has been left blank for double-sided copying.

I. INTRODUCTION

The Wage and Hour Division (WHD) of the U.S. Department of Labor (DOL) protects and enhances the welfare of the nation's workforce by promoting and achieving compliance with labor standards. WHD enforces statutes to support these standards—statutes that give core protections to at least 143 million workers in more than 9.8 million establishments throughout the United States and its territories.[4] Data and research inform WHD compliance strategies and helps the agency monitor and evaluate how effectively it uses those strategies.

DOL's Chief Evaluation Office contracted with Mathematica to conduct the study titled Evaluation Research on Wage and Hour Division's Compliance Strategies. The goal of the study was to support WHD in continuing to build evidence on the effectiveness of opportunities for evaluation of WHD's compliance strategies. Part of that effort involved assessing the opportunities and challenges of conducting an impact evaluation that could estimate the outcomes of a specific WHD compliance strategy, called the "directions for future research" study. The study is the subject of this report.

The directions for future research study grew out of a plan earlier in the contract to assess the potential to evaluate the impact of a high-priority WHD strategy. Mathematica learned from WHD leadership that, while the agency has a range of tools and practices in place to assess strategies and compliance and has engaged in many types of studies of compliance, impact evaluations could provide valuable information. Working with WHD, Mathematica examined a range of initiatives meeting specific criteria and consistent with WHD's multipronged approach, and prioritized the ones that might be valuable to study. These strategies aimed to leverage the overall structure of industries in order to influence compliance beyond just those employers WHD could investigate. WHD relied on combinations of strategies that could include press releases and drop-in articles tailored to industry, stakeholder engagement, and a range of compliance assistance tools and resources. One of the strongest strategies we considered was leveraging voluntary cooperation of a lead entity or brand within several industry subsectors, a strategy we refer to as strategic partnerships in this report. To help WHD think systematically about the challenges to conducting impact evaluations of their strategies, Mathematica developed options and considerations for conducting monitoring and evaluation of WHD's strategies.

This report on directions for future research develops a framework for WHD to consider in constructing monitoring and evaluation processes for one or more of its strategies. It builds on information gained from (1) a literature and database review that identified the knowledge gaps a study like this might fill (Dolfin et al. 2018); (2) discussions with WHD about compliance strategies; and (3) discussions with a panel of experts about compliance strategies, including strategic partnerships. The report uses that framework to illustrate how WHD might assess the evaluability of compliance strategies and design strong monitoring and evaluation processes. It

[4] See https://www.dol.gov/agencies/whd/workers.

goes on to present steps that could be considered to ensure potential future monitoring and evaluation yields useful information.

The directions for future research study is intended as a resource for WHD and other agencies considering how to evaluate the impacts of their activities. It is meant to highlight good research practices and is not a reflection on what WHD has done or may be doing; WHD is already conducting many of these activities. This report does not identify or build a design for any specific monitoring or evaluation activity. Instead, it provides a theoretical framework and important considerations that can assist in ensuring that strategies are implemented in a way that makes them better suited to monitoring and evaluation. It is important to note, however, that the options, ideas, and illustrations discussed here are not intended for use as-is. Should WHD decide to conduct an evaluation, a specific design would be necessary to address the specific circumstances of that evaluation.

This chapter provides context for the report by briefly summarizing the research on employer compliance with laws and regulations (Section A), describing how monitoring and evaluation can be used to assess a strategy's potential (Section B), and giving a roadmap to the report (Section C).

A. WHD's approaches to building compliance with laws and regulations

WHD works to improve employers' compliance with labor standards. In this context, compliance means taking steps to understand labor standards and follow them in good faith. Compliance is often measured in terms of avoiding violations of labor laws and regulations. To promote and enforce compliance, WHD pursues both enforcement strategies like investigations and compliance assistance strategies designed to encourage employers to voluntarily comply with labor standards. Both approaches have support in the research literature.

1. **Voluntary compliance strategies (compliance assistance)** give employers information and tools to promote their compliance. Grounded largely in social theories on employer behavior, these strategies place employers' decision making about noncompliance in the larger social context, in which employers are actors within society and not just calculators of economic benefits and costs. This context can include the employer's organizational environment (Sutton 1998; Barnes and Burke 2006), social norms (Cialdini et al. 2006; Parker 2006; Behavioural Insights Team 2012), expectations or perceptions about behavior (Friedrichs 2009; Gray and Silbey 2014), and ethical views about behavior (Calavita 1990; Parker 2006). Strategies like public awareness and self-monitoring programs that are thought to influence employer behavior could be effective in this larger context, although evidence on the effectiveness of voluntary compliance strategies is mixed. Educating employers by raising public awareness can increase knowledge about laws and regulations, but it is unclear whether this knowledge translates into compliance (Sneed et al. 2007; Evans et al. 2011), and

self-monitoring programs have not been shown to consistently improve worker rights and working conditions on their own (Locke et al. 2009).

2. **Enforcement strategies** are designed to uncover noncompliance with labor regulations and enforce consequences. Although enforcement could lead employers to comply through a variety of mechanisms—for example, because they want to follow the law, or because the strategies make compliance more salient or overcome other behavioral bottlenecks—researchers often study these strategies in the context of rational choice theory. That is, they assume that employers make decisions about compliance and achieve compliance outcomes after making rational calculations of the perceived benefits and costs (Ashenfelter and Smith 1979; Kagan and Scholz 1984; Chang and Ehrlich 1985). Such calculations can incorporate the perceived certainty of detection, size of penalties, and damage to reputation, which are determined by the enforcement strategies of investigations and the threat of penalties and damages. Research support for enforcement strategies is strong, showing that investigations increase compliance (Johnson et al. 2017; Levine et al. 2012; Gray and Mendeloff 2005; Gray and Scholz 1991, 1993; Jin and Lee 2014), especially when accompanied by penalties and damages (Galvin 2016). Research also reveals that the severity and certainty of penalties drive the cost of noncompliance (van Rooji and Fine 2017), that penalties imposed on one employer can deter violations on the part of other employers by increasing the perceived probability of detection (Braithwaite and Makkai 1991; Gray and Scholz 1991; Gray and Shadbegian 2005), and that damage to reputation increases the cost of noncompliance for an employer (Weil 2012). A body of work suggests that enforcement tools might be better understood within a social context. For example, enforcement tools seem to be more effective when the relationship is a simple employer-employee one, compared to a subcontracting relationship, for example (Weil 2014).

Strategic partnerships: Example of a voluntary compliance strategy

WHD's strategic partnerships involve leveraging the cooperation of a lead entity or brand in an industry. Partnerships make industry leaders, who may not have an employment relationship with the employees of the investigated establishments, a primary actor in compliance by encouraging them to work proactively with WHD to improve compliance with labor laws and regulations. within their network. The strategy engages a single key entity (such as a company headquarters, franchisor, industry association, or related entity) to broadly promote compliance-related activities to the many establishments it interacts with. The partnership strategy grew out of the agency's emphasis on compliance assistance as part of a multipronged approach to securing compliance. and the strategy has been part of its performance plan since 2003. When partnering with brands, WHD typically provides training, enforcement history data, and information on compliance assistance to support the brand's voluntary compliance efforts. Evidence on the effectiveness of partnerships is mostly qualitative (Fine and Gordon 2010; Kagan et al. 2003).

B. Monitoring and evaluation

Monitoring and evaluation help guide WHD's decision making by building a better understanding of the outcomes and impacts of the strategies it uses. We define monitoring as ongoing tracking of a strategy's components, including activities conducted during implementation and outcomes observed afterwards. We define evaluation as impact evaluation. Impact evaluation builds on the knowledge gleaned from monitoring and establishes whether, and how much, a strategy caused outcomes to change. As we will discuss, to develop these deeper causal insights, evaluation has more requirements than monitoring. WHD can use these tools not only to help assess the value of strategies, but also to potentially identify whether a strategy's expected outcomes did not materialize because of issues with execution (implementation failure) or because the strategy itself was ineffective (theory failure) (see Wandersman 2009). The discussion of monitoring and evaluation in this report can help guide WHD as it engages in these activities and provide a starting point for developing the specific designs that would be needed.

- **Monitoring can help identify implementation failure.** Monitoring may help WHD assess whether the components and outcomes of implementation are unfolding the way they were expected to. It can answer questions about whether the strategy is meeting its targets and, if it is not, why (see key research questions in the sidebar). This can allow for continuous quality improvement, and the strategy could achieve more of its potential because monitoring has yielded nuanced information about how to adjust (Tatian 2016). For example, knowing there is progress and change in some outcomes and not others can help identify which program components are working well and which ones might need to be adjusted, or which program components bring about faster or slower changes. Monitoring can also help identify whether processes and other environmental and institutional factors might impede how a strategy works. WHD may or may not be able to overcome the challenges uncovered as part of monitoring; if it cannot, it might conclude that the strategy cannot be implemented in a way that allows it to achieve desired outcomes. By the same token, monitoring can help identify aspects of local conditions such as industry composition or population characteristics that may be associated with effective implementation (Patnaik 2020).

- **Evaluation can help identify theory failure.** If monitoring shows that the strategy is largely being implemented as intended, and expected outcomes follow, an evaluation could

determine whether the strategy had caused the desired change in outcomes and reveal the magnitude of the change that it caused (see key research questions in the sidebar). If, instead, an evaluation showed no changes in outcomes despite good implementation, it would imply that the theory about what a strategy would do was wrong: the strategy was not effective. Although monitoring can, at best, show that the implementation of a strategy was *correlated with* some change in outcomes, an evaluation can produce evidence that the strategy *caused* the intended change in outcomes (Wandersman 2009). For an evaluation to provide this evidence, however, it must be designed so that WHD can have confidence that the information it provides is relevant and relatively free from inaccuracies and bias.

Although WHD extensively monitors its strategies, this report considers monitoring and evaluation in a common framework to support the agency in thinking about what additional efforts could help build both stronger monitoring and rigorous evaluation.

C. A roadmap to this report

This report builds an understanding of the factors to consider when developing monitoring and evaluation plans for a strategy designed to improve compliance. It does not identify or build a design for any specific monitoring or evaluation activity. Instead, it provides a theoretical framework and important considerations that can assist in ensuring that strategies are implemented in a way that makes them better suited to monitoring and evaluation. It is important to note, however, that the options, ideas, and illustrations discussed here are not intended for use as-is. Should WHD decide to conduct an evaluation, a specific design would be necessary to address the specific circumstances of that evaluation. The report begins with a framework for monitoring and evaluation (Chapter II), illustrating considerations and challenges with the strategy of strategic partnerships. The framework is then used to illustrate how WHD might assess whether a compliance strategy is suitable for monitoring and evaluation (Chapter III), and to discuss activities that could strengthen the conditions supporting successful monitoring and evaluation of strategic partnerships and other strategies. Next, the report discusses opportunities and challenges WHD faces in designing and supporting monitoring and evaluation of its compliance strategies (Chapter IV). It concludes with steps that could be considered to ensure potential future monitoring and evaluation yields useful information (Chapter V).

This page has been left blank for double-sided copying.

II. A FRAMEWORK FOR DESIGNING MONITORING AND EVALUATION

Not all monitoring and evaluation produces accurate, relevant information. When done with care, monitoring and evaluation can systematically provide ongoing information about a strategy's performance and potential that WHD can use to assess the strategy's general effectiveness and the context it would be most effective in. In contrast, if monitoring and evaluation are based on incomplete information and not thought through, findings can be misleading.

This chapter presents the key considerations to designing a strong monitoring and evaluation process for a strategy, using strategic partnerships to illustrate them. Section A discusses the importance of a well-defined theory of change (TOC) in articulating and documenting the implementation and impact of a strategy. Section B describes the five factors that may be considered to develop monitoring and evaluation processes that can provide relevant, reliable information. These factors help frame the discussion in Chapter III about how WHD might assess the suitability of a compliance strategy for monitoring and evaluation and the discussion in Chapter IV on opportunities that might be available for monitoring and evaluating a strategy.

A. The value of a well-defined, measurable, and documented theory of change

A well-articulated TOC will build understanding of how a strategy is expected to link activities to change employer behavior in a way that improves compliance with the laws and regulations WHD enforces. It details how the activities undertaken as part of a strategy will increase compliance. Indeed, the process of developing a theory of change can help WHD clarify its strategy and construct measures that can be used to assess whether the strategy would lead to expected outcomes. (Hodges et al. [2002] provides an example.) A well-articulated TOC ensures that all WHD stakeholders—from administrators and policy advisers to district and regional directors and community outreach staff—share an understanding of how a strategy should unfold to achieve expected outcomes, and an understanding of the factors that affect its implementation.

A TOC identifies the human, financial, and physical resources (**inputs**) required to implement critical components of the strategy (**activities**). These activities change the conditions (**outputs**) that ultimately improve the **long-term outcome** of compliance with the laws and regulations WHD enforces. To increase compliance, behaviors targeted by the strategy's activities might need to change (which are the TOC's **intermediate outcomes**), and these behavioral changes might unfold in stages (which are the TOC's **short-term outcomes**). These four components of a strategy's TOC (inputs, activities, outputs, and outcomes) are often influenced by situations, forces, or circumstances that exist within or outside WHD (**contextual factors**). Figure II.1 builds on WHD's earlier work on a TOC, and uses the example of a strategic partnership strategy to show how a TOC might be structured. The hypothesized links between components in the example would in practice be supported by data or literature. In the discussion that follows, the

examples and suggestions we present are intended to be broadly illustrative rather than comments on any work WHD has done.

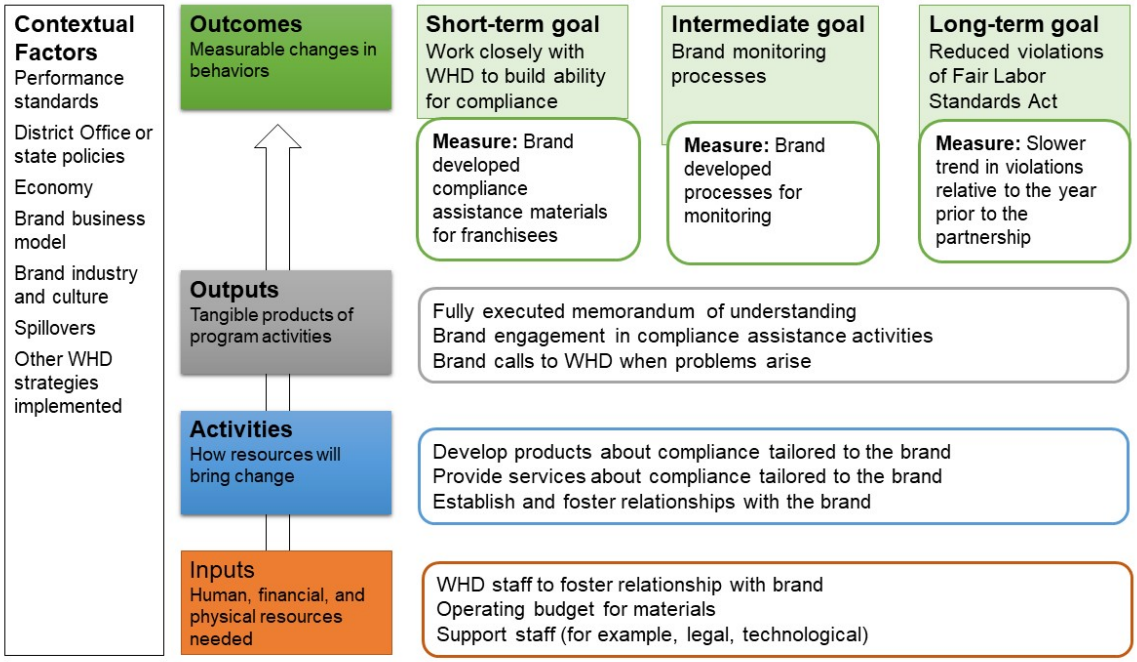
1. **Inputs** are the human, financial, and physical resources needed to perform the activities (core components) that are critical to a strategy's success. Examples include staff, time, money, equipment, facilities, supplies, software, and written materials. The TOC illustrates a partnership strategy that uses staff and financial resources to work with a brand, as shown to the right of the orange portion of the figure.
2. **Activities** are the things done through a strategy to bring about change. Although the complexity of some strategies can make individual activities difficult to capture, core activities should be identified in a TOC. If activities vary by region (for example), these variations should be recognized and documented to build an understanding of how contextual factors influence a strategy and its outcomes. In addition, a well-specified TOC would identify who delivers and engages in the activities and what the target audience for them is. For example, WHD leaders could develop compliance assistance materials and services tailored to a brand's needs and give them to leaders at the brand headquarters to foster a close relationship, as shown in the blue portion of the figure.
3. **Outputs** are direct, tangible products of a strategy's activities. An output describes who is affected—directly and indirectly—by the strategy and what products of the strategy lead to the desired outcomes. The outputs are used to monitor and report progress toward outcomes because they provide evidence that the strategy's activities are being implemented as planned. For example, as part of its working relationship with WHD, the brand could call WHD when issues arise, engage in activities to promote compliance assistance, and execute a memorandum of understanding (MOU) with WHD. Outputs are shown in the gray portion of the figure.
4. **Outcomes** are the measurable changes in employer behavior that are expected to occur as a result of the strategy. The outcomes are often the same as the program goals, and must clearly articulate the baseline (that is, starting point) against which the strategy's progress is assessed. WHD's overarching goal for a strategy is clear: its mission is to promote and achieve compliance with labor standards to protect and enhance the welfare of the nation's workforce. In practice, it can be difficult to assess whether compliance is increasing. As a result, short-term and intermediate outcomes are often used to assess whether the strategy is increasing compliance. This approach works if the shorter-term outcomes (1) have been shown to be associated with compliance and (2) could justify the cost of conducting an evaluation into whether a strategy reduces violations. In the short term, the brand might work closely with WHD to develop compliance assistance materials for its franchisees that address common violations. In the intermediate term, the brand might adopt an internal process for monitoring compliance, as shown in the green portion of the figure. These outcomes might translate into a long-term outcome of fewer FLSA violations across the brand's establishments.

5. **Contextual factors** can influence both a strategy's implementation of activities, outputs, and outcomes and its effects. General examples might include institutional, community, and public policies that could support or impede either the strategies themselves, or the business model or industry of establishments targeted by the strategy. Importantly, these factors might be at different levels; for example, the regional economy is at the broadest level whereas the dominant business strategy in the industry and an employer's culture are narrower. In the strategic partnership example, factors could include the brand's business model or culture, the consequences faced by other firms in the industry (which could create spillover effects on the brand), WHD District Office policies or state laws that support the strategy, or other strategies that WHD may be implementing (for example, targeted enforcement activities). Such factors can influence strategy inputs (for example, the skills of the field staff needed to implement the strategy might vary with the culture of the industry), the activities undertaken (for example, some brand's cultures might be more receptive to incumbent worker trainings than others), expected outputs (for example, the incidence of WHD complaints may be greater when enforcement activities are strong), or expected outcomes (for example, intermediate outcomes of brand processes to resolve complaints might be put into place quicker when the labor market is tight).

Identifying the components of a TOC makes measurement and evaluation relatively straightforward. Articulating the assumptions for how change will occur makes it easier to identify what needs to be measured, both to monitor the strategy's performance and to determine if it is achieving desired outcomes.

WHD could consider following this example to build TOCs for strategies in the future. Although this example was described chronologically, starting with inputs, developing a theory of change generally begins by thinking about a strategy from the point of the desired goal. WHD could then proceed to identify all the assumptions about how the strategy will achieve that goal (Hernandez and Hodges 2003; Yampolskaya et al. 2004). Resources to consider include reviews of literature related to the strategy or the long-term outcomes, as well as documentation and data on WHD's implementation of the strategy.

Figure II.1. Theory of change: Example



B. Factors for successful monitoring and evaluation

Monitoring and evaluation work together, with an informative evaluation building on the knowledge generated from ongoing monitoring of a strategy. Monitoring the inputs and activities conducted as part of the strategy, as well as the outputs and outcomes that are produced, can help an agency identify implementation challenges and failures. Examples of the differences between actual and expected implementation abound. Consider that:

- An agency might not have the kind of human or physical resources needed to execute the strategy. Resources may not be available to fully pilot or experiment with strategies that go beyond the existing staff workload, or to build capacity to fully implement a strategy. This could be a problem for the strategic partnership strategy.
- An agency might not have the infrastructure to facilitate the implementation activities. For example, changing resource availability or priorities can make it challenging to carry out the planned partnership activities over the long term.
- Expected outputs and outcomes might follow only in some circumstances, highlighting the importance of accounting for context. For example, a strong economy might help a strategy achieve expected outputs and outcomes because employers have trouble hiring workers and view WHD as an ally in building the labor conditions that might attract the workers. In the example of strategic partnerships, some partner brands could have been partly motivated to engage with WHD to support plans for expansion due to strong economic conditions.

Such "real world" implementation challenges might leave the strategy's outcomes unachievable, and the information from an impact evaluation would be misleading. Indeed, evaluating a strategy that has not been implemented in a way that maximizes its potential would likely produce results that also understate the strategy's potential value (Fixsen and Blase 1993; Weiss et al. 2014; Institute of Medicine 2001). Findings might suggest that the strategy is not worth pursuing when the appropriate action might be to invest in the strategy so its potential can be realized and then appropriately assessed.

To determine how responsible the strategy actually was for changes in the outcomes, an evaluation must go beyond monitoring the outcomes that follow a strategy (as we will discuss). It must examine a strategy that is implemented with maturity, be able to distinguish a strategy's core activities from other factors that might influence outcomes, and approximate what would have happened if the strategy had not been implemented using data from a comparison group. As a result, rigorous evaluations are typically more resource-intensive than monitoring and often take several years, and because of this, they are usually conducted only periodically.[5] It can be faster and cheaper to determine whether a strategy is not achieving its goals if monitoring data reveal failures in its theory of change early on. Once a strategy achieves success on measures of implementation, it could be worth considering whether it is feasible to evaluate the impact of the program on long-term goals such as compliance.

In this section we discuss the five factors that WHD and other agencies might consider in selecting strategies for and designing monitoring and evaluation to ensure they are successful—that is, that they provide useful information. The factors were developed based on the literature on implementation science and impact evaluation design, including Tatian (2016), Fixsen et al. (2005), Fixsen and Blase (1993), Clearinghouse for Labor Evaluation and Research (n.d.), and What Works Clearinghouse (n.d.), as well as other studies referenced below. The first three factors apply to both monitoring and evaluation, and the last two apply only to evaluation (see Figure ES.1. We discuss each factor in terms of an ideal to help illustrate the trade-offs that must be made when developing monitoring and evaluation in a world that is far from ideal. We continue to use the strategic partnership example introduced above to illustrate the trade-offs, but this should not be interpreted as a comment on what WHD has done or may be doing.

1. Documented and supported core activities

If either monitoring or evaluation are to give WHD an assessment of a strategy's potential, the strategy must have (1) well-defined activities that are viewed as key strategy components and are consistently implemented across employers, and (2) stakeholders who support the monitoring and evaluation. Fixsen et al. (2005, Chapter 4 in particular) discuss the factors at length, and we present them all here in turn.

[5] Because evaluations are conducted only periodically and (typically) on a sample of entities participating in the strategy, contextual factors often make findings difficult to apply to other time periods, geographies, or entities. Extrapolating findings to different kinds of employers might be a particularly important concern for WHD if the evaluation is available only for a select, nonrepresentative set of establishments.

- **Well-documented activities.** A well-articulated TOC will identify the core activities that define the strategy—and who is to deliver them—and these activities must be described with enough detail that they are undertaken consistently in implementation. General descriptions of core activities could cause staff to perform them in different ways or not complete all of them. As a result, not all employers would experience the strategy in the same way, and it could blur what is being evaluated.

  Importantly, however, there is room for variation in implementation, provided that the key features of core activities are defined and implemented consistently. In the strategic partnership example, the TOC defines core activities to be products and services tailored to the brand and relationship building. Other helpful details that could be specified are the types of products (for example, informational materials, worksheets, or tools) and services (for example, analysis of compliance data), how often they are provided, and the mode of interaction (for example, email, phone, or in person). WHD could decide that variation within the informational material or types of analyses, or greater frequency of interaction, are not enough to change the general strategy and can be left to the discretion of WHD leaders working with the brands.

- **Stakeholder support.** Staff who are implementing the strategy—or managing those who are—are often critical to successful monitoring and evaluation. They are the ones who must be relied on to change their procedures if necessary. For example, if investigators were asked to conduct more investigations of the potential partner during the planning phase, this could have put pressure on investigative resources, and staff might have to scale back other work in response. Should staff—or other key stakeholders in implementing a strategy—be unwilling and unable to participate in an evaluation, an evaluation might not be feasible.

2. Measurable outcomes

It is impossible to assess the value of a strategy without examining its outcomes. To see how outcomes might follow or be caused by a strategy, they must be quantifiable and observable during the period of monitoring and evaluation. WHD has long worked to develop measures of violations that relate to compliance outcomes. Other outcomes may be challenging to measure, as we will discuss in the next chapter.

- **Quantifiable.** Outcomes used for monitoring and evaluation are usually expressed in quantifiable terms and should be objective and measurable (numeric values, percentages, scores, and indices). Outcome measures should ideally include a baseline and can include a target or goal (such as a performance standard) so a strategy can be assessed in terms of its contribution to compliance goals, and not just used to track activities or inputs. Measures can thus be used to observe progress and quantify actual results compared to expected results. In our partnership example, we quantified the long-term outcome as a slowdown or decrease in the observed trend in violations over the one-year period before the strategic partnership began. In this example, the trend in violations in the year before the strategic partnership began is the baseline, and a slowdown in violations is the target. Other strategic partnership

outcomes of interest can include brand buy-in and stakeholder awareness, but these are difficult to quantify.

- **Observable.** An evaluator must be able to capture (or observe) outcomes—both before the strategy's activities began (that is, baseline) and after the strategy has been implemented—in order to monitor progress or determine an impact. For outcomes to be observable, enough time must elapse between baseline and the time the expected outcomes would be expected to occur. If, for example, it takes about two years for the activities from a partnership strategy to be put in place and for outputs from those activities to be realized, any monitoring or evaluation at the six-month mark could not observe outcomes. Because the outcome of increased compliance occurs over a fairly long period, WHD could consider structuring monitoring and evaluation based on short-term and intermediate outcomes in the TOC as indicators of progress toward its long-term goal. Information gained from these efforts would be grounded in the assumption that short-term and intermediate outcomes are associated with increased compliance—as the TOC lays out—and this assumption may or may not be valid.

3. Available, appropriate data

The most informative monitoring and evaluation efforts rely on data that capture all the elements in the TOC at the appropriate unit of analysis. This means the data must include measures of:

- **Activities and inputs.** The intensity of activities (often called dosage) or the level of resources put into a strategy can affect outcomes, as the TOC suggests. Having data on dosage and resources could provide information about why a strategy was not effective, how much the outcomes improve as dosage increases, or what amount of dosage or resources are needed to produce desired outcomes. In the strategic partnership example, data on the types of products and services provided, how they were tailored to the brand, how often they were used, and the type of interactions might provide information about how activities are associated with fewer violations. Data could be reported directly by those interacting with the brand representatives.
- **Short-term and intermediate outcomes and outputs.** These measures are crucial to monitoring and evaluation; they can help WHD understand if a strategy is being implemented as intended. If monitoring suggests that a strategy is not increasing compliance, WHD would want to know why. Conversely, if an evaluation finds that a strategy improved compliance or improved it only for some entities or in some circumstances, WHD would probably want to know the conditions it worked best in. In the strategic partnership example, it could be that the employers established internal monitoring processes, and well-structured MOUs reduced violations, as the TOC predicted. Without data on monitoring processes or MOU content, the evaluation would not be able to assess this. Data could be collected from the partner in a variety of ways, including observations, documents, interviews, or surveys; and collected in a module that is linkable with the Wage and Hour Investigative Support and Reporting Database (WHISARD) using brand, employer, and/or establishment identifiers.

- **Long-term outcomes** (that is, compliance measures). To allow for a determination of how they changed with the implementation of the strategy, outcomes should be captured both before and after a strategy is implemented. Using the example of strategic partnerships in Figure II.1, the appropriate outcome measure would be the number of violations of the Fair Labor Standards Act, because the targeted goal for the strategy was to slow the trend of increased violations. However, available WHD administrative data on long-term outcomes are not ideal. WHD does not consistently conduct follow-up investigations for a range of reasons, particularly given limited resources. Moreover, non-random selection of establishments for investigations means that WHD cannot determine based on violation measures whether brand establishments are committing more violations or whether investigators are getting better at identifying those with violations.
- **Context.** A strategy might be more effective in some contexts than others. Having data on contextual factors would allow WHD to identify patterns that could suggest the situations in which the strategy tends to be more and less effective. (For conclusive evidence, however, a rigorous impact analysis would be required.) It may be that strategic partnerships, for example, are effective in noncompetitive industries or when labor markets are tight.

Unless data were collected for the purposes of monitoring and evaluating a specific strategy, appropriate data might not be available. This could apply to WHD's administrative data, as we will discuss in Chapter III, even though WHD has enhanced data reporting and visualization systems and built data modules with new content, among other data capacity building efforts. We summarize reasons why appropriate data may not be available and elaborate on each in the discussion of Factor #3 in Chapter III. First, any one data source is unlikely to be contain all the information identified in the TOC. For the strategic partnership example, administrative data may not contain all the detailed, strategy-specific information needed to confidently assess the outcomes following a strategy's implementation, and external data are unlikely to contain measures of components of the TOC. Under such circumstances, a survey or other primary data collection method specifically designed with the TOC for strategic partnerships in mind could dramatically enhance the information available on the strategy's activities, outputs, and outcomes. Second, data quality could potentially be strengthened by enhanced quality assurance procedures. Third, WHD's administrative data lack some appropriate information because they cannot provide measures of how prevalent violations are. They have only limited information on the characteristics of establishments, and the process of selecting employers for investigation also limits the appropriateness of the information.

4. Implementation maturity

Evaluation methods can be used at various stages of a strategy's implementation. However, to interpret the findings, it is critical to understand what is being evaluated—in other words, what was implemented. Formative evaluation, defined as research about how a program is designed or carried out with the goal of improving implementation and results, would be appropriate at early stages to address questions about effective approaches to implementation (Tatian 2016). When a strategy is first put in place, local and regional staff might resist change, especially if incentives

are not aligned with the desired implementation. Attempts to implement new practices can sometimes end at this point, overwhelmed by the demands on staff procedures and management (Macallair and Males 2004). Formative evaluation can help researchers understand what practices are being implemented and what factors support or hinder their use. However, evaluating the impact of the strategy during this period would not produce an accurate assessment of its full potential.

An impact evaluation should be conducted after the implementation of the strategy is "mature," meaning that its inputs and activities are integrated into infrastructures and supported as part of a regular routine. A mature strategy has local and regional buy-in, with the new learning integrated into organizational practices, policies, and procedures. The strategy is fully operational with full staffing complements, full client loads, and all the realities of the "doing business" part of WHD policies and procedures. Procedures are routinized, staff implementing the strategy are proficient and skilled, and managers and administrators support and facilitate the strategy. In sum, it has become "accepted practice," and a new operationalization of "business as usual" has evolved (Bertram et al. 2011), with outputs and outcomes realized. With a mature strategy, it is clear what is being evaluated, and the strategy has its best opportunity to reveal impacts.

Importantly, maturity does not necessarily require a long time to achieve or perfect consistency in delivery across locations, especially if the strategy is relatively simple and narrowly focused. For example, a compliance assistance strategy that involves emailing information to a list of contacts can mature rapidly. A mature strategy is not necessarily implemented in the same manner in each location because the context for implementation across locations—like District Offices or states—will differ. Some adaptations to local conditions will be desirable and become part of the "standard model" or agreed-upon articulation of the strategy (Winter and Szulanski 2001). Other adaptions will be undesirable, as they create drift from the envisioned strategy (Mowbray et al. 2003). When attempting to discriminate between drift and adaptation to local context, WHD might consider first implementing the strategy in one place as envisioned, then adapting it to other local areas. With this process, the adaptation is likely to be consistent with the TOC, which can make the adaptation more successful than it would be if the strategy were modified before it was implemented with fidelity to the vision (Winter and Szulanski 2001).[6] Another option for discriminating between drift and adaptation is for WHD to identify key features of implementation activities that define the strategy, features without which it would be something different. That can reveal where there is room for local adaptation.

5. Internal validity

If an evaluation has internal validity, it can produce evidence on whether the strategy caused the outcomes. To do this, the evaluation must be able to (1) distinguish a strategy's core activities from other factors that might influence outcomes, which are sometimes called confounding factors (MacKinnon et al. 2012; What Works Clearinghouse n.d.); and (2) carefully construct a

[6] Of course, at some point alterations might be large enough that the strategy must be redefined. Under such circumstances, monitoring and evaluation must be tailored to the "new" strategy.

counterfactual condition; that is, an approximation of what would have happened if the strategy had not been implemented.

- **No confounding factors.** If outcome-influencing factors other than the strategy cannot be disentangled from the strategy's activities, an evaluation will not be able to reveal whether the strategy, the (confounding) factor, or both have caused the outcome. This can lead to bias in estimates of impacts. In our strategic partnership example, if WHD increased its use of enforcement strategies when developing strategic partnerships, it would be impossible to tell which strategy was associated with changed behavior or increased compliance. Similarly, if strategic partnerships were only established with two different brands, and both those brands had similar emerging internal cultures and ethos, it would be impossible to determine whether it is the emerging culture and ethos or partnership that is associated with increased compliance. Confounds can be particularly problematic for strategic partnerships (and other strategies) that have only been used a few times because it could be more likely that the partners share certain characteristics that make them more or less likely to comply. Confounds cannot be adjusted for in analysis; researchers must either identify other instances of a strategy that are not affected by a confound or else acknowledge that evaluation findings do not provide as strong evidence as if there were no confound (What Works Clearinghouse n.d.).
- **Counterfactual.** A comparison condition or group (for example, of similar brands) that has not been exposed to the strategy is one way to capture the counterfactual. There are two broad types of research designs that would allow WHD to ascertain that the strategy being evaluated caused a change in compliance:
    1. **A randomized controlled trial (RCT)**—also called an experiment—uses randomization to determine which entities would be involved in the strategy (and make up the treatment group) and which would not (and make up the control group). Because entities are randomly assigned to either the treatment group or the control group, the members of these two groups will, on average, have similar observed and unobserved characteristics before receipt of the strategy. Because nothing else about the two groups should be different except exposure to the strategy, comparing outcomes after the strategy has been implemented for the treatment group should provide an unbiased assessment of the strategy's impacts. The RCT is the most scientifically rigorous method of testing available and is regarded as the gold standard for evaluations.
    2. **A quasi-experimental design (QED)** uses a method other than random assignment to form a comparison condition or group. The strongest QED studies select the treatment and comparison groups in a way that makes them as similar to each other as possible at baseline, before the evaluation begins. The key to ensuring that the QED design can produce evidence that the strategy caused increased compliance is baseline equivalence, which refers to a lack of difference in characteristics between members in the treatment and comparison groups before the treatment group engages with the strategy. It is a major

concern for QEDs because any initial dissimilarities—and not the program—could be the underlying reason for observed differences in outcomes.

Of note, an evaluation design that can provide causal evidence may pose logistical challenges (Heard et al. 2017). For an RCT, implementing randomization can be difficult in the context of limited resources and performance measures that provide incentives to find violations; it can also be challenging to ensure that entities assigned to the comparison group (and therefore ineligible for the strategy) do not somehow receive the strategy or become influenced by it. For a QED, the data requirements are greater because information must capture differences between the treatment and comparison groups before the strategy is implemented as well as after.[7] For both designs, all elements of the TOC must be captured with data for members of the treatment and comparison groups. These challenges mean that internal validity depends both on how well the evaluation's design constructs the counterfactual and how successfully that study design is carried out. As WHD has found in many past studies of compliance, there may be trade-offs between rigor and flexibility and resources on the ground. Thus, when considering an evaluation of a strategy and assessing the potential for internal validity, WHD could continue to consider not only the rigor of the evaluation design but also whether it can be implemented well.

In the strategic partnership example, it is difficult to identify a counterfactual condition. WHD cannot randomly assign brands to a partnership to conduct an RCT because a partner must agree to engage. Moreover, strategic partnerships are too resource-intensive to have large numbers of partner brands who could make up the treatment and control groups. QEDs may be challenging as well, because the unobserved factors that motivate the brand to engage in the partnership make it difficult to identify a group of similar brands to represent the counterfactual. One feasible option could be to consider the counterfactual condition of the brand before the partnership (as in an interrupted time series design, described in Chapter III), and compare trends in brand compliance before and after the partnership engagement begins. To be rigorous, such an evaluation would need to have many observations of the outcomes for several strategic partnerships that were initiated at predetermined times.

[7] The exception is regression discontinuity designs, discussed in Chapter III.

This page has been left blank for double-sided copying.

III. CONSIDERATIONS IN ASSESSING THE EVALUABILITY OF A STRATEGY

To yield relevant, precise, and actionable information on a strategy's effectiveness in changing outcomes, it is important for agencies such as WHD to not only build sound monitoring and evaluation processes, but also to select the most suitable strategy. We refer to the suitability of a strategy for monitoring and evaluation as its "evaluability" (that is, the strategy's potential to produce useful information about its outcomes and effectiveness through monitoring and evaluation).

In assessing a strategy's evaluability, agencies such as WHD could consider the five factors described in Chapter II. This applies whether the agency is seeking to retrospectively evaluate a strategy that has already been implemented, or is prospectively developing a plan to implement and evaluate a strategy. As the discussion will show, a prospective approach that considers these five factors in building a monitoring and evaluation design before implementing a strategy increases the agency's options for conducting successful, rigorous monitoring and evaluation that can yield useful insights. By planning ahead, the agency might be able to improve or strengthen the conditions that support this success, such as gathering or enhancing documentation and data and constructing a counterfactual condition.[8]

In this chapter, we use strategic partnerships and other examples to illustrate how WHD can consider these factors when assessing the evaluability of a strategy for monitoring and evaluation. These examples are intended to be illustrative and should not be interpreted as a comment on what WHD has done or may be doing. We discuss the circumstances for each factor that are ideal for evaluation and some of the challenges and tradeoffs that WHD might face as it tries to achieve that ideal. We focus on prospective evaluations, but also make note of potential retrospective evaluations. Chapter IV has more details on the opportunities and challenges that accompany monitoring and evaluation of strategies.

A. Factor #1: Documented and supported core activities

To provide useful information, a strategy's activities must be well documented and supported. The documents that describe these components and activities can be collected and enhanced to support several critical roles they play in monitoring and evaluation:

- **Informing evaluation design**. The TOC articulates how WHD expects a strategy to unfold. Documentation for core components in the TOC can therefore help guide the identification

[8] Evaluability is another possible factor to consider when designing a strategy, and when selecting a strategy to use in a given situation. However, there are many other factors to consider, such as the strategy's feasibility, resource needs, alignment with WHD's mission and current agency priorities, and the potential for doing the most good. Evaluability is probably not the deciding factor. This report does not speak to those decisions because it focuses on how to monitor and evaluate strategies instead of how to create and use them.

and prioritization of research questions, selection of outcome measures, identification of data needs, and selection of data sources or design of data collection tools.

- **Providing a benchmark against which to assess fidelity and attribution for the impact.** If the key activities of core components are not carefully delineated in written documents (for example, procedure manuals and analysis plans), it is difficult to determine whether the activities are being conducted in a way that is consistent with the TOC (that is, with fidelity). Without this documentation, there can be major variations in the way District Offices select and approach state agencies, which might affect their rates of success in obtaining state data, their selection of outcome measures and their methods of constructing such measures, their analysis methods, and how they use the findings from any analysis of these data.

  With major inconsistencies across District Offices (that is, low fidelity to the TOC), information from an evaluation is not likely to be informative and will likely understate the potential value of the strategy (Fixsen and Blase 1993; Weiss et al. 2014; Institute of Medicine 2001). WHD District and Regional Offices often have a degree of flexibility, autonomy, and discretion in their compliance activities that can enable them to implement strategies smoothly. Consistency in implementation does not preclude such flexibility; it merely requires that the core content of the key components and any adaptions in activities that are part of them are consistent with the TOC.

Alignment with WHD performance standards

Any evaluation must consider the performance standards and targets that WHD is held accountable to. These can change over time (GAO 2008), and they influence the extent to which an evaluation can be supported. Research could be most feasible when the goals of the strategy are aligned with performance goals, because it is more likely that resources such as staff time and careful documentation will be in place when the strategy being evaluated is a priority for the agency. For example, if WHD's current priority is to increase the amount of back wages that are collected, then an evaluation is likely to be well supported if the strategy being examined focuses on increasing the collection of back wages, instead of being, for example, one that aims to increase workers' awareness of the Fair Labor Standards Act.

- **Providing context.** Documenting activities can sometimes provide contextual information by describing the observed need that motivated the development of these activities to begin with; this can be useful for interpreting the findings of an evaluation and placing them in the relevant context. An evaluation would (ideally) build on monitoring efforts with a process study. Interviews with field staff would be part of such a study, and they would help WHD understand how a strategy was implemented in practice, assess fidelity to the original design of the strategy, and identify factors that facilitated or impeded implementation. Interviews with key personnel associated with the employers that received the strategy could elucidate their experiences with implementation, especially for voluntary compliance assistance strategies. For example, when the strategy involves leveraging other federal government

agencies, the evaluation team might want to talk with key staff in the field offices and state agencies and request documents and data from them.

Given the need for such documentation, it is important to have the availability and buy-in of key stakeholders to monitor and evaluate a strategy. Key stakeholders may include external entities, such as partners, community leaders, agency directors, supervisors, industry associations, practitioners, advocates, and policymakers, along with WHD staff at all levels (local, regional, and national) who are designing, implementing, and managing the strategy. Ideally, during the planning process, these stakeholders will provide feedback on all aspects of the monitoring or evaluation effort, including the significance, reach, and evaluability of a strategy; the suitability of selected outcomes and their alignment with the logic model for the strategy; and the suitability of any applicable comparison groups and their similarity to the group that received the strategy.

B. Factor #2: Measurable outcomes

The availability of quantifiable and observable measures of outcomes is critical to whether a strategy can be monitored or evaluated to an informative end. Several types of compliance measures fit these needs. The most common are those centered on the incidence of violations, severity of violations, nature of violations (for example, the type of law that is violated or the recurring nature of violations), and complaint measures. For example, a strategy that aims to reduce the prevalence of violations of minimum wage laws is suitable for monitoring and evaluation, because this outcome is relatively easy to quantify and convert into empirical measures—such as the number of violations per 1,000 workers in a given year.

Although some outcomes might be relatively easy to measure, WHD could also consider whether they can be captured in the time frame of the monitoring and evaluation. Some strategies such as outreach on compliance and tools facilitating compliance for employers might affect measurable outcomes immediately, whereas other strategies like strategic partnerships could take years of implementation before the measures for the long-term outcomes outlined in the TOC can become available.[9] For the latter, in the meantime, WHD could monitor short-term or intermediate outcomes that are associated with long-term outcomes. For example, strategic partners might establish compliance procedures or staff training within a year after implementation. In such situations, WHD could decide which outcomes of a strategy it is most interested in, and whether monitoring and evaluation that can only capture the partial effects of a strategy are worthwhile.

Some strategies may target outcomes that are difficult to quantify, however, which make them ill-suited for monitoring and evaluation. Outcomes might be difficult to measure because they

[9] The length of time it takes for long-term outcomes to occur depends on the strategy. Some strategies might achieve the long-term outcomes specified in the TOC quickly. For example, a strategy that targets a specific establishment for an employment relationship might result in swift changes in compliance outcomes for that establishment. In comparison, a strategy that involves general outreach and education to a wider swath of employers could take longer to produce substantial reductions in compliance violations; such a strategy is not targeted to specific entities, and it can take time for general awareness and knowledge to spread across employers and translate into changes in actions.

are subjective (for example, employers' attitudes toward compliance assistance) or multidimensional (for example, employees' working conditions or the quality of relationships between WHD staff and the business community). These measurement challenges might be overcome, but solutions are often complicated, and the findings can be more susceptible to measurement error or bias. For example, WHD could measure employer attitudes with a carefully designed survey administered to a large, representative sample of employers, but such an effort would be resource-intensive to design and field, and the resulting data can suffer from several biases. For example, employers might respond to survey questions so as to portray themselves in the most positive light, or employers with the most positive attitudes about compliance assistance could be most likely to respond to the survey. An alternative option might be to focus on related quantifiable measures such the number of mentions of compliance in industry trade publications or web forums which could potentially be gathered from the internet using web scraping methods (automated processes of extracting data from websites; see Hoynes et al. 2011 for an example).

C. Factor #3: Available, appropriate data

A strategy is suitable for monitoring and evaluation only if the necessary data to capture inputs, outputs, and outcomes are available in an appropriate form. When assessing whether a strategy is evaluable, WHD could consider whether relevant and high quality data would be available or could be made available. Here, we discuss what might constitute appropriate data and then discuss their actual and potential availability for monitoring and evaluation of a strategy. More broadly, the discussion below also highlights how data collection could improve the evaluability of strategies.

1. Appropriate

There are several aspects of appropriate data for monitoring and evaluation: they are comprehensive, at the appropriate unit of observation, high quality, capture appropriate outcomes, and contain appropriate supplementary information.

a. Comprehensive

Fully comprehensive data contain measures of all components of the TOC and support informative monitoring and evaluation. Unless data were collected specifically for monitoring or evaluation, however, they are unlikely to contain all the relevant information. It may be possible to link multiple data sources using a common identifier to create a more complete set of information. For example, WHD might consider linking establishment-level data on outcomes with MSA or county-level data on unemployment to account for local economic context. Identifying information such as establishment addresses can facilitate this linkage (as illustrated by Patnaik [2020]). As another example, WHD might consider linking a database on outcomes to another database to find out whether a brand received a strategy. For example, a roster of brands that participated in strategic partnerships with WHD could be used to find those brands in a database containing the establishment characteristics that were used to define targeting criteria

for the strategy. Comprehensiveness of data might often depend on data being linkable. Therefore, in assessing the evaluability of a strategy, WHD could consider whether multiple data sources might need to be used, and whether the information needed to link them exists.

b. Appropriate unit of observation

For a strategy to be evaluable, there must be data available at the unit of observation the strategy is focused on—often establishments, firms, or brands. Use of higher-level data can produce misleading results. For example, if the goal is to determine whether compliance at the employer level changed following implementation of a partnership with the employer, WHD may not want to examine whether average wages of workers change in all establishments operating in a county where partnerships are pervasive, because other contextual factors in the county (such as unemployment levels) may have contributed to those outcomes. Granular, or lower-level data, can often be aggregated up to a higher level, but higher-level data generally cannot be disaggregated to more granular units. For example, if a strategy is targeted at franchised establishments, data should be at that level. Using data at the brand level (for example) is not sufficiently granular and, as a result, could mask outcomes that exist at the establishment level. Using data at the establishment level would allow an assessment of whether targeted establishments had reduced compliance and, by aggregating establishments to the brand level, allow for a comparison of all employers under one brand with all employers under another.

c. High quality

For the most accurate findings, data should be of high quality. Quality is reflected in a number of dimensions. Ideal data should be reliable and reflect the intended concept. They should be complete, so data fields do not have missing values. Data should be free of data entry errors, such as illogical values or typographical errors. Detailed documentation should describe how data were generated and processed. While ideal data is the goal, it may not be possible to attain. Robust data quality assurance processes can help ensure that data quality is as high as possible. Appropriate data analysis methods can be used to address issues such as missing data (see Deke and Puma 2013).

d. Appropriate outcomes

Critically, a strategy is only evaluable if there are data that will allow measurement of expected outcomes. Next, we describe three key types of compliance outcomes that are of particular importance and pose particular challenges:

- **Prevalence of an outcome**. Prevalence refers to the proportion of a population that has a certain outcome within a specified time. For example, prevalence of Fair Labor Standards Act (FLSA) violations in an industry could be captured by the number of establishments in the industry that violate the law divided by the total number of establishments in that industry. Examining this outcome can enhance the generalizability of findings, improving external validity. Monitoring and evaluating this outcome would require data on not only the

number of establishments that have violated the FLSA, but also the total number of establishments in the industry (the population). In the absence of data on the population, monitoring and evaluation could use data on a large, representative sample of entities to measure prevalence.

WHISARD is the most accurate source of information on violations, but it does not currently provide prevalence measures because it does not include a population or random sample of establishments. Only establishments that have been investigated by WHD, either because of a WHD-directed strategy or initiative or because of a complaint filed against the establishment, are in WHISARD. Further, WHISARD data suffer from selection bias because the establishments with worse compliance outcomes than average are more likely to be recorded in WHISARD. This means that violation measures among a group of establishments in WHISARD may not necessarily reflect the actual prevalence of violations; they could also reflect investigators' skill in identifying establishments in violation. The ideal data would be collected from a sizable, representative sample of businesses or workers so there would be a large enough sample not only to be able to detect impacts (called statistical power) but also to provide accurate measures of the prevalence or distribution of outcomes. As discussed in Patnaik (2020), external data could potentially enable WHD to create a sampling frame for future monitoring and evaluation efforts; that is, a master list of all entities from which to draw when deciding upon entities to investigate. This would allow random sampling; in other words, WHD could choose to investigate a small subset of all the entities that could be investigated but choose them in a manner such that each entity in the sampling frame has an equal probability of being chosen for investigation.

- **Incidence of an outcome.** Incidence refers to the rate of occurrences of new cases within a specified time frame. For example, if a strategy involved running regular radio ads that encouraged people to report FLSA violations, a relevant outcome could be the number of worker complaints that WHD receives per month. To measure the incidence of an outcome, there would have to be data on the population of entities or, at a minimum, a large, representative sample of entities. For example, it may not be possible to monitor the incidence of complaints that are received by WHD because only complaints that pass an initial screening are entered into WHISARD.[10] If data on the full universe of complaints (or a representative sample) are unavailable, a robust investigation of whether a strategy changes the incidence of valid complaints would be difficult. Instead, a feasible investigation would need to focus on a research question about an alternative measure, such as whether a strategy changes the incidence of valid complaints tracked by WHD.

[10] Complaint-based investigations entail a screening process, so not all complaints are entered into WHISARD. If complaints are not under WHD's purview, they are not pursued. If the complaint could relate to a violation of the laws and regulations that WHD enforces, the complaint is assigned a complaint identifier and entered into WHISARD. A decision is then made about whether to create a case. Some complaints (for example, those isolated to a single type of violation for a single employee) are conciliated. Others lead to an investigation covering all employees in the establishment.

- **Spillover effects**. A strategy might change outcomes outside the entities that were directly involved with it, producing spillover effects. For example, a strategy that publicizes violations of a specific firm through press releases is likely to increase not only compliance in that firm but also the compliance of other firms in the same industry or county.[11] To examine these effects, data must include both the targeted entities and those entities that might experience spillover effects. Accordingly, if a strategy is expected to produce spillover effects, accurately estimating its full effects on compliance would require data on the outcomes for the targeted firm (for example, using data on the firm's violation rates and severity of violations) and for the industry or the county as a whole before and after strategies were adopted in order to capture both the direct and spillover effects associated with the strategy.
- **Change in outcomes:** Because effective monitoring involves measuring changes in outcomes over time, it requires data before and after a strategy is implemented. For example, as part of WHD's YouthRules! Initiative, in 2005, the division launched a nationwide outreach campaign to increase awareness about youth employment laws in construction (U.S. Department of Labor 2005). To monitor and evaluate this campaign, a study examined the change in compliance over time by comparing subsequent violations among employers who had been previously investigated (Eastern Research Group 2009). To understand changes in specific entities' outcomes over time, data would need to contain information on the same entities before and after a strategy is implemented (panel data). As another example, WHD has negotiated enhanced compliance agreements (ECAs) in cases that required litigation by the solicitor. The ECAs typically include requirements for the investigated entity to take actions, such as establishing new positions to oversee compliance, training management personnel, and providing means for workers to lodge complaints internally on a confidential basis. To monitor or evaluate these ECAs would require panel data that contained outcomes of the same firms or individuals over multiple time periods. Such panel data would capture the extent of recidivism (that is, the rate at which entities that committed violations in the past run afoul of the law again). As described later in Section E, panel data can also provide internal validity for evaluation.

e. Appropriate supplementary information

When assessing the evaluability of a strategy, WHD could consider whether other data are available that may not be strictly necessary but could substantially enhance the value of monitoring and evaluation activities in several ways. Such supplementary data can:

- **Enable rigorous evaluation designs.** Data on factors other than outcomes can facilitate more rigorous evaluation designs that can provide more accurate and credible findings (see section on internal validity). As just one example, data capturing descriptive characteristics

[11] Johnson (2018) studied a targeted disclosure policy in which the Occupational Safety and Health Administration issued press releases about facilities that were assessed penalties above a certain threshold for safety and health violations; it found that publicizing the violations of one facility led to improved compliance and fewer occupational injuries among nearby facilities in the same sector.

of establishments (such as the number of employees, annual revenue, age, and location) could enable WHD to identify a credible comparison group of establishments that are similar in those characteristics to the establishments that were involved in a strategy.

- **Provide answers to a broader set of questions.** Monitoring and evaluation are designed to answer the question, "Does the strategy improve compliance?" However, additional data can enable WHD to answer other important questions, such as, "Is the strategy more effective for certain types of entities than others?" For example, data on the ownership characteristics of franchised establishments could enable a subgroup analysis that examines whether the strategy is more or less effective at improving compliance among establishments whose owners also own other franchised establishments. The answers to such questions can help WHD focus its strategies on entities that could most benefit from them.
- **Provide context.** Additional data can provide important context for a study and help us understand the effectiveness of a strategy.

Example of context: Industry structure and business models

Research and WHD's experience indicate that industry structures and business models create incentives and opportunities for employers that influence compliance (Weil 1996, 2005, 2008, 2009, 2010, 2012, 2014; Ji and Weil 2009, 2015; Weil and Mallo 2007; Weil and Pyles 2005). For example, the use of subcontracting within an industry or company makes it more likely that workers work as independent contractors at piece rate with few labor law protections (or none). A wider understanding of the potential effectiveness of a strategy might require an understanding of a given industry structure and business model. Therefore, even if monitoring and evaluation do not hinge on having data on industry structure and business models, these data could still significantly improve the value of these efforts by placing the findings in context.

2. Availability: Potential data sources

WHD has or could acquire a variety of data that can potentially be used for monitoring and evaluation, including its own administrative data; Patnaik (2020) includes a detailed discussion of such data and their uses. In assessing the suitability of a strategy for monitoring and evaluation, WHD could consider the types of data that are available or could be made available. The upcoming discussion highlights how data collection efforts (that may not be specific to a given strategy) can improve the evaluability of a strategy.

a. WHD administrative data

WHD's case management system, WHISARD, tracks the assignment, investigation, management, resolution and closing of investigations and records case history, including the findings of any violations found by the investigators and any penalties assessed. Unique strengths of WHISARD data are that they provide direct measures of the incidence and number of violations found, covering a wide range of violations reported by WHD's investigative staff; track results from investigations; and provide context about the investigations and their

outcomes. However, WHISARD also has some notable limitations for the purposes of an evaluation.

- First, it includes only some establishments, and those establishments are not selected at random. This means that analyses relying on WHISARD data may suffer from selection bias because the sample of entities being analyzed will not be representative of the broader target population of entities for which we wish to measure the effect of a strategy.

- Second, WHISARD contains fairly small samples relative to the full population of establishments, which may limit the potential statistical power of some evaluations. For example, in 2016 the Bureau of Labor Statistics reported 231,632 establishments operating in the "limited service restaurant" industry, but WHISARD contained only 554 cases of limited service restaurants that were investigated in the same year.[12]

- Third, WHISARD contains limited information about the characteristics of the establishments that are investigated, and WHISARD and other WHD administrative data also contain limited information about the inputs, activities, and outputs of strategies. This lack of descriptive information may constrain the types of questions that the data can answer and limit the analytic methods that can be used.

- Finally, WHISARD has limited follow-up observations of given entities at multiple points in time. Data on the same entity both before the strategy's implementation and at multiple time points afterward are rarely available, and thus do not consistently capture long-term outcomes or facilitate comparisons of outcomes before and after a strategy.

Some of WHISARD's shortcomings as a monitoring and evaluation tool could be addressed by collecting or identifying additional data. Those data could come from primary data collection or from sources housed outside WHD (hereafter referred to as external data). We discuss these in turn.

b. Primary data collected by WHD

Information related to a strategy's implementation could potentially be collected by WHD (if it is not already). These data could potentially fill gaps in several TOC components, including activities, outputs, and intermediate outcomes. They could potentially be collected through a variety of means, including observations, document collection, interviews, or surveys. To take the example of strategic partnerships, WHD might consider gathering information on activities engaged in during implementation, such as dates and topics of meetings, which data were shared with the partner and when, which compliance tools were shared and when, whether phone or email check-ins were conducted—all of which could be observed by WHD staff. It might also

[12] The Quarterly Census of Employment and Wages calculates an annual average of 231,632 privately owned establishments falling under the North American Industry Classification System (NAICS) code 722513 (U.S. Bureau of Labor Statistics 2019).

consider collecting information on outputs, such as whether a partner used the compliance tools WHD provided and whether a partner put in place specified compliance-related procedures. WHD staff could request this information from the partner brand headquarters or franchisees, or ask headquarters staff about it in an interview. Finally, WHD might consider gathering information on intermediate outcomes such as partner attitudes and perceptions about the usefulness the compliance tools; this information could be collected through online surveys or interviews with headquarters and/or franchisees. To plan for these efforts, WHD could consider thinking carefully about the measures it would like to collect, with reference to the TOC; developing data collection protocols and procedures; piloting and revising the protocols; creating a database or data modules in WHISARD; and creating a data entry system, including instructions, procedures, and guidance on entering data.

c. External data

When assessing the evaluability of a strategy, WHD can explore whether relevant national- or state-level databases can add value to monitoring and evaluation by providing data that capture various elements of the TOC for that strategy. These data could potentially be linked to WHISARD by establishment, industry, or geographic area (for example). Dolfin et al. (2018) explores a variety of external data that may be relevant. An accompanying brief to this report, "Data for Monitoring and Evaluation of WHD's Compliance Strategies," explores how external data can be integrated with WHISARD to meet the data needs of monitoring and evaluation (Patnaik 2020). We summarize some key findings here. Broadly, external data could provide the following:

- **Population of employers**. Databases such as CHDExpert or Dun & Bradstreet provide a census of a population of employers and could be linked to WHISARD using establishment name and address.
- **Characteristics of business entities**. For example, CHDExpert is an organization that collects, tracks, and analyzes data in the food service industry, including the characteristics of restaurants (such as years in business, average check size, number of employees, and annual sales) and characteristics of the area they are located in (such as restaurant density and average household income). Such data could be linked to WHISARD by industry or geographic area.
- **Contextual information on local economic, industry, and population conditions.** Data about the industry or business model can describe the business landscape. For example, a proprietary data set, Construction Market Data, has information on the construction market, such as units and value; an expansion index on whether a location's construction volume is expected to expand or shrink in the upcoming 12 months; and information on upcoming bids for projects at the federal, state, and local levels. External data can also describe the prevailing economic or social conditions that exist when a strategy is implemented. For example, the Quarterly Census of Employment and Wages tracks the number of employees, number of establishments, total wages, average weekly wage, and average annual pay in each

industry segment and county in the nation; these data provide important, nuanced information about the employment dynamics of an establishment's local context. Contextual data could be linked to WHISARD by industry or geographic area.

Such descriptive and contextual data could support monitoring and evaluation by enabling WHD to do the following:

1. Plan future investigations based on a universe of data. External data could enable WHD to create a sampling frame from which to select a random subset of establishment for investigation, from which it could estimate measures of violation prevalence, as the agency has done in previous work. Alternatively, WHD could use such external data prospectively by attempting to match an establishment to external data before investigating it, enabling data validation.

2. Assess the extent of WHISARD's selection bias among establishments that have been investigated, which could inform WHD's thinking about whether and how it may be possible to design monitoring and evaluation activities to produce thoughtful evidence on the effectiveness of a strategy.

3. Identify entities similar to those that received the strategy in order to create comparison groups for an evaluation design with internal validity.

4. Account for differences in the characteristics of entities that did and did not receive a strategy using a statistical method called covariate adjustment that isolates how much of the difference in outcomes between the two groups can be attributed to the strategy rather than to differences in characteristics.

5. Account for differences in implementation and results related to local economic, industry, or population context.

6. Examine whether a strategy was more effective for some subgroups of entities than others.

In sum, external data can offer information that correlates with industry behavior, describes relevant industry and economic trends, signals changes in compliance, and helps in interpreting and understanding findings. When combined with WHISARD or other WHD administrative data, they can potentially be used to address research questions of interest, including the effectiveness of WHD efforts for certain subgroups and risk factors associated with violations. When assessing how suitable a strategy is for monitoring and evaluation, WHD could consider whether WHISARD data are sufficient or whether other data are available that could complement WHISARD data and enrich the findings. At present, however, the low match rates found in an exercise linking external data to WHISARD highlight a potential challenge in using such data (Patnaik 2020), so WHD might consider acquiring small samples of external data and

testing the feasibility of linking processes while considering whether and how to monitor and evaluate a strategy.

D. Factor #4: Implementation maturity

There is an inherent tension in determining how mature a strategy must be before it should be evaluated. On the one hand, it is desirable to wait until the strategy is implemented as intended before evaluating it. On the other hand, it is likely that WHD would want some credible evidence of the effectiveness of a strategy before making it a part of "business as usual" and investing considerable resources in it. WHD might have to balance these two competing needs to choose the right timing for the evaluation. To obtain credible evidence, WHD could look at results of monitoring the strategy.

Changes in agency policy, priorities, and staffing can pose a challenge to a strategy developing and maintaining maturity. For example, if WHD asks Community Outreach Resource and Planning Specialists (CORPS) to prioritize certain compliance assistance activities, reach out to new stakeholder populations, or aim for more frequent employer outreach, this could change the way a given compliance assistance strategy is implemented. Changes in agency performance measures of, for example, efficiency or targeting success, or in goals for District Offices can affect implementation activities indirectly by changing the incentives that District Offices face. Finally, changes in specific staff or in staff level of effort in implementation can lead to activities being carried out differently or to a different extent. In each example, the changes in implementation can increase the time needed for a strategy to reach maturity; if the changes are extensive, they could turn a mature strategy into a new strategy. When considering whether a strategy is mature enough for a potentially lengthy evaluation process, WHD might consider whether any upcoming agency changes of this kind could disrupt the evaluation, and whether it could be worthwhile to hasten or delay the evaluation.

It is important to note that the consideration of maturity for evaluation does not preclude formative evaluation and pilot testing for strategies that are not yet mature. This is an important part of strategy development, a full treatment of which is outside the scope of this report. Formative evaluation can address questions about effective approaches to implementation by investigating how a program is designed or carried out. Pilot tests, or evaluations conducted in a small number of locations or for a small sample of employers as a precursor to a larger study, can offer a useful way to refine a developing strategy by testing whether changes to certain components of the strategy have the potential to be effective. For example, if WHD wishes to identify an effective format for employer invitations to compliance assistance seminars, they could consider developing several formats, identifying a relatively small random sample of employers, conducting an RCT by randomly assigning employers to receive each format, and comparing the responses of employers in each group. The geographic variation in implementation of WHD's strategies due to innovations and adaptations to local conditions offers rich opportunities for such testing.

E. Factor #5: Internal validity

In Chapter II, we discussed how an evaluation must have internal validity for WHD to be able to directly attribute the change in outcomes to the strategy. Accordingly, a strategy is suitable for evaluation only if it has been implemented in a way that makes it feasible to execute a rigorous experimental or quasi-experimental evaluation design. Next, we discuss two requirements for internal validity: the strategy must have no confounding factors, and it must be implemented in a way that provides a reasonable understanding of the counterfactual.

1. Confounding factors

A confounding factor refers to something observable or unobservable that influences both the outcome and the strategy. Confounding factors cause an association between the outcome and strategy that can be mistaken for a true effect of the strategy on the outcome. For example, if a strategy was rolled out in counties in states with no state minimum wage, an evaluation design could not determine the effects of the strategy by comparing compliance outcomes across the counties that were and were not targeted: some of the differences in outcomes could be caused by existing state minimum wage laws. The following are examples of confounding factors:

- **Bundling of strategies.** Bundling refers to the practice of combining strategies and activities into a single strategy that is delivered to a targeted entity. With bundling, it is difficult if not impossible to isolate the effect of a single strategy because the receipt of any one strategy also includes receipt of other bundled components. Because we never observe an instance in which only the strategy to be evaluated has been implemented, it is impossible to distinguish the effects of the strategy from the effects of other bundle components. This concern of bundling is especially relevant for WHD because it takes a multipronged approach to compliance, so it may rarely use one strategy in isolation.[13]
- **Matching variation in other relevant policy.** Ideally a strategy would be targeted and implemented in a way that enables an evaluation to account for federal and state laws. For example, if a strategy is targeted to different geographical regions, but the variation in the strategy implementation is mirrored closely by variation in states' laws and enforcement practices, it becomes difficult to attribute differences in regions' outcomes to the effects of the strategy and not to those practices. Importantly, changes to applicable federal and state laws over time also need to be accounted for. For example, under a new federal rule, on January 1, 2020, most salaried workers who earn less than $684 per week ($35,568 per year) became eligible for time-and-a-half overtime pay for any hours they work beyond 40 hours a week (U.S. Department of Labor 2019). This was a significant increase from the prior threshold of $455 per week ($23,660 per year), and it was estimated that nearly 3.5 million workers would be impacted in some way by the new rule. As it may take some time for awareness and understanding of any rule change to spread across all employers (and

[13] When a combination of different interventions is deployed into very different contexts, a factorial or multi-arm RCT could test components of well-understood interventions, implemented as intended, where they are expected to have an impact. (Deutsch et al. [2019] provide an example.)

workers) and for them to adjust their behavior accordingly, any evaluation encompassing data from this period may need to account for this adjustment period.

- **Changes in employer business culture.** For strategies in which employers choose to participate, such as by entering into strategic partnerships or using an online compliance tool, concurrent changes in business culture could confound the strategy. An emerging culture that emphasizes compliance and employee rights could lead the employer to engage in the strategy, so that an evaluation of the strategy could confound the effects of the strategy with the effects of the cultural change.
- **Concurrent shocks.** A strategy may be confounded by sudden events at the brand or national level. For example, the onset of a national recession can affect all employers, including both those who receive a strategy and those who do not, muddling the findings from an evaluation focused on changes in outcomes over time. Brand-level shocks—such as union actions or corporate policy changes (for example, Target's voluntary commitment to a higher minimum wage than required by state or federal law)—can undermine an evaluation if the strategy targeted a small number of entities.

2. Counterfactual

To determine the extent to which a strategy caused a change in outcomes, we need to understand the counterfactual—that is, what would have happened in the absence of the strategy. Unfortunately, for any given strategy, the true counterfactual cannot be known, because it is not possible to observe the same entity under the two alternate scenarios (with the strategy and without it). However, a well-designed evaluation can "mimic" or approximate information on the counterfactual if the implementation of the strategy meets two criteria. First, there must be a comparison group—an identifiable group of entities that can be used to approximate what outcomes would have been without the strategy. This may not be possible for every strategy, depending on how entities were selected and targeted for receipt of a strategy. Second, data must be available for both entities engaged in the strategy (the treatment group) and the comparison group. For example, if a strategy was enacted at the start

Comparison groups for a study of WHD partnerships

Selection of brands to join a WHD partnership was targeted and based on a range of data-driven criteria. In the restaurant industry, WHD focused on brands for which the nature of violations could be influenced systemically at the brand level. Among the brands identified as being suitable for a partnership, only some ultimately entered into a partnership. An evaluator can match partnership brand establishments to non-partnership brand establishments based on characteristics like industry subsegment. However, partnership brands will likely differ from other brands in ways that data cannot capture, such as level of commitment to starting a partnership. This makes it difficult to approximate the counterfactual and to estimate the causal effects of partnerships, especially if the characteristics that made brands eligible for or attractive for a WHD partnership are also those that are associated with compliance.

of every WHD investigation, then WHISARD data would contain no comparison group and therefore be inadequate for an evaluation.

Many designs can provide an evaluation with a nuanced understanding of the counterfactual. The gold standard is an RCT, in which entities are randomly assigned to either receive an intervention or not. Such a design could provide the best estimates of the causal effect of a strategy, because they should result in two groups of entities that are near-identical in their observed and unobserved characteristics and only differ in their receipt of the strategy. Quasi-experimental designs using methods other than random assignment to form treatment and comparison groups may prove more feasible and sometimes more ethical to implement than an RCT, and could still provide credible evidence of the effectiveness of strategies. The strongest designs have treatment and comparison groups that are very similar to each other, particularly in any characteristics that can influence the receipt of the strategy and in any characteristics that independently influence compliance outcomes.[14]

Below we provide an overview of several designs, focusing on features of suitable strategies, data requirements, and examples. The examples are intended to illustrate potential ideas to consider and are not based on statistical analysis of specific data or calculations of statistical power.[15] Of note, all approaches can face threats to internal validity from confounding factors (especially those that are unobservable).

a. Randomized controlled trial

An RCT provides the most rigorous evidence of a strategy's effectiveness. It needs to be planned prospectively; it is not a design that can be used to evaluate a strategy retrospectively. It might not always be feasible for WHD to implement random assignment. An RCT is likely to be implemented well when WHD can precisely direct the strategy to certain entities (no potential for spillover from the treatment to the comparison group), when all the entities the strategy is directed to engage in the strategy (that is low risk of "no-shows"), and when there are no ethical concerns about assigning entities to either a treatment or control group. Pre-strategy data are not necessary, although it is preferable to have them to account for chance variations between the two groups and to improve the precision of impact estimates.

[14] The Causal Evidence Guidelines developed by the Clearinghouse for Labor Evaluation and Research (CLEAR) describe CLEAR's system for rating the strength of causal evidence of different impact evaluation designs, including many of those discussed here (CLEAR, n.d.).

[15] These technical considerations of statistical feasibility should be considered as part of a plan for an evaluation. This report does not address this topic.

EXAMPLES

- WHD could randomly select establishments to be required to post different materials about FLSA laws in various specific locations, such as employee restrooms and break areas. This would provide several treatment groups that could be compared to each other and to a control group to learn about the effectiveness of each material and location. WHD could compare violations among each group through follow-up investigations. Such a strategy is not likely to result in spillovers and poses low burden on employers, thus increasing the likelihood of their engagement.
- Random assignment does not require starting with a random sample. WHD could use a nonrandom sample of establishments selected for investigations and randomly assign half to a treatment group and half to a control group for a valid RCT design, especially if the establishments were selected in the same way, such as by a given District Office.
- Randomly selecting past violators and publicizing their violations in a press release is likely to have spillover effects and is therefore a poor candidate for an RCT.
- Randomly selecting worker complaints to pursue may be a poor candidate for an RCT because there could be

A related experimental design involves randomly assigning groups (or "clusters") of entities to an intervention or a control group in order to estimate the impacts of programs designed to affect entire groups. The advantage of such a clustered random assignment design is that it can be more feasible to implement, but it still produces an accurate impact estimate. A strategy is suited for a clustered design if one expects a high likelihood of spillover effects within clusters. Consider a strategy that targeted establishments for education and awareness campaigns and mailed flyers to establishment owners. If an evaluation design involved randomly assigning establishments to either receive this strategy or not, there may be cases where a franchisee owns an establishment in both the intervention group and the control group, and thus there will be a crossover in knowledge across the establishments in the two groups that share an owner. It may therefore be better to consider the franchisees to be clusters and randomly assign franchisees (and thus all the establishments they own) to an intervention or a control group.

b. Regression discontinuity

A regression discontinuity approach involves determining who can and cannot receive a strategy nonrandomly, but based on some precise assignment mechanism. This assignment can then be exploited to identify the causal relationship between receiving a strategy and outcomes. Often, assignment to a strategy is determined based on a cutoff value of a continuous variable (the "running variable"), and then the effects of the strategy can be estimated by comparing the outcomes of the entities that had just qualified for the strategy (the treatment group) to the entities that had just missed qualifying for the strategy (the comparison group). Examples of running variables could be the date employers enroll in a webinar or the level of violations found (see below). This approach can be useful for scenarios in which the strategy can be precisely doled out according to some transparent and consistent assignment mechanism, and there is low risk of spillovers across the cutoff, and can potentially be used retrospectively. This design requires larger samples than an RCT for adequate statistical power (see Schochet 2008 and Deke and Dragoset 2012).

EXAMPLES

- WHD might provide compliance assistance to the first 1,000 employers who sign up for a webinar (the treatment group) and place the next 1,000 employers on a waitlist for 12 months (the comparison group). Comparing the average outcomes of the two groups would not reveal the causal effects of the strategy if the groups differ—for example, because more highly motivated employers are the first to sign up for the webinar. A regression discontinuity design can lessen this problem. The last 50 of the first 1,000 employers who sign up for a webinar are probably relatively similar in motivation to the first 50 of the second group of employers. By comparing the two groups—one that just qualified to receive compliance assistance and one that just missed enrollment—the evaluation can mitigate the concern about differences between the groups and derive a causal estimate of the impact of the strategy.
- WHD might have assessed liquidated damages to establishments found to have back wages or percentages of employees in violation that exceed a certain cutoff value. Given a large sample of investigations, WHD could conduct follow-up investigations of those establishments just above and just below the cutoff to assess whether the assessment of liquidated damages had an impact on compliance.

c. Matched comparison group

When it is not feasible to implement the strategy in a manner that allows WHD to manipulate who is assigned to receive it, a comparison group can be constructed using statistical techniques for "matching" the two groups if the observable factors that determine receipt of a strategy are well understood. This approach is often used in retrospective evaluations. The most straightforward case of matching can be conducted when WHD intentionally targets a strategy to some entities based on measurable, observed criteria. For example, if a strategy was designed for establishments with a history of a certain type of repeat violation, WHD could use WHISARD data to match those establishments to other establishments that had similar histories but did not receive the strategy. If WHD did not intentionally target strategies at certain entities, matching could also be conducted if it were well known that certain factors predict treatment. For example, if WHD opens a webinar series to all employers, those who attend are likely to be more motivated than those who do not, so WHD could match these employers to others who are likely to be highly motivated, such as those who signed up and did not attend, are enrolled to receive a WHD newsletter, or have participated in other voluntary compliance activities in the past.

More often, the receipt of the strategy can be determined by a multidimensional set of pretreatment characteristics rather than a single characteristic. For such a strategy, simple matching would not be appropriate, but WHD might match entities that received the strategy to entities that did not by using methods such as propensity score matching (PSM) (see, for example, Rosenbaum and Rubin 1983). PSM involves calculating a predicted probability (propensity score) of receiving treatment based on observed pretreatment characteristics and then creating a comparison group by matching entities with similar propensity scores. Matching techniques such as PSM can be suitable when data contain a rich set of characteristics that could predict receipt of the strategy. However, if the characteristics that predict treatment are not observable, PSM can produce a comparison group that appears on the surface to be similar to the group that received the strategy, but differs in some important ways that influence compliance,

such as employer commitment levels to voluntary compliance assistance. Therefore, when assessing the suitability of a strategy for an evaluation using matching methods, it is crucial to have a thorough understanding of which types of entities have received the strategy and why.

EXAMPLE

- WHD might consider evaluating a compliance assistance webinar series by comparing a treatment group of employers who attended with a comparison group of those who did not. Attendees could differ from non-attendees on certain characteristics, however. An analysis of enrollment attendance data might indicate that employers are more likely to enroll and attend the series when they have certain characteristics (for example, they are franchisees, and their establishments are less than five years old, located in small towns, and have recently been investigated by WHD). Data on these characteristics could be used to construct a propensity score of how likely it is for an employer to attend the series. Then, the outcomes of each employer in the treatment group are compared to the outcomes of an employer (or several) who did not attend the series but had a similar propensity score. If WHD thinks the groups could still differ in ways that cannot be observed, then the results from the PSM approach may be biased. For example, if the employers in the treatment group place higher priority on being in compliance than the comparison group does, then impacts would look more favorable than they really are.

d. Difference-in-difference

Using a difference-in-difference (DD) design, WHD could estimate the impact of a strategy by looking at whether the treatment group had a greater change in their outcomes after implementation of a strategy than a comparison group did. (These two differences—between treatment and comparison groups and before and after—are what give the approach its name.) This design is appropriate for evaluating strategies implemented at a specific known time that are expected to affect one easily identified group of employers, but not another. They can be used for retrospective evaluations. Outcomes data are required for the treatment and comparison groups both before and after the implementation of a strategy. However, DD designs can face multiple threats to internal validity (for example, the two groups could have been on different outcomes trajectories before the strategy, making it difficult to separate the effects of the strategy from pre-strategy trends in outcomes).

EXAMPLE

- WHD might issue press releases, such as about certain types of violations found in local establishments or guidance for particular industries, within a certain geographic area on a specific date. To evaluate the impact of the press releases, WHD could compare changes in compliance before and after the specific date in the designated geographic area, and compare them to changes in other geographic areas. Because differing economic conditions in the geographic areas could affect compliance at the same time, the design could be strengthened by rolling out the tool in multiple geographic areas at different dates, and incorporating this additional variation in treatment and time into the analysis.
- This approach could be used to examine geographic variation in a strategy's implementation. CORPS may use a compliance assistance strategy of direct outreach to employers in the construction industry. Given a TOC that specifies key activities of the strategy, WHD might work with District Offices to identify and classify variations in these activities. Through discussions and document review with CORPS about how they conduct outreach, WHD might reach consensus that the key dimensions along which implementation differs are intensity, targeting, delivery, and content. The work with the CORPS might classify differences along these dimensions as the following. WHD may find that intensity can be defined as the average hours spent on outreach per week. Type of targeting could be defined based on geography (employers closest to the office), employer interest (employer whom CORPS have interacted with in the past), or likely violations (based on tips or other observations). Mode of delivery may be defined as email, phone, or in-person. Content may be defined as notifications about WHD events, guidance on common FLSA compliance issues in the industry, or general offers of assistance. To evaluate the impact of the outreach strategy on the prevalence of compliance using data at the CORPS level, WHD could compare changes in compliance before and after implementation, accounting for each District Office's intensity, targeting, delivery, and content.

e. Interrupted time series

An interrupted time series (ITS) design compares outcomes for a treatment group over time before and after implementation of a strategy, essentially allowing the members of the treatment group to serve as their own comparison group. The ITS design requires data on outcomes of multiple uses of the strategy and at multiple data points as observed over long periods both before and after the strategy. Using this approach, WHD could assess whether the outcomes of

EXAMPLE

- To evaluate a small number of (for example, four) strategic partnerships, WHD could identify a measure of compliance and examine it at multiple times during the years (for example, two years) before and after embarking on each partnership through the collection of panel data. This approach would be appropriate if the timing of engagement in the partnerships was driven by factors unrelated to the partners' compliance, such as the level of interest in strategic partnerships at WHD or the amount of agency resources available to launch a partnership.
- To evaluate a nationwide compliance assistance strategy, such as national press releases or online compliance assistance resources, WHD could observe compliance outcomes for a long period both before and after several press releases or the introduction of several resources. For best results, WHD could plan the timing of implementing the strategies in advance.

the treatment group are different from what would be predicted based on the trend in outcomes that existed before the strategy. Importantly, the timing of the intervention should be predetermined and not chosen based on compliance trends. This approach is well suited to strategies in which the employers receiving the intervention are likely to differ from others, and data can be collected frequently over a long period of time. It can potentially be used for retrospective evaluations.

When assessing the evaluability of a strategy, WHD could consider the types of evaluation designs that could feasibly be employed to study its effectiveness, the resources (for example, data or logistics) that would be needed to execute the feasible designs, and how the findings from the design could be interpreted and leveraged to facilitate the best use of WHD's resources.

IV. OPPORTUNITIES AND CHALLENGES IN MONITORING AND EVALUATION OF STRATEGIES

In this chapter, we describe the potential opportunities and challenges that WHD or other agencies might face when structuring a strong monitoring and evaluation process for its strategies, a process that addresses each of the five factors for success. We consider opportunities to be features of a strategy, agency processes, or resources that an agency might build on to design and support meaningful monitoring and evaluation. We consider challenges to be features of a strategy or of agency processes or resources that can make monitoring or evaluation unproductive. We identified general potential opportunities and challenges that WHD and other agencies might face in continuing to strengthen and build monitoring and evaluation efforts (Table IV.1) based on knowledge development Mathematica engaged in with WHD. The discussion is intended as a general resource to highlight good research practices. We present examples relevant to WHD's work to illustrate the general points, but it is important to note that they should not be interpreted as reflections on what WHD has done or may be doing.

Agencies such as WHD can build on many processes and resources to design and support meaningful monitoring and evaluation, including coordination and targeting processes, documentation and guidance, violation and performance measures, administrative data systems and data collection efforts, and data quality assurance processes, as the opportunities section of Table IV.1 shows. In addition, some strategies, such as those that are well-established or include geographic variation, have features that may make them well suited for evaluation, in particular for RCTs. However, designing and supporting monitoring and evaluation can be challenging due to stakeholder disagreements and desires, agency constraints, or features of strategies or their outcomes.

A. Factor #1: Documented and supported core activities

Documenting and supporting a strategy's core activities is fundamental to any monitoring or evaluation. Existing agency processes and resources provide much-needed documentation and support and have features that can facilitate additional development, but building a comprehensive TOC could be expensive for some strategies.

Opportunity 1.a: Internal coordination efforts offer opportunities to define and gain agreement on the TOC and to consistently implement core activities.

If a hierarchical structure and communication process exists within an agency, it can facilitate documentation and support. Leaders at the multiple levels in the hierarchy may work together to develop annual plans, providing opportunities to also discuss the components of strategies' TOC. In addition, implementing strategies can include multiple levels of the hierarchy, which facilitates the standardization of materials and implementation. For example, the WHD national office works with District Offices to implement strategic enforcement strategies and could

observe and help ensure that the core strategic partnership activities of investigating and communicating with selected brands and their establishments are carried out consistently.

Table IV.1. Potential opportunities and challenges in monitoring and evaluation of WHD compliance strategies

Factors	Opportunities	Challenges
1. Documented and supported core activities	a. Internal coordination efforts offer opportunities to define and gain agreement on the TOC and to consistently implement core activities.	a. Complex strategies, or those that have extensive data collection needs, are often those in most need of stakeholder agreement on the TOC, and often require more resources to gain needed agreement.
	b. Existing documentation and guidance can provide a foundation for a well-articulated TOC.	
2. Measurable outputs and outcomes	a. Violation and performance measures can help structure measurable outputs and outcomes.	a. Stakeholders may disagree on which measures to examine.
3. Available, appropriate data	a. Data quality assurance procedures could be formalized, including by aligning performance standards with data quality.	a. Modifying existing data systems to collect additional data for monitoring and evaluation may be difficult.
	b. Administrative data might be modified to capture additional elements in the TOC and thereby become the basis for a monitoring and evaluation data collection system.	b. Spillovers, which are often outcomes of strategies, can be difficult to capture.
	c. Electronic metadata might be leveraged as an inexpensive source of data.	
	d. Existing interactions with entities receiving strategies offer opportunities to collect data.	
	e. Follow-up investigations could provide valuable information to enforcement agencies.	
	f. Investment in an external sampling frame could strengthen monitoring and evaluation efforts.	
4. Maturity	a. Well-established strategies can serve as a starting point to develop evaluations.	a. Stakeholders may want evidence on strategies that do not have all components in place; the results of evaluating such strategies might understate their potential value.
	b. Agencies can use monitoring of strategies as a way to ensure TOC components are in place.	

Factors	Opportunities	Challenges
5. Internal validity	a. Performance standards aligned with evaluation goals may create incentives that support evaluation.	a. Spillover effects make it difficult to capture a counterfactual.
	b. Agencies could build on existing processes to develop RCTs and other rigorous designs for evaluation.	b. Random assignment may not be feasible for evaluation of some strategies.
	c. Geographic variation in strategies could be exploited in developing evaluation designs.	

RCT = randomized control trial. TOC = theory of change.

Opportunity 1.b: Existing documentation and guidance can provide a foundation for a well-articulated TOC.

Written and codified documentation and guidance, including educational materials, procedure manuals, and templates for activities, are a solid starting point for identifying the core activities of a strategy and illustrating how they can be adapted to local contexts without compromising the strategy. For example, WHD compliance assistance webinars and presentations are typically based on standardized presentation materials including slide decks and handouts, which document the content and provide a template that can help ensure consistency across presentations.

Challenge 1.a: Complex strategies, or those that have extensive data collection needs, are often those in most need of stakeholder agreement on the TOC, and often require more resources to gain needed agreement.

Complex strategies include core activities that are difficult to standardize and quantify, so reaching agreement on what they are and documenting and supporting them can be a costly and lengthy process. For example, WHD strategic partnerships are complex because they rely on establishing and fostering relationships; moreover, if they are tailored to the partner's specific compliance challenges, it could be difficult to characterize the core activities. Stakeholders may disagree on the specific activities that are and should be used to implement the strategy, as well as how to measure their content, quantity, quality, or delivery (Weiss et al. 2014). They can also disagree on how similar the activities must be to be considered consistent implementation. Reaching agreement may require more discussion and documentation for complex strategies than for other strategies.

B. Factor #2: Measurable outcomes

Monitoring and evaluation must have measurable outcomes to be feasible. Existing measures, even if not developed for research purposes, can be used and modified to support monitoring and evaluation, although reaching agreement on specific measures may not be easy.

Opportunity 2.a: Violation and performance measures can help enforcement agencies structure measurable outputs and outcomes.

Measures that enforcement agencies have already developed related to agency performance and investigative findings can offer a solid starting point for identifying measurable outputs and outcomes of strategies. The advantage of these measures is that they have already been reviewed and agreed on and thus can, in all likelihood, be readily used. For example, WHD performance measures for fiscal year 2018 (U.S. Department of Labor 2018) included measures related to outputs, such as the number of outreach hours to employers and number of compliance actions (investigations). Adaptations such as the number of outreach hours to businesses engaged in strategic partnerships or the number of compliance actions among the establishments of a brand that is a strategic enforcement target could potentially be used to examine those strategies.

Challenge 2.b: Stakeholders may disagree on which measures to examine.

It can be challenging for agency stakeholders to determine a set of measures to examine to support the TOC. For example, they may face choices about which outputs may be both closely tied to outcomes and measurable, or what relevant outcomes related to organizational behaviors may be measurable. This challenge may be heightened given finite resources that may limit the number of outcomes that can be collected and require agencies to prioritize some measures over others. For example, for a given strategy, investigation of a branded fast-food establishment may lead to changes in violation rates among other fast-food establishments in the local area, but reaching agreement on how to define a local geographic area (perhaps by local population density or commuting patterns) might not be easy.

C. Factor #3: Available, appropriate data

Available, appropriate data are key to informative monitoring and evaluation. Existing agency administrative and other data as well as additional data collection could help fill some current data gaps, particularly if supported through enhanced data quality control processes.

Opportunity 3.a: Data quality assurance procedures could be formalized, including by aligning performance standards with data quality.

Given the central role that data play in monitoring and evaluation, it is critical to ensure data quality. As a specific example, employer contact information in WHISARD must be of high quality to be able to match to external data, as findings in Patnaik (2020) suggest. For this reason, agencies that have not formalized data quality assurance procedures could consider doing so. For example, agencies could consider enhancing procedures for accurate data entry, data verification, and quality control reviews. Instruction manuals for data entry and data entry systems could describe ways to check and standardize establishment contact information, indicate specific ranges of data field values to use where possible (instead of open-ended responses), and provide guidance on what to do when the value of a data field is unknown. They could also provide guidance on how to code information correctly or how to identify the correct information to include. Data verification processes could periodically be used check the accuracy

of a small fraction of cases in key data fields against another source or could potentially build checks into the data entry system. Quality control reviews could check the extent of missing values, values out of range in data fields, and discrepancies between information provided in different fields, for example. Furthermore, agencies could consider creating performance measures aligned with data quality to create incentives for quality assurance. Such measures might reflect contributions to data quality, such as number of cases that went through quality control review or percentage of cases assessed as complete and accurate by quality control reviewers.

Opportunity 3.b: Agency data systems might be modified to capture additional elements in the TOC and thereby become the basis for a monitoring and evaluation data collection system.

Agency case management systems can be a key source of some needed data for monitoring and evaluation, and may also be modified to capture or link to additional elements of the TOC, either in existing modules or additional modules. Information on other outcomes, outputs, and activities could potentially be gathered from other sources internal to the agency and included in a data module. Examples of potential modifications include:

- **Adding modules to track activities.** For example, for a strategic partnership strategy, a module could track core activities (such as whether WHD representatives provided information in a meeting, including the date and duration of the meeting), outputs (such as whether the partner provided required staff trainings and when), and potentially even outcomes (such as measures of partner attitudes, if observation or survey data collection instruments to do so were developed). Some of this information may already be readily linkable to agency case management systems. Establishment identifiers could be created to facilitate this.
- **Adding establishment and geographic characteristics**. For an enforcement agency, information on establishment characteristics and context could be linked to a case management system to yield substantial benefits to monitoring and evaluation. External data could contain rich information about establishments, including information on an establishment's role in a business model such as franchising or supply chain, parent company ownership, product and customer characteristics, and local geographic and industry features like local density of similar industry establishments or the population's average earnings. Such data could be used in monitoring and evaluation to provide context on industry trends and operator density that could influence a strategy's effects or to define subgroups that might have different outcomes from the strategy.

Opportunity 3.c: Electronic metadata might be leveraged as an inexpensive source of data.

For strategies involving web and email communications, electronic metadata can provide accurate output and outcomes data quickly at a relatively low cost. Electronic metadata can be defined as information about electronic records, such as email communications, that are gathered and stored by electronic systems, such as email delivery systems. Examples of electronic

metadata include information on whether a recipient opened an email that is available from the email delivery system, and whether users registered for an online event that is available from webinar software. They are is based on system data, so they are accurate. They are available through the agency email platform provider, so they can be readily accessed. For strategies including activities that take place via email or web, electronic metadata could provide output and outcomes information.

Opportunity 3.d: Existing interactions with entities receiving strategies offer opportunities to collect data.

To collect useful data needed for monitoring and evaluation, agencies could take advantage of existing touch points with entities receiving strategies. For example, in the course of implementing strategic partnerships, WHD staff could observe implementation activities, collect and review documentation of outputs, and interview the partner about outputs and even outcomes related to the partner's perceptions or learning. Similar interviews could be conducted in the course of an investigation as well.

Opportunity 3.e: Follow-up investigations could provide valuable information to enforcement agencies.

Follow-up investigations can be valuable in both monitoring and evaluation because they can show changes in outcomes for specific employers who were the focus of a strategy. This can be especially valuable when analyzing non-random samples of employers who were investigated. For example, if an agency is interested in evaluating a strategy used in a particular industry and compares compliance outcomes of employers investigated before and after the strategy was used, it may be difficult to determine whether an increase reflects improvement in employer compliance behavior or better detection of violators by investigators. In contrast, if the agency compares outcomes of the same employers before and after the strategy was used, it could identify improvement (or a lack thereof). This would support the internal validity of the evaluation.

Opportunity 3.f: Investment in an external sampling frame could strengthen monitoring and evaluation efforts.

Enforcement agencies might consider strengthening monitoring and evaluation by investing in external data covering a population or random sample of employers to use as a sampling frame. Investment in such data could yield great benefits by allowing the agency to prospectively design monitoring and evaluation efforts that measure the prevalence of outcomes as well as RCT evaluations with internal validity. To create a sustainable sampling frame infrastructure suitable for broad monitoring and evaluation, agencies could explore cost of investing in an initial database and periodic updates.

Challenge 3.a: Modifying existing data systems to collect additional data for monitoring and evaluation may be difficult.

Information that may be useful to an evaluation may not be included among required data elements in the case management system. Opportunity 3.b presented examples of such information and how it could be incorporated. However, modifying legacy systems or introducing evaluation protocols can be resource-intensive. In addition to resources needed to develop, pilot, implement, and train for such changes, time and effort may be needed to reach agreement on taking such a step. Moreover, for enforcement agencies, data collected beyond the scope of what is required during an investigation may be more at risk for data quality issues such as completeness. Aligning incentives such as investigation performance measures with the goals of additional data collection could help ameliorate these potential issues.

Challenge 3.b: Spillovers, which are often outcomes of strategies, can be difficult to capture.

Spillover outcomes are outcomes for entities who do not receive a strategy directly, but defining and identifying this population can be challenging. Even if the population is clear conceptually—for example, certain enforcement strategies may spill over to other establishments in a brand rather than to other establishments in the industry—it can be challenging to measure. For example, in a WHD strategy involving conducting and publicizing enforcement of small mom-and-pop coffee shops, available data may not include these establishments. Specifically, available data may exclude establishments with number of employees below a given threshold, or may not provide detailed enough classification to identify the industry segment.

D. Factor #4: Implementation maturity

A strategy's implementation must be mature for an evaluation to produce accurate findings on causality or impact. Well-established strategies can provide a starting point for an evaluation, and monitoring can help other strategies become mature, but there could be pressure from stakeholders to push immature strategies ahead for evaluation.

Opportunity 4.a: Well-established strategies can serve as a starting point to develop evaluations.

Well-established, mature strategies could be examined for evaluation without delay. Such strategies have inputs and activities that are part of an agency's typical routines, and staff follow established implementation processes. Agencies could use these mature strategies to help build and strengthen their monitoring and evaluation processes, instead of trying to develop both a strategy and a monitoring and evaluation process at the same time.

Opportunity 4.b: Agencies can use monitoring of their strategies as a way to ensure TOC components are in place.

For strategies that are not yet ready for evaluation, monitoring offers an opportunity to help bring them to maturity and thus develop a pipeline of mature strategies for evaluation. Monitoring can help identify whether specific outputs and outcomes are being realized; this can help shed light on whether specific inputs and activities are being implemented and supported so that agencies can consider providing additional implementation support and help make the strategy mature. Consider an example compliance assistance strategy in which WHD provides web-based tools employers can use to diagnose and fix compliance issues. If monitoring shows low usage rates of particular components of the tool (perhaps through examination of electronic metadata), WHD could check for and address any problems accessing the tool through certain web browsers or broken links.

Challenge 4.a: Stakeholders may want evidence on strategies that do not have all components in place; the results of evaluating such strategies might understate their potential value.

New and innovative strategies are often of great interest, but if they are not mature, an evaluation can produce inaccurate and misleading information. To meet the needs for evidence about strategies that are still under development, agencies can consider whether other types of research besides impact evaluation may be appropriate. For example, monitoring could show whether the strategy may be on its way to create impacts by producing expected outputs; formative evaluation could shed light on factors that support the implementation of key strategy activities. Stakeholders would need to interpret the findings appropriately and not as evidence of whether the strategy works.

E. Factor #5: Internal validity

To understand whether strategies caused outcomes and to estimate their impacts, evaluations must have internal validity. Certain agency processes and resources provide opportunities for RCTs, the most rigorous type of evaluation design. Several types of strategies may be well suited for evaluation using RCTs, but this design is not always feasible for evaluation of others.

Opportunity 5.a: Performance standards aligned with evaluation goals could create incentives that support evaluation.

Agency staff often play a key role in implementing strategies. Agency performance measures create strong incentives for them to do this work in particular ways, which could make it challenging for staff to support evaluation activities. For example, performance measures related to efficiency (output per labor hour) could discourage staff from spending additional time to engage in implementation activities required for an evaluation. Agencies could consider creating performance measures aligned with monitoring and evaluation goals to support staff and ensure the production of high-quality evidence. Such measures might reflect specific activities

conducted for an evaluation or specific contributions to data quality (such as the percentage of cases that went through quality control review).

Opportunity 5.b: Agencies could build on existing processes to develop RCTs and other rigorous designs for evaluation.

Enforcement agencies can take advantage of existing processes to identify employers to receive strategies—targeting procedures—and adapt them to support evaluation designs with internal validity. For example, an agency may select a set of establishments for enforcement that are at risk for violations. This existing targeting procedure makes it challenging to identify a comparison group of establishments that are similar in terms of compliance and other characteristics before the strategy's application. One option to address this issue is to randomly assign establishments from this set to receive enforcement. For example, within WHD, District Office managers could identify a group of establishments with violations that all meet high priority criteria for liquidated damages (LDs), but instead of assessing LDs for all of them, assess LDs on a randomly selected fraction, similar to what the Occupational Safety and Health Administration did in its site-specific targeting (Johnson et al. 2019). This change in targeting procedures would likely yield a valid comparison group for evaluation. Other changes to targeting procedures might include creating rosters of establishments to receive a strategy ordered by priority (for example, number of employees or annual revenue) and designating a cutoff priority level that would be used to select targets. If the roster were extensive, this procedure could support regression discontinuity designs (described in Chapter III).

Opportunity 5.c: Geographic variation in strategies could be exploited in developing evaluation designs.

Agency strategies are sometimes implemented by local staff. These staff often have a degree of discretion in how they implement the strategies and may make adaptations to fit local conditions. For example, within WHD, District Office staff could leverage a strong relationship with an industry association to disseminate guidance or promote the use of compliance tools through its network. Their innovations can be valuable in detecting and remedying more violations, and in developing insights about potential new approaches. Agencies could consider developing evaluation designs that leverage this geographic variation, such as the difference-in-difference design discussed above. They may also consider taking advantage of it when conducting pilot tests of implementation activities to develop strategies, although this topic was not covered in this report.

Challenge 5.a: Spillover effects make it difficult to capture a counterfactual.

Because many agency strategies can affect entities that were not targeted directly—through press attention, word of mouth, or even treatment group entities sharing information with potential comparison group members—it can be challenging to identify a counterfactual for evaluation. Spillover effects can be a challenge in an RCT as well. For example, consider the example of conducting an RCT of a compliance assistance strategy that involves sending email invitations to

a webinar by randomly assigning stakeholders on WHD's Key News electronic mailing list to treatment and control groups. It can be difficult to prevent some control individuals from receiving the strategy. Some individuals or organizations could have multiple email addresses on the list that could potentially be assigned to different conditions. Additionally, recipients could forward the email to their contacts, who may have been assigned to a different condition. These crossovers can dilute the evidence of the strategy's impacts.

Challenge 5.b: Random assignment may not be feasible for evaluation of some strategies.

One challenge to evaluation is that RCT designs are not feasible for some strategies. For example, the use of some strategies—such as enforcement agencies opening investigations into and levying penalties on establishments meeting certain conditions—may be required by law, so that assignment to a control condition is not possible. As another example, for WHD, random assignment is not feasible for evaluation of strategic partnerships, because the partner must not only be targeted for the partnership by WHD but must also agree to engage in the partnership; moreover, the substantial resources required to develop a partnership can prohibit the creation of a treatment group of multiple partners. Other rigorous designs may be appropriate to evaluate such strategies, however.

V. STEPS TO CONSIDER FOR SUCCESSFUL MONITORING AND EVALUATION

The previous chapter identified opportunities and challenges that WHD and other agencies might face in monitoring and evaluating their strategies. In this chapter, we summarize steps that emerged from that discussion that can be considered to help ensure that potential future monitoring and evaluation of strategies yields useful information. The steps are based on the discussion in Chapters II and III and the opportunities and challenges presented in Chapter IV. They are intended as a general resource reflecting good research practices and should not be interpreted as a reflection on what WHD has done or may be doing.

A. Build data infrastructure suitable for monitoring and evaluation

Agencies could consider investing in several efforts that may ensure that all data for monitoring and evaluation are of high quality.

1. **Build capacity for ensuring data quality and conducting data analytics**. Potential approaches to consider include investing in training for staff who report and analyze data, and developing procedures to verify data and ensure their quality.
2. **Consider ways that administrative data could be collected and used to support the internal validity of evaluations**. Enforcement agencies might consider conducting follow-up investigations to develop panel data on establishments and estimate changes in compliance. They could use external data to develop a sampling frame, from which they could select establishments to investigate using random sampling and estimate violation prevalence. Alternatively, agencies could establish a protocol of pro-actively linking establishments to external data before an investigation begins, which would facilitate efforts to validate both external and administrative data and enhance the value of the external data by improving its match rate to the administrative data.
3. **Develop and maintain data sets that could be linked to agency data systems**. For example, these could come from primary data collection on strategy implementation and intermediate outcomes, or from acquiring external data that reflect, for example, a population of establishments and their characteristics. Building capacity around the ongoing maintenance and statistical modeling of these data sets could help evolve monitoring and evaluation activities and improve the quality of analysis possible.

B. Specify performance measures and data collection needs to support evaluation goals

Agency staff often play a key role in implementing strategies. Agency performance measures create strong incentives for them to do this work in particular ways, which could make it challenging for staff to support evaluation activities. For example, performance measures related to efficiency (output per labor hour) could discourage staff from spending additional time to

engage in implementation activities required for an evaluation. Agencies could consider creating performance measures aligned with monitoring and evaluation goals to support staff and ensure the production of high-quality evidence. Such measures might reflect specific activities conducted for an evaluation or specific contributions to data quality (such as the percentage of cases that went through quality control review).

C. Develop a system for monitoring and evaluation of strategies consistent with the evaluation design

It may be helpful to consider developing a system for monitoring and evaluation of strategies. The advantage of such a system is that monitoring and evaluation processes could become more efficient and consistent through repeated use, and that all strategies would have the opportunity for monitoring and evaluation. It would be important to build consensus within an agency and among stakeholders at each stage to strengthen the monitoring and evaluation designs and help reach agreement about how to interpret the findings. The monitoring and evaluation process could include the following steps based on a summary of the discussion of five factors for consideration:

1. **Build a detailed, evidence-driven TOC for each strategy.** To describe the components of the TOC—inputs, activities, outputs, and outcomes—in detail, it is important to have data on how the strategy is actually implemented and the specific goals it pursues. These should be specific data on who is involved in the component; when, where, and how the component happens; and potentially on how often it happens and how long it lasts (frequency and duration). There are many ways to collect these data, including by reviewing documents, observing activities, and interviewing the field staff who deliver the strategy and the key personnel associated with the entities that receive the strategy.
2. **Monitor strategies to determine their maturity.** To gauge whether a strategy is being implemented as planned, it is important to identify and analyze data measuring the components of the TOC. The maturity of implementation and the extent to which the TOC has been implemented will determine the appropriate evaluation design. Both qualitative and quantitative data can be collected through these activities using a range of techniques. For example, interviews with staff and stakeholders, administrative data analysis, and research can all support learning and ongoing improvements during program implementation.
3. **Develop appropriate evaluation designs.** Integrating evaluation design with planning for implementation could offer the best chance for a successful evaluation. By planning an evaluation before implementation, agencies may be able to improve or strengthen the conditions that support this success, such as collecting or enhancing documentation and data during implementation and constructing a counterfactual condition.

REFERENCES

Ashenfelter, Orley, and Robert S. Smith. "Compliance with the Minimum Wage Law." *Journal of Political Economy,* vol. 87, no. 2, 1979, pp. 333–350. Available at https://www.journals.uchicago.edu/doi/abs/10.1086/260759. Accessed March 14, 2018.

Barnes, Jeb, and Thomas F. Burke. "The Diffusion of Rights: From Law on the Books to Organizational Rights Practices." *Law & Society Review,* vol. 40, no. 3, 2006, pp. 493–524. Available at https://pdfs.semanticscholar.org/2497/1717a32657f6ff3f1387f91dacc99876b695.pdf. Accessed March 14, 2018.

Behavioural Insights Team. "Applying Behavioural Insights to Reduce Fraud, Error and Debt." London: Cabinet Office, 2012. Available at http://38r8om2xjhhl25mw24492dir.wpengine.netdna-cdn.com/wp-content/uploads/2015/07/BIT_FraudErrorDebt_accessible.pdf. Accessed May 22, 2020.

Bertram, Rosalyn, Karen Blase, David Shern, Pat Shea, and Dean Fixsen. "Policy Research Brief: Implementation Opportunities and Challenges for Prevention and Promotion Initiatives." Alexandria, VA: National Association of State Mental Health Program Directors, 2011. Available at https://www.researchgate.net/publication/233885310_Policy_Research_Brief_Implementation_Opportunities_and_Challenges_for_Prevention_and_Promotion_Initiatives. Accessed May 22, 2020.

Braithwaite, John, and Toni Makkai. "Testing an Expected Utility Model of Corporate Deterrence." *Law and Society Review,* vol. 25, no. 1, 1991, pp. 7–40. Available at https://www.semanticscholar.org/paper/Testing-an-Expected-Utility-Model-of-Corporate-Braithwaite-Makkai/11b072bbdd61b3180fb19ee8ff11b56b11176996. Accessed May 22, 2020.

Calavita, Kitty. "Employer Sanctions Violations: Toward a Dialectical Model of White-Collar Crime." *Law & Society Review,* vol. 24, no. 4, January 1990, pp. 1041–1069. Available at https://www.researchgate.net/publication/272593736_Employer_Sanctions_Violations_Toward_a_Dialectical_Model_of_White-Collar_Crime. Accessed April 17, 2018.

Chang, Yang-Ming, and Isaac Ehrlich. "On the Economics of Compliance with the Minimum Wage Law." *Journal of Political Economy,* vol. 93, no. 1, 1985, pp. 84–91. Available at https://www.researchgate.net/publication/24108358_On_the_Economics_of_Compliance_with_the_Minimum_Wage_Law_. Accessed May 22, 2020.

Cialdini, Robert B., Linda J. Demaine, Brad J. Sagarin, Daniel W. Barrett, Kelton Rhoads, and Patricia L. Winter. "Managing Social Norms for Persuasive Impact." *Social Influence,* vol. 1, no. 1, 2006, pp. 3–15. Available at https://www.tandfonline.com/doi/abs/10.1080/15534510500181459 . Accessed May 22, 2020.

Clearinghouse for Labor Evaluation and Research. "Causal Evidence Guidelines, Version 2.1." Washington, DC: U.S. Department of Labor, n.d. Available at https://clear.dol.gov/sites/default/files/CLEAR_EvidenceGuidelines_V2.1.pdf . Accessed February 26, 2020.

Deke, John and Lisa Dragoset. "Statistical Power for Regression Discontinuity Designs in Education: Empirical Estimates of Design Effects Relative to Randomized Controlled Trials." Working paper 8. Washington, DC: Mathematica, June 2012. Available at https://www.researchgate.net/publication/241761734_Statistical_Power_for_Regression_Discontinuity_Designs_in_Education_Empirical_Estimates_of_Design_Effects_Relative_to_Randomized_Controlled_Trials_Princeton_NJ_Mathematica_Policy_Research. Accessed May 22, 2020.

Deke, John and Michael Puma. "Coping with Missing Data in Randomized Controlled Trials." Evaluation Technical Assistance Brief #3. Washington, DC: U.S. Department of Health and Human Services, Office of Adolescent Health, 2013. Available at https://pdfs.semanticscholar.org/d866/53ca30cdc392daac1d8ead2385b7879af5c0.pdf. Accessed May 22, 2020.

Deutsch, Jonah, Naihobe Gonzalez, and Nan Maxwell. "Behavioral Interventions for Compliance Assistance: Design Report." Report submitted to U.S. Department of Labor. Washington, DC: Mathematica, March 27, 2019.

Dolfin, Sarah, Nan Maxwell, Alix Gould-Werth, Armando Yañez, Jonah Deutsch, and Libby Hendrix. "Compliance Strategies Evaluation Literature and Database Review." Report submitted to U.S. Department of Labor. Washington, DC: Mathematica, October 18, 2018.

Eastern Research Group. "The Social and Economic Effects of Wage Violations: Estimates for California and New York." Washington, DC: U.S. Department of Labor, 2014. Available at https://selfsufficiencyresearch.org/content/social-and-economic-effects-wage-violations-estimates-california-and-new-york. Accessed May 22, 2020

Eastern Research Group. "Evaluation of DOL's Wage and Hour Division Child Labor Program." Washington, DC: U.S. Department of Labor, 2009. Available at https://www.dol.gov/sites/dolgov/files/WHD/legacy/files/ChildLaborEvalFinalReport.pdf. Accessed May 22, 2020.

Evans, Mary F., Lirong Liu, and Sarah L. Stafford. "Do Environmental Audits Improve Long-Term Compliance? Evidence from Manufacturing Facilities in Michigan." *Journal of Regulatory Economics*, vol. 40, no. 3, 2011, pp. 279–302. Available at https://www.researchgate.net/publication/225675117_Do_environmental_audits_improve_long-term_compliance_Evidence_from_manufacturing_facilities_in_Michigan . Accessed May 22, 2020.

Fine, Janice, and Jennifer Gordon. "Strengthening Labor Standards Enforcement through Partnerships with Workers' Organizations." *Politics & Society*, vol. 38, no. 4, 2010, pp. 552–585. Available at http://journals.sagepub.com/doi/pdf/10.1177/0032329210381240. Accessed March 14, 2018.

Fixsen, Dean. L., and Karen A. Blase. "Creating New Realities: Program Development and Dissemination." *Journal of Applied Behavior Analysis*, vol. 26, 1993, pp. 597–615. Available at https://www.researchgate.net/publication/14891138_Creating_new_realities_Program_development_and_dissemination. Accessed May 22, 2020.

Fixsen, Dean L., Sandra F. Naoom, Karen A. Blase, Robert M. Friedman, and Frances Wallace. "Implementation Research: A Synthesis of the Literature." FMHI Publication no. 231. Tampa, FL: University of South Florida, Louis de la Parte Florida Mental Health Institute, The National Implementation Research Network, January 2005. Available at https://nirn.fpg.unc.edu/sites/nirn.fpg.unc.edu/files/resources/NIRN-MonographFull-01-2005.pdf. Accessed October 11, 2018.

Friedrichs, David O. *Trusted Criminals: White Collar Crime in Contemporary Society.* Belmont, CA: Wadsworth Cengage Learning, 2009. Available at https://www.researchgate.net/publication/40936285_Trusted_Criminals_White_Collar_Crime_in_Contemporary_Society. Accessed May 22, 2020.

Galvin, Daniel J. "Deterring Wage Theft: Alt-Labor, State Politics, and the Policy Determinants of Minimum Wage Compliance." *Perspectives on Politics,* vol. 14, no. 2, 2016, pp. 324–350. Available at https://www.scholars.northwestern.edu/en/publications/deterring-wage-theft-alt-labor-state-politics-and-the-policy-dete. Accessed March 14, 2018.

Government Accountability Office. "Fair Labor Standards Act: Better Use of Available Resources and Consistent Reporting Could Improve Compliance." GAO-08-962T. Testimony before the Committee on Education and Labor, House of Representatives, 2008. Available at https://www.gao.gov/new.items/d08962t.pdf. Accessed November 20, 2019.

Gray, Garry C., and Susan S. Silbey. "Governing Inside the Organization: Interpreting Regulation and Compliance." *American Journal of Sociology,* vol. 120, no. 1, 2014, pp. 96–145. Available at https://www.researchgate.net/publication/272749528_Governing_Inside_the_Organization_Interpreting_Regulation_and_Compliance_1 . Accessed May 22, 2020.

Gray, Wayne B., and John M. Mendeloff. "The Declining Effects of OSHA Inspections on Manufacturing Injuries, 1979–1998." *ILR Review*, vol. 58, no. 4, 2005, pp. 571–587. Available at https://www.researchgate.net/publication/5119573_The_Declining_Effects_of_Osha_Inspections_on_Manufacturing_Injuries_1979-1998. Accessed May 22, 2020.

Gray, Wayne B., and John T. Scholz. "Analyzing the Equity and Efficiency of OSHA Enforcement." *Law & Policy,* vol. 13, no. 3, 1991, pp. 185–214. Available at https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9930.1991.tb00066.x. Accessed April 17, 2018.

Gray, Wayne B., and John T. Scholz. "Does Regulatory Enforcement Work? A Panel Analysis of OSHA Enforcement." *Law & Society Review*, vol. 287, no. 1, 1993, pp. 177–214. Available at http://www.jstor.org/stable/3053754. Accessed March 14, 2018.

Gray, Wayne B., and Ronald J. Shadbegian. "When and Why Do Plants Comply? Paper Mills in the 1980s." *Law & Policy,* vol. 27, no. 2, 2005, pp. 238–261. Available at https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9930.2005.00199.x. Accessed April 17, 2018.

Heard, Kenya, Elisabeth O'Toole, Rohit Naimpally, and Lindsey Bressler. "Real-World Challenges to Randomization and Their Solutions." Cambridge, MA: Abdul Latif Jameel Poverty Action Lab (J-PAL) North America, 2017. Available at https://www.povertyactionlab.org/sites/default/files/resources/2017.04.14-Real-World-Challenges-to-Randomization-and-Their-Solutions.pdf. Accessed April 8, 2020.

Hernandez, Mario, and Sharon Hodges. "Building Upon the Theory of Change for Systems of Care." *Journal of Emotional and Behavioral Disorders*, vol. 11, no. 1, 2003, pp. 19–26. Available at https://www.researchgate.net/publication/247784972_Building_Upon_the_Theory_of_Change_for_Systems_of_Care. Accessed May 22, 2020.

Hodges, Sharon, Mario Hernandez, Teresa Nessman, and Lodi Lipien. "Creating Change and Keeping it Real: How Excellent Child-Serving Organizations Carry Out Their Goals. Cross-site Findings for Phase I Community-Based Theories of Change." Tampa, FL: Louis de la Parte Florida Mental Health Institute, Research and Training Center for Children's Mental Health, University of South Florida, December 2002. Available at http://rtckids.fmhi.usf.edu/rtcpubs/creatingchange/CreatingChange.pdf. Accessed November 5, 2019.

Hoynes, Hilary, Marianne Page, and Ann Huff Stevens. "Can Targeted Transfers Improve Birth Outcomes? Evidence from the Introduction of the WIC Program." *Journal of Public Economics*, vol. 95, nos. 7-8, pp. 813–827, 2011. Available at https://www.sciencedirect.com/science/article/pii/S0047272710002082?via%3Dihub. Accessed April 5, 2020.

Institute of Medicine, Committee on Quality of Health Care in America. *Crossing the Quality Chasm: A New Health System for the 21st Century*. Washington, D.C.: National Academy Press, 2001. Available at http://www.ihi.org/resources/Pages/Publications/CrossingtheQualityChasmANewHealthSystemforthe21stCentury.aspx. Accessed May 26, 2020.

Ji, Min Woong, and David Weil. "Does Ownership Structure Influence Regulatory Behavior? The Impact of Franchising on Labor Standards Compliance." Working Paper. Boston, MA: Boston University, 2009. Available at https://www.researchgate.net/publication/228200171_Does_Ownership_Structure_Influence_Regulatory_Behavior_The_Impact_of_Franchising_on_Labor_Standards_Compliance. Accessed May 22, 2020.

Ji, Min Woong, and David Weil. "The Impact of Franchising on Labor Standards Compliance." *ILR Review*, vol. 68, no. 5, 2015, pp. 977–1006. Available at http://journals.sagepub.com/doi/full/10.1177/0019793915586384. Accessed March 14, 2018.

Jin, Ginger Z., and Jungmin Lee. "Inspection Technology, Detection, and Compliance: Evidence from Florida Restaurant Inspections." *Rand Journal of Economics*, vol. 45, no. 4, 2014, pp. 885–917. Available at https://pdfs.semanticscholar.org/b186/4090867270bd24a9d214517de721b754dd2b.pdf. Accessed March 14, 2018.

Johnson, Matthew S. "Regulation by Shaming: Deterrence Effects of Publicizing Violations of Workplace Safety and Health Laws." Duke University, Working Paper, 2018. Available at https://www.semanticscholar.org/paper/Regulation-by-Shaming-%3A-Deterrence-Effects-of-of-%E2%88%97-Johnson/9a0c4eec304fd84dc145eb05002e84d712a45808 . Accessed May 22, 2020.

Johnson, Matthew S., David I. Levine, and Michael W. Toffel. "Organizational and Geographic Spillover Effects of Regulatory Inspections: Evidence from OSHA." DOL Scholars Final Report, 2017. Available at https://www.dol.gov/sites/dolgov/files/OASP/legacy/files/Organizational-and-Geographic-Spillover-Effects-of-Regulatory-Inspections-Evidencce-from-OSHA.pdf. Accessed May 22, 2020.

Johnson, Matthew S., David I. Levine, and Michael W. Toffel. "Improving Regulatory Effectiveness through Better Targeting: Evidence from OSHA." Harvard Business School Technology and Operations Management Unit, Working Paper No. 20-019, 2019. Available at https://pdfs.semanticscholar.org/88b4/a63f544b2d9fcc797dde768a813458c1dcce.pdf. Accessed May 26, 2020.

Kagan, Robert A., Neil Gunningham, and Dorothy Thornton. "Explaining Corporate Environmental Performance: How Does Regulation Matter?" *Law & Society Review*, vol. 37, no. 1, 2003, pp. 51–90. Available at http://onlinelibrary.wiley.com/doi/10.1111/1540-5893.3701002/full. Accessed March 14, 2018.

Kagan, Robert A., and John T. Scholz. "The 'Criminology of the Corporation' and Regulatory Enforcement Strategies." In *Enforcing Regulation*, edited by K. O. Hawkins and J. M. Thomas, pp. 67–95. Boston, MA: Kluwer-Nijhoff, 1984. Available at https://www.researchgate.net/publication/289982560_The_Criminology_of_the_Corporation_and_Regulatory_Enforcement_Strategies. Accessed March 14, 2018.

Lee, Joanne, Peter Z. Schochet, and Jillian Berk. "The External Review of Job Corps: Directions for Future Research." Washington, DC: Mathematica, March 2018. Available at https://ideas.repec.org/p/mpr/mprres/376221bbee0d4b40bda431a169ad6a7a.html . Accessed May 22, 2020.

Levine, David I., Michael W. Toffel, and Matthew S. Johnson. "Randomized Government Safety Inspections Reduce Worker Injuries with No Detectable Job Loss," *Science*, vol. 336, no. 6083, 2012, pp. 907–911. Available at http://science.sciencemag.org/content/336/6083/907. Accessed October 10, 2018.

Locke, Richard, Matthew Amengual, and Akshay Mangla. "Virtue Out of Necessity? Compliance, Commitment, and the Improvement of Labor Conditions in Global Supply Chains." *Politics & Society*, vol. 37, no. 3, 2009, pp. 319–351. Available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1286142 . Accessed May 22, 2020.

Macallair, Daniel, and Mike Males. "A Failure of Good Intentions: An Analysis of Juvenile Justice Reform in San Francisco during the 1990s." *Review of Policy Research*, vol. 21, no. 1, 2004, pp. 63–78. Available at https://www.researchgate.net/publication/23999903_A_Failure_of_Good_Intentions_An_Analysis_of_Juvenile_Justice_Reform_in_San_Francisco_during_the_1990s. Accessed May 22, 2020.

MacKinnon, David P., Stefany Coxe, and Amanda N. Baraldi. "Guidelines for the Investigation of Mediating Variables in Business Research." *Journal of Business Psychology,* vol. 27, 2012, pp. 1–14. Available at https://www.researchgate.net/publication/228079846_Guidelines_for_the_Investigation_of_Mediating_Variables_in_Business_Research. Accessed May 22, 2020.

Mowbray, Carol. T., Mark C. Holter, Gregory B. Teague, and Deborah Bybee. "Fidelity Criteria: Development, Measurement, and Validation." *American Journal of Evaluation*, vol. 24, no. 3, 2003, pp. 315–340. Available at https://www.researchgate.net/publication/242079822_Fidelity_Criteria_Development_Measurement_and_Validation. Accessed May 22, 2020.

Parker, Christine. "The 'Compliance' Trap: The Moral Message in Responsive Regulatory Enforcement." *Law & Society Review,* vol. 40, no. 3, 2006, pp. 591–622. Available at http://onlinelibrary.wiley.com/doi/10.1111/j.1540-5893.2006.00274.x/full. Accessed March 14, 2018.

Patnaik, Ankita. "Exploring External Data to Enhance Monitoring and Evaluation of WHD's Compliance Strategies." Report submitted to U.S. Department of Labor. Washington, DC: Mathematica, January 14, 2020.

Rosenbaum, Paul and Donald Rubin. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika*, vol. 70, no. 1, 1983, pp. 41–55. Available at https://www.researchgate.net/publication/243082748_The_Central_Role_of_the_Propensity_Score_in_Observational_Studies_For_Causal_Effects. Accessed April 1, 2020.

Schochet, Peter Z. "Technical Methods Report: Statistical Power for Regression Discontinuity Designs in Education Evaluations (NCEE 2008-4026)." Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education, 2008. Available at https://www.researchgate.net/publication/234752602_Technical_Methods_Report_Statistical_Power_for_Regression_Discontinuity_Designs_in_Education_Evaluations_NCEE_2008-4026. Accessed May 22, 2020.

Sneed, Jeannie, Catherine H. Strohbehn, and Shirley A. Gilmore. "Impact of Mentoring on Food Safety Practices and HACCP Implementation in Iowa Assisted-Living Facilities." *Topics in Clinical Nutrition,* vol. 22, no. 2, 2007, pp. 162–174. Available at https://journals.lww.com/topicsinclinicalnutrition/Abstract/2007/04000/Impact_of_Mentoring_on_Food_Safety_Practices_and.8.aspx. Accessed March 14, 2018.

Sutton, Stephen. "Predicting and Explaining Intentions and Behavior: How Well Are We Doing?" *Journal of Applied Social Psychology,* vol. 28, no. 15, 1998, pp. 1317–1338. Available at https://onlinelibrary.wiley.com/doi/full/10.1111/j.1559-1816.1998.tb01679.x. Accessed March 14, 2018.

Tatian, Peter. "Performance Measurement to Evaluation." Washington, DC: Urban Institute, 2016. Available at https://www.urban.org/research/publication/performance-measurement-evaluation-0. Accessed April 2, 2020.

U.S. Bureau of Labor Statistics. "Quarterly Census of Employment and Wages." Washington, DC: U.S. Bureau of Labor Statistics, 2019. Available at https://data.bls.gov/cew/apps/table_maker/v4/table_maker.htm#type=0&year=2016&qtr=A&own=5&ind=722513&supp=0. Accessed January 2, 2020.

U.S. Department of Labor. "FY 2018 Annual Performance Report." Washington, DC: U.S. Department of Labor, 2018. Available at https://www.dol.gov/sites/dolgov/files/general/budget/2020/CBJ-2020-V1-01.pdf. Accessed May 22, 2020.

U.S. Department of Labor, Wage and Hour Division. "U.S. Department of Labor Launches Nationwide Campaign Focusing on Youth Working in Construction." Release Number 05-1382-NAT. Washington, DC: U.S. Department of Labor, 2005. Available at https://www.dol.gov/newsroom/releases/esa/esa20050718. Accessed November 20, 2019.

U.S. Department of Labor, Wage and Hour Division. "Defining and Delimiting the Exemptions for Executive, Administrative, Professional, Outside Sales and Computer Employees." Rule 84 FR 51230. Washington, DC: U.S. Department of Labor, 2019. Available at https://www.federalregister.gov/documents/2019/09/27/2019-20353/defining-and-delimiting-the-exemptions-for-executive-administrative-professional-outside-sales-and . Accessed November 20, 2019.

van Rooij, Benjamin, and Adam Fine. "How to Punish a Corporation: Insights from Social and Behavioral Science." New York: Program on Corporate Compliance and Enforcement, New York University School of Law, 2017. Available at https://wp.nyu.edu/compliance_enforcement/2017/09/01/how-to-punish-a-corporation-insights-from-social-and-behavioral-science/. Accessed March 14, 2018.

Wandersman, Abraham. "Four Keys to Success (Theory, Implementation, Evaluation, and System/Resource Support): High Hopes and Challenges in participation." American Journal of Community Psychology, vol. 43, no. 1-2, 2009, pp. 3–21. Available at https://onlinelibrary.wiley.com/doi/pdf/10.1007/s10464-008-9212-x. Accessed April 8, 2020.s

Weil, David. "If OSHA Is So Bad, Why Is Compliance So Good?" *RAND Journal of Economics*, vol. 27, no. 3, 1996, pp. 618–640. Available at http://www.fissuredworkplace.net/assets/Weil.If-OSHA-So-Bad.RAND.1996.pdf. Accessed March 14, 2018.

Weil, David. "Public Enforcement/Private Monitoring: Evaluating a New Approach to Regulating the Minimum Wage." *ILR Review*, vol. 58, no. 2, 2005, pp. 238–257. Available at http://journals.sagepub.com/doi/pdf/10.1177/001979390505800204. Accessed March 14, 2018.

Weil, David. "A Strategic Approach to Labour Inspection." *International Labour Review*, vol. 147, no. 4, 2008, pp. 349–375. Available at http://onlinelibrary.wiley.com/doi/10.1111/j.1564-913X.2008.00040.x/full. Accessed March 14, 2018.

Weil, David. "Rethinking the Regulation of Vulnerable Work in the USA: A Sector-based Approach." *Journal of Industrial Relations*, vol. 51, no. 3, 2009, pp. 411–430. Available at http://journals.sagepub.com/doi/pdf/10.1177/0022185609104842. Accessed March 14, 2018.

Weil, David. "Improving Workplace Conditions Through Strategic Enforcement." Boston University School of Management Research Paper No. 2010-20. Boston, MA: Boston University, 2010. Available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1623390 . Accessed March 14, 2018.

Weil, David. "Examining the Underpinnings of Labor Standards Compliance in Low Wage Industries: Final Report." Report to the Russell Sage Foundation, Grant RSF 85-08-06. Boston, MA: Boston University, 2012. Available at http://www.russellsage.org/sites/all/files/Weil.Final%20Report%202012.pdf. Accessed October 9, 2018.

Weil, David. *The Fissured Workplace: Why Work Became So Bad for So Many and What Can Be Done to Improve It*. Cambridge, MA: Harvard University Press, 2014. Available at https://www.researchgate.net/publication/278020172_The_Fissured_Workplace_Why_Work_Became_So_Bad_for_So_Many_and_What_Can_be_Done_to_Improve_it. Accessed May 22, 2020.

Weil, David, and Carlos Mallo. "Regulating Labour Standards Via Supply Chains: Combining Public/Private Interventions to Improve Workplace Compliance." *British Journal of Industrial Relations*, vol. 45, no. 4, 2007, pp. 791–814. Available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1028385 . Accessed May 22, 2020.

Weil, David, and Amanda Pyles. "Why Complain? Complaints, Compliance, and the Problem of Enforcement in the U.S. Workplace." *Comparative Labor Law & Policy Journal*, vol. 27, no. 59, 2005–2006, pp. 59–513. Available at https://www.researchgate.net/publication/267260153_Why_Complain_Complaints_Compliance_and_the_Problem_of_Enforcement_in_the_US_Workplace . Accessed May 22, 2020.

Weiss, Carol. "Nothing as Practical as Good Theory: Exploring Theory-Based Evaluation for Comprehensive Community Initiatives for Children and Families." In *New Approaches to Evaluating Community Initiatives*, edited by James Connell, Anne Kubisch, Lisbeth Schorr, and Carol Weiss. Washington, DC: Aspen Institute, 1995. Available at https://www.semanticscholar.org/paper/Nothing-as-Practical-as-Good-Theory-%3A-Exploring-for-Weiss/ed98a1ac4b7b54ef4854b7b7a802db7b3e46ae02. Accessed March 31, 2020.

Weiss, Michael, Howard Bloom, and Thomas Brock. "A Conceptual Framework for Studying the Sources of Variation in Program Effects." *Journal of Policy Analysis and Management*, vol 33, no. 3, 2014, pp. 778–808. Available at https://www.researchgate.net/publication/261190699_A_Conceptual_Framework_for_Studying_the_Sources_of_Variation_in_Program_Effects. Accessed May 22, 2020.

What Works Clearinghouse. "Confounding Factors." WWC Standards Brief. Washington, DC: U.S. Department of Education, Institute of Education Sciences, n.d. Available at https://www.ies.ed.gov/ncee/wwc/Docs/referenceresources/wwc_brief_confounds_101117.pdf. Accessed November 5, 2019.

Winter, Sidney G., and Gabriel Szulanski. "Replication as Strategy." *Organization Science*, vol. 12, no. 6, December 2001, pp. 730–743. Available at https://www.researchgate.net/publication/238337669_Replication_As_Strategy. Accessed May 22, 2020.

Yampolskaya, Svetlana, Teresa M. Nesman, Mario Hernandez, and Diane Koch. "Using Concept Mapping to Develop a Logic Model and Articulate a Program Theory: A Case Example." *American Journal of Evaluation,* vol. 25, no. 2, 2004, pp. 191–207. Available at https://www.researchgate.net/publication/230584099_Using_Concept_Mapping_to_Develop_a_Logic_Model_and_Articulate_a_Program_Theory_A_Case_Example. Accessed May 22, 2020.

This page has been left blank for double-sided copying.

Mathematica

Princeton, NJ • Ann Arbor, MI • Cambridge, MA
Chicago, IL • Oakland, CA • Seattle, WA
Tucson, AZ • Woodlawn, MD • Washington, DC

EDI Global, a Mathematica Company

Bukoba, Tanzania • High Wycombe, United Kingdom

mathematica-mpr.org