University of Denver
Digital Commons @ DU

Electronic Theses and Dissertations | Graduate Studies

2022

Examining the Credibility of Story-Based Causal Methodologies

Megan E. Kauffmann
University of Denver

Follow this and additional works at: https://digitalcommons.du.edu/etd

 Part of the Other Statistics and Probability Commons, and the Statistical Methodology Commons

Recommended Citation

Kauffmann, Megan E., "Examining the Credibility of Story-Based Causal Methodologies" (2022). *Electronic Theses and Dissertations*. 2026.
https://digitalcommons.du.edu/etd/2026

This Dissertation is brought to you for free and open access by the Graduate Studies at Digital Commons @ DU. It has been accepted for inclusion in Electronic Theses and Dissertations by an authorized administrator of Digital Commons @ DU. For more information, please contact jennifer.cox@du.edu,dig-commons@du.edu.

Examining the Credibility of Story-Based Causal Methodologies

A Dissertation

Presented to

the Faculty of the Morgridge College of Education

University of Denver

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

by

Megan E. Kauffmann

March 2022

Advisor: Bruce Uhrmacher, Ph.D.

©Copyright by Megan E. Kauffmann 2022

All Rights Reserved

Author: Megan E. Kauffmann
Title: Examining the Credibility of Story-Based Causal Methodologies
Advisor: Bruce Uhrmacher, Ph.D.
Degree Date: March 2022

ABSTRACT

The purpose of this study was to explore how evaluators justify using story-based methodologies when examining causality. The two primary research questions of the study included: 1) what arguments are made by evaluators to justify the credibility of story-based causal methodologies to evaluation stakeholders; and 2) from the perspective of evaluators, how do contextual factors influence whether story-based causal methodologies are perceived as credible by evaluation stakeholders? A case study was conducted to examine the cases of four evaluators who had experience implementing a story-based methodology in an evaluation. Data collection procedures included two interviews with each participant and a review of materials related to their story-based evaluations. Analysis of the data revealed cross-case themes highlighting participants' arguments for how they justify using story-based methodologies for causal examination. The first argument was that these methodologies can credibly be used to examine causality when the evidence needed aligns with the type of evidence produced by these methodologies. The second argument was that these methodologies can add credibility to a causal study because they reduce evaluator bias by elevating participant voice in the data collection process. The third argument was that participants can generally be trusted to provide true accounts; thus, their accounts of how change occurred for them are credible

forms of evidence. And finally, the fourth argument was that these methodologies include procedures to triangulate participant accounts with other sources of data, thereby enhancing the credibility of the evidence produced. The study also included an examination of contextual factors that may contribute to these methodologies being perceived as credible by evaluation stakeholders. Findings revealed that stakeholders in the following contexts might be more likely to perceive story-based methodologies as credible: learning contexts, low-risk contexts, multi-cultural contexts, and contexts that value centering participant voice. As a contribution to the evaluation field, this study also provided a practitioner's guide for evaluators needing to justify using story-based methodologies for causal examination.

ACKNOWLEDGMENTS

Many people supported me throughout this work and I am grateful for all of their assistance. Thank you to Dr. Bruce Uhrmacher for all of the support and guidance as my dissertation director. Thank you as well to my committee members Dr. Nicholas Cutforth and Dr. Paul Michalec for the thoughtful feedback.

I would also like to thank Dr. Antonio Olmos and Dr. Robyn Thomas Pitts for their guidance on this work. Thank you as well to the study participants for taking part in the study. Finally, I want to thank my friends and family for their support along the way.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER I: INTRODUCTION

Background

Directors of small social programs want to demonstrate that their programs lead to intended outcomes. Within the social science quantitative tradition, the design deemed most credible for creating strong evidence for causal inference is the randomized controlled trial (RCT). While the RCT can be utilized with small social programs, a number of issues may arise when attempting an RCT in small social program settings. This chapter explores the main issues encountered by evaluators seeking to implement RCTs with small social programs. This chapter also provides additional context to frame the widely held view that RCTs are the most credible research design to demonstrate evidence of causality within the quantitative tradition.

The quantitative research design seen to produce the strongest evidence of causality in social science is the RCT (Gao et al.2017; Scriven, 2008). At the beginning of an RCT, control and treatment groups are formed and measured on a set of variables that an intervention is trying to impact. The treatment group receives the intervention while the control group doesn't. After the intervention, the treatment group and control group are measured on those variables again. Any statistically significant difference in the average score between the two groups on those variables is deemed to be due to the intervention. By creating a control group, the researcher is estimating what would have

happened if the group of people receiving the intervention had not received it, a concept that is referred to as the counterfactual (Gates & Dyson, 2017). The counterfactual is only conceptual; it is not physically possible for a group to receive and not receive an intervention at the same time. (Shadish et al.2002). Instead, to approximate the counterfactual researchers create a control group that is probabilistically similar to the treatment group on key variables (Shadish et al., 2002). Then, researchers compare effects from the control group to those from the treatment group.

The RCT design has its foundation in the philosophy of John Stuart Mill (Shadish et al., 2002). In the 19th century, Mill's theory of causal relationships stated that three conditions needed to be met to make an argument that causality was present (Shadish et al., 2002). These three conditions were: 1) the cause had to occur before the effect (temporal precedence); 2) the cause had to be related to the effect; and 3) there should not be any plausible alternative explanations for how the effect occurred. Mill's theory of causal relationships is the cornerstone for how causal inquiry is approached within the social sciences. In examining the structure of the RCT design, it is evident how Mill's theory has been applied: RCT designs include a requirement of temporal precedence, correlation between cause and effect, and examination of other plausible explanations. The process of meeting the third condition can also be known as addressing threats to internal validity (Shadish et al., 2002).

Internal validity refers to the ability to demonstrate that variations in a treatment can be attributed to changes in an outcome (Shadish et al., 2002). This type of validity is the most relevant to causal inference. The strength of evidence for causal inference is

higher the more that the threats to internal validity are addressed (Shadish, et al., 2002). Gliner et al. (2016) discuss two groups of threats that need to be addressed to strengthen evidence of internal validity: 1) differences in participant characteristics prior to an intervention, and 2) extraneous or confounding variables that occur during the course of an experiment that could offer alternative plausible causal explanations. The RCT is considered to be the strongest quantitative research design because the design component of random assignment serves to rule out the first group of internal validity concerns.

Random assignment means that there is a randomization of individuals with equal likelihood of receiving an intervention into treatment and control groups (Shadish et al., 2002). This randomization ensures that the differences that were present in the two groups prior to the intervention were due to random chance and not due to the researcher's systematic selection bias (Shadish et al., 2002). Because random assignment effectively manages the first group of threats to internal validity (pre-intervention differences in groups), RCTs are considered the strongest quantitative causal research design. As for the second group of internal validity threats (extraneous factors that occur during the course of the experiment), the RCT is strengthened when these extraneous factors can be addressed, controlled for, or explained as non-significant through a process of applying reason, logic, and statistical correction (Shadish et al., 2002).

The RCT is a useful quantitative causal research design to identify social programs that are effective. However, the RCT design components require specific contextual and environmental characteristics to be in place for the design to be implemented; these are explored fully in the next section. These characteristics are not

present in certain types of social programs, particularly small social programs (see Definitions section for a definition of small social programs). The next section explores issues that evaluators encounter when attempting to conduct an RCT design with small social programs.

Problem Statement

Given that RCTs are the preferred method to demonstrate causal inference within the quantitative social sciences, it is understandable that small social programs would request the design from evaluators. However, when an evaluator tries to implement an RCT with a small program, they may run into barriers. This section details several issues that emerge when attempting to implement an RCT with small social programs, including: 1) the cost may be prohibitive; 2) small sample sizes within small social programs may pose a challenge; 3) small social programs may be unwilling to withhold services to a control group; 4) small social programs may have a strong interest in the "how" or "why" an intervention is working (information a conventional RCT may not provide); and, 5) the complex and dynamic nature of small social programs may pose a challenge for linear analytic methods. Each of these issues is explored in depth below.

RCTs May Be Cost Prohibitive

RCTs may be more costly than other evaluation designs (Azzam & Christie, 2007; Christie & Fleischer, 2010; Gao et al., 2017; Linden et al., 2006). This is because of the expenses needed to maintain a large sample size and provide incentive payments to participants. Some scholars state that a large sample size is crucial to achieve the statistical power required in an RCT (Hawkins, 2016; Scriven, 2008). Recruiting and maintaining a

large sample size through financial incentives to participants adds to the cost of the design (Scriven, 2008). Additionally, the researcher expertise needed to design and implement an RCT can also be expensive. The scale of these expenses is reflected in the Corporation for National Community Service's (n.d.) estimate that a typical RCT costs more than 25% of the entire project budget. Given that small social programs are often operating on a limited budget that prioritizes service provision, dedicating 25% of that budget to an evaluation is likely not feasible.

However, it is important to note that the expense of an evaluation can vary widely depending on the choices made by the evaluator and program. For example, one may be able to conduct an RCT with a sample size as small as 60 (Glass & Hopkins, 1996), which would reduce the costs of incentives. Additionally, some methodologies that may generally be considered to be less expensive than the RCT may not actually be so in practice. For example, one may be conducting a mixed methods study including a quantitative pre/post survey and qualitative interviews at the end of an intervention to gather feedback on the program. Within a mixed methods study, one must follow standards of rigor not just for quantitative methods but also for qualitative methods (Creswell & Plano Clark, 2017). This may prove to be costly as the time spent on procedures for rigor may be greater than it would have been with only one method utilized. Additionally, qualitative methods can also prove to be costly as qualitative research can require many hours of labor to conduct interviews, transcribe the data, code the data, and write the analysis while attending to standards of rigor (A. Olmos, personal communication, September 29, 2020).

Small Sample Sizes May Pose Challenges

The definition utilized for this study for small social programs includes any programs with fewer than 100 participants. According to Glass and Hopkins (1996), a sample size of at least 30 for both the treatment and control groups is necessary for statistical inferences from an RCT to be valid. Thus, for small social programs with 60 or more participants, RCTs may be feasible. For programs with a smaller population than that, the RCT would not likely be feasible. For programs with just above 60 participants, the RCT may also not be feasible as the program (and/or the RCT study) may experience attrition, bringing the sample size to a number too low to conduct the study with validity.

Small Social Programs May Be Unwilling to Withhold Services

The random assignment procedures in an RCT require the creation of a control group that does not receive services. A number of evaluation theorists and researchers state that, in some circumstances, withholding treatment is unethical (Assam & Christie, 2007; Boruch, 2007; Hawk, 2015). For example, imagine a small social program is distributing free vaccines within their community via a medical van. It would be unethical for them to attempt to achieve randomization by withholding services from eligible community members who approach the van for a vaccine. Some authors state that withholding services is immoral; Boruch (2007) writes that RCTs violate human rights when some groups receive treatment and others don't. Azzam and Christie (2007) also state that it is unethical to withhold services to a group that could benefit. For small social programs operating within their communities, one can picture the ethical

quandaries that could emerge when attempting to select who among friends, family, and neighbors should receive treatment.

However, a quasi-experimental design called the wait-list design may be able to address these ethical concerns. In an evaluation of an HIV-prevention intervention among African American women, Hawk (2015) and the community partner chose a wait-list design. In a wait-list design, the treatment group receives the treatment first during the same period of time when the comparison group does not receive the treatment; after the period ends, the comparison group then receives the treatment. Thus, the wait-list quasi-experimental design is an alternative for organizations that do not want to implement an RCT due to the ethical concerns of withholding treatment. It may be feasible for small social programs to implement this quasi-experimental design. However, there may be issues that arise for evaluators if is not possible to delay services for political reasons (for example, if it is perceived that some groups are being favored by receiving treatment first) or for ethical reasons (for example, if the treatment is needed to save lives, and decisions about who receives it should be based on criteria that would be difficult to randomize).

Small Social Programs Are Interested in the "How" and "Why"

RCTs are effective at informing programs about whether there is evidence for causal inference. However, the conventional RCT that does not include a qualitative component cannot explain how or why the program is working. Oftentimes, small social programs are in need of qualitative evidence demonstrating how or why a program is working in order to improve programming. Chen (as cited in Vingilis & Pederson, 2001)

states that RCTs can identify that an intervention had a significant effect for a treatment group but they cannot explain the mechanisms of how the program brought about the desired change. Christie and Fleischer (2010) concur with Chen, stating that RCTs may not provide the kind of information that programs need to make improvements. If a RCT were to include a line of qualitative inquiry to uncover what elements of an intervention triggered change, small social programs could intentionally adjust their programming to achieve more consistent results. Additionally, the context in which an intervention occurred may not always stay the same. As Hawkins (2016) states, "RCTs may show us how a program or intervention was useful in a past context, but this context will never again occur" (p. 277). Qualitative inquiry may be able to explain what contextual factors were important in triggering the causal chain of events within a particular place and time and with a particular group of people. This information may be important for small social programs to be aware of as they continuously improve their programming.

Small Social Programs are Dynamic, Complex, and Evolving

Small social programs often implement complex, evolving, and dynamic interventions; this may pose challenges for evaluators implementing RCTs. Chatterji (2007) argues that programs in the social sector don't operate in the same way as medical interventions; rather, they operate in dynamic, complex systems. As such, within social science the units of randomization are seldomly chosen correctly (Chatterji, 2007). Lehmann (2015) provides an example of the challenge of applying a single-cause model within the complex setting of a school-based intervention. Lehmann illustrates that school performance is impacted by many complex factors, such as parental employment,

parental stressors, and substance use, among others. If one were to design a study examining a causal link between only one of those factors and school performance, significant results could lead to the development of interventions targeting that one cause (Lehmann, 2015). Lasting impact would not be attained because of the complex interrelation of factors leading to school performance (Lehmann, 2015).

Mason et al. (2018) similarly argue that an RCT approach is not preferred in applied service settings where a web of individuals may be impacted. The authors state that “the applied service setting encourages multicomponent and adaptive interventions that can generate desired changes within a heterogeneous population of individuals, often involving intact families, schools, peer networks, and communities" (Mason et al., 2018, p. 164). The complex causal mechanisms within social services interventions, which sometimes involve entire communities interacting in synergistic ways, may be difficult to understand using the linear cause and effect modeling of the RCT.

Finally, because small social programs operate in dynamic systems, the program theory behind their services may need to remain dynamic and evolving in order to respond to stakeholders’ needs. In these scenarios, it may mean that the program theory does not reach the stage of stability necessary for an implementation fidelity model to be developed. A stable program theory and implementation fidelity measures must be in place to carry out an experimental design. As Rossi et al. (2019) state, an experimental design cannot be implemented when there are not sufficient ways to monitor implementation fidelity.

An Alternative to the RCT: The Quasi-Experimental Design

The previous section has illustrated issues that evaluators may encounter when attempting to implement an RCT with a small social program. It is worth noting that the quasi-experimental design (QED) can address many of these issues. The quasi-experimental design is similar to the RCT in that John Stuart Mill's three criteria for causality must be met (Shadish et al., 2002). The key difference between the RCT and the QED is that within the RCT units are randomly assigned to treatment and condition groups, whereas within a QED there is no random assignment (Shadish et al., 2002). In many interventions or treatment situations, there is non-random "self-selection" into groups (individuals have chosen their own conditions) or non-random "administrator selection" (program administrators have selected the conditions individuals receive) (Shadish et al., 2002, p.14). These situations still provide opportunities to compare groups, but random assignment is no longer possible. A QED can be utilized in these situations because a comparison group is created to serve a similar function as a control group would (Azzam & Christie, 2007). However, because the comparison group is not created through random assignment, there are more threats to internal validity within a QED then there are within an RCT (Shadish et al., 2002). To account for this, it is recommended that the researcher gain an understanding of which alternative explanations are plausible and then systematically rule out whether these explanations could be causing the effect (Shadish et al., 2002).

Some features of QEDs make these designs compatible with the contexts of small social programs. As QEDs can often be done with observational data and do not include

random assignment, they can be done with less expense than an RCT (Linden et al., 2006) and can be completed in a shorter time frame. Additionally, no one is denied treatment for the purposes of an experiment within a QED, alleviating some ethical concerns. Treatment-only quasi-experimental designs, in which every participant receives the treatment, may work particularly well to address ethical concerns in small social program settings. These include: 1) the one-group pretest-posttest design, 2) the one-group pretest-posttest design using a double pretest; and 3) the one-group pretest-posttest design using a nonequivalent dependent variable (Shadish et al., 2002). In these designs, the same group of individuals are compared at pre- and post-test.

Thus, a quasi-experimental design may be an option when evaluators cannot conduct a RCT in small social program settings. Cost is less of a concern because QEDs can be conducted using existing data. Additionally, the ethical concerns that sometimes arise with withholding services in a RCT are not present within QEDs that utilize treatment-only designs. However, two issues that arise when using RCTs with small social programs still arise when using QEDs. Similar to RCTs, QEDs are not compatible with dynamic, changing contexts; Marchal et al. (2012) state that because QEDs follow a linear model of analysis they don't work well to evaluate complexity. Additionally, QEDs without a qualitative component (just like RCTs without a qualitative component) cannot explain the "how" or "why" behind a causal change.

Central Problem Addressed by this Study

As detailed in the previous section, evaluators may encounter a number of issues when attempting to implement a RCT with a small social program, including: costs may

be prohibitive, sample sizes may be too small, withholding services may be unethical, information about the "how" or "why" a program worked may not be uncovered, and the complexity of the small social program may not fit with a linear model of evaluation. The quasi-experimental design may resolve the issues of prohibitive cost and the withholding of services in some cases. However, there are still some scenarios in which evaluators will encounter too many of these issues and will not be able to implement a RCT or a QED. Consequently, small social programs in these scenarios face two main problems: 1) they have no recognized credible alternative to the RCT or QED to demonstrate that their programs are working for beneficiaries; and, 2) they may not be competitive for funding if they cannot show evidence of program effectiveness.

However, alternative causal evaluation designs do exist. Gates and Dyson (2017) discuss the success case method and most significant change methodologies, which are two designs under the narrative approach to evaluating causality. I have termed the success case method and the most significant change methodologies "story-based causal evaluation methodologies" because both include the collection of individual accounts as the primary form of data collected (Brinkerhoff, 2003; Dart & Davies, 2003). These methodologies may be more feasible for small social programs because collecting individual accounts may be less expensive than the data collection procedures included in experimental designs, which often include incentive costs for a larger sample of participants. Secondly, these methodologies may be more feasible for small social programs because they can be implemented with small (<60) sample sizes, whereas experimental designs cannot. Thirdly, because these methodologies capture participants'

stories, they capture the "how" and "why" change occurred from participants' perspective; this is evidence that small social programs need in order to make changes to enhance the effectiveness of their programs.

Despite story-based evaluation methodologies being feasible for small social programs, they may not be seen as credible by evaluators for use in causal inquiry. Part of the reason these methodologies may not be seen as credible is because of what evaluators refer to as the "hierarchy of methods"; within the hierarchy, RCTs are seen as the top method to evaluate causality, and other methods are seen as less credible (Gates & Dyson, 2017, p. 31). Thus, the central problem this research study addresses is the need to raise awareness of the credibility of story-based causal evaluation methodologies in the eyes of evaluation stakeholders.

Purpose of Study and Research Questions

The purpose of this study is to examine how evaluators who have practiced story-based causal methodologies justify their use within evaluation studies that were intended to produce evidence of causality. The purpose of this study responds to Gates and Dyson's (2017), suggestion that "future discussion within the evaluation community is needed regarding how evaluators construct relevant and defensible causal arguments and how evaluators can justify the causal approach taken to multiple audiences" (p. 43). Audiences or stakeholders in evaluation typically include the following groups: evaluation clients, donor or funders, program beneficiaries, program implementers, program participants, and other/peer evaluators. The first research question for this study maps directly onto this suggestion from Gates and Dyson. The second research question

expands this line of inquiry to consider how relational, contextual, or environmental factors may influence perceived credibility of story-based causal methodologies. The following two research questions are the central research questions for this study:

RQ1: What arguments are made by evaluators to justify the credibility of these story-based causal methodologies to evaluation stakeholders?

- Do the arguments that evaluators make to justify the credibility of these methodologies differ depending on which evaluation stakeholder is the audience for the argument? If so, how do they differ?

RQ2: From the perspective of evaluators, how do contextual factors influence whether story-based causal methodologies are perceived as credible by evaluation stakeholders?

Significance of Study

While story-based causal methodologies are a strong design choice for small social programs seeking to demonstrate effectiveness of their program, these designs may not be seen as credible alternatives to the RCT because they are not grounded in the quantitative social science traditions of causal research. Thus, it is important to raise awareness of the credibility of story-based causal methodologies among evaluators and evaluation stakeholders. This study articulates the arguments made by evaluators to justify the use of story-based methodologies in causal evaluation. In Chapter V, these arguments are integrated with the literature to develop a practitioner's guide for evaluators that are working in contexts in which they will need to justify why story-based methodologies are credible for use in causal research. This practitioner's guide is useful

for evaluators who believe that story-based causal methodologies are the best fit for the evaluation context, but need to justify their use with evaluation stakeholders.

Definition of Terms

The following key terms are used throughout this study.

- Causal mechanism –an idea or opportunity that is introduced into a context that triggers a causal reaction (Pawson &Tilley, 1997).
- Causal ways of thinking – Gates and Dyson (2017) present multiple ways of thinking about causality. Another way to refer to a way of thinking about causality is to refer to it as a causal view.
- Credibility – There is no one definition for credibility in the literature as it applies to causal evaluation. *The Program Evaluation Standards* (Yarbrough et al., 2010) outline a set of standards around utility, feasibility, propriety, accuracy, and accountability which speak to different facets of the credibility of evaluation studies. Within qualitative research, the term credibility refers to the trustworthiness of research findings (Anfara et al., 2002). This definition, although crafted to apply specifically to qualitative research, can also be utilized to discuss credibility in causal evaluation generally. For the purpose of this study, the term credibility refers to the extent to which evidence can be trusted.
- Internal validity – this facet of validity refers to the ability to attribute variations in a treatment to changes in an outcome (Shadish et al., 2002). This facet of validity is the most relevant to causal inference (Shadish, et al., 2002). Two groups of threats to internal validity are characterized by Gliner et al. (2016):

threats referring to group characteristics before an experiment and threats referring to extraneous factors occurring during an experiment.

- Randomized controlled trial (RCT) – the RCT is the only true experimental design (Shadish et al., 2002). It begins with a randomization of individuals with equal likelihood of receiving an intervention into treatment and control groups. This action of random assignment ensures there is no threat of selection bias. RCTs must also include a design element to control for the effect of extraneous variables (Gliner et al., 2016).
- Small social programs – A small social program is one that provides a focused intervention to small group of people, often in cohorts, in small doses and with a small budget. While there is no set standard for what constitutes "small," for the purpose of this study "small" refers to any program that serves at any given time a population size that may be too small to be used as a sample for population statistics (<100). Typically, these programs target a particular group of people who are experiencing a problem or a set of problems, and the intervention is intended to ameliorate those problems. The program may be run by a non-profit or may be a small initiative piloted by a county, state, or federal agency. Some examples of these types of programs include educational pilot programs offered to small cohorts in a school, after-school programs for at-risk youth, short-term training seminars for adults, or summer programs. Westhorp (2008) defines "small community-based" programs as:

> [Programs that] generally offer a particular sub-set of programs, often a combination of information, advice and referral, education or skills development, personal development, and 'support'... They typically offer those services in a particular sub-set of contexts: small organizations (or 'projects' managed by larger organisations), located in small, often stand-alone buildings such as houses or shopfronts *[sic]*, relatively widely geographically dispersed, often with only a handful of staff on site, and often (but not exclusively) staffed by a combination of professional, para-professional and volunteer workers (pp. 77-78).

Westhorp aptly defines small social programs as providing services in specific contexts particular to their communities and relying upon a small number of staff and volunteers; these characteristics are also used to define small social programs for the purpose of this study. Additionally, for the purpose of this study the term "small social programs" also includes smaller components of larger programs. An example of a smaller component of a larger program may include a training program for teachers in which there are several models of training being offered to a large group of teachers, but the number of teachers engaged in each training model is small. Another example may be if a subgroup of individuals is engaged in an intervention and an evaluator wants to specifically examine the experience of the smaller subgroup rather than the entire large group receiving the intervention. Smaller components of large studies are similar to small social programs because they have a smaller sample size and often may be dynamic or complex in nature, particularly if they are pilot programs.

- Story-based causal evaluation methodologies: Throughout this study, success case method and most significant change are referred to as story-based causal evaluation methodologies because they both utilize individual accounts to detail how change

occurred for an individual within their context (Brinkerhoff, 2003; Dart & Davies, 2003).

CHAPTER II: LITERATURE REVIEW

The literature was reviewed to determine how evaluators have justified the credibility of story-based causal evaluation methodologies in their published work. The first section of the literature review presents the concept of causal pluralism. This is the idea that there are multiple ways of thinking about causal evaluation (Gates & Dyson, 2017). The second section highlights the advantages of conducting causal evaluation within the narrative tradition. The advantages include assigning more value to the individual's account of what causal factors are at play, being able to describe in detail the individual's motivation, and capturing the context in which causal change occurs. The third section summarizes the key arguments that evaluators have made to justify that story-based methodologies are credible in causal evaluation However, there were not many published works that discussed the arguments that evaluators make; thus, this gap in the literature is discussed at the end of the chapter.

Different Ways of Thinking About Credibility in Causal Evaluation

There are several methodological resources that outline what the terms validity and credibility mean within the social sciences (Anfara et al., 2002; Creswell & Miller, 2000; Gliner et al., 2016; Shadish et al., 2002). Within the quantitative tradition, research validity refers to the quality of the entire study (Gliner et al., 2016). Within the qualitative tradition, the word credibility replaces the word validity (in recognition that not all

quantitative concepts transfer neatly into qualitative concepts), and refers to the degree to which findings from a study can be trusted (Anfara et al., 2002). Both terms refer to whether or not the findings/results of a study can be utilized to inform decisions and practice.

There are procedures within both the qualitative and quantitative traditions that enhance the validity or credibility of a particular study. However, these procedures are tradition-specific, and do not apply to all causal evaluation as a whole. While the set of procedures to increase evidence of internal validity within a RCT is well-known (Gliner et al., 2016), there are not well-known procedures for increasing evidence of credibility or validity of causal evaluation in general (across the quantitative, qualitative, or mixed methods traditions). The RCT design and quasi-experimental designs have a set of procedures to enhance evidence of internal validity, a long history of acceptance of their credibility, and a basis in philosophical thought (Shadish et al., 2002). Story-based causal evaluation methodologies do not have such a set of procedures or a history of acceptance; however, there is burgeoning philosophical and rational thought justifying their use in evaluation that produces evidence of causality.

Several prominent evaluators acknowledge that the RCT is not the only research design that can yield credible evidence of causality. Patton (2015), Scriven (2008), and Shadish et al. (2002) offer the "common sense" argument: some cause and effect relationships are obvious to us if we simply observe. We do not, for example, need to conduct a RCT to understand that an egg fries when hitting a hot pan, as Scriven (2008) notes. In line with this common sense reasoning, Scriven states that critical observation is

the best method to determine that a causal relationship exists, not the RCT. Gao et al. (2017) also make an argument that some qualitative research approaches, such as Mohr's causal reasoning approach and Scriven's modus operandi approach, are also credible ways to investigate causality. Gao et al.'s argument welcomes the possibility that there are different credible ways of thinking about causality. Gates and Dyson (2017) advocate for an acceptance of multiple ways of thinking about causality; this approach may be best described as causal pluralism. They also present a framework to summarize and communicate recent thinking about this phenomenon.

The Framework of Causal Pluralism

Gates and Dyson's (2017) work on causal pluralism presents a comprehensive summary of contemporary thinking on divergent ways of thinking about causality. Gates and Dyson reflect that the debate over the philosophy of science and how causal claims are made is alive and well and believe that the question of how to warrant causal claims currently vexes the evaluation field. The conventional way of thinking about causality (quantitative experimental designs are the only designs that can be utilized to provide evidence of causality) has recently been challenged by growth in contribution analysis, theory-based approaches, and non-linear, systems-based modeling, according to the authors. Gates and Dyson embrace and promote a more flexible way of thinking about causality that can be thought of as causal pluralism. Like other types of pluralism, causal pluralism advances the co-existence and equality of different types of causal views and rejects the hierarchy of one type of causal view over the other (Gates & Dyson, 2017).

It is worth noting that the view of causal pluralism shares commonalities with the concept of dialectical pluralism (Johnson, 2017). Johnson states that dialectical pluralism "recommends that one concurrently and equally value *multiple* perspectives and paradigms" (p. 159). In accordance with dialectical pluralism, evaluators operating from different paradigms should converse and expand on each other's thinking, particularly in mixed methods research settings where both the qualitative and quantitative components are to be valued as equal (Johnson, 2017). By emphasizing that different perspectives should be valued equally, dialectical pluralism has much in common with causal pluralism. Both perspectives acknowledge that qualitative methods can produce credible evidence in research contexts where quantitative methods may have historically been considered the only acceptable method.

Five different causal views are defined within the Gates and Dyson (2017) causal pluralism framework: the successionist, narrative, causal package, generative, and complex systems views. Each of the views is substantiated by a distinct line of reasoning that is summarized in this paragraph; the following paragraphs provide comparisons and contrasts between the views. The successionist view, according to the authors, holds that a cause is hypothesized to precede an effect, have a relationship with that effect, and be the likely only explanation to why that effect occurred (it is the view that supports the causal logic behind the RCT). The narrative view affirms that the narrative we create as human agents about how change occurs is valid. The causal package view holds that causality occurs in packages of factors, rather than individual factors. The generative view holds that there are multiple causal pathways at play and that the causal mechanism

within these pathways is triggered only if the right context and actors are present. Finally, the complex systems view holds that causal pathways may be non-linear and may take the form of feedback loops. Table 1. provides a summary of these views and the particular research approaches and methodologies associated with each causal view, according to the authors.

Table 1.
Causal Ways of Thinking Aligned with Approach and Methodology

Way of Thinking	Design Approach	Methodology
Successionist	Experimental	RCT, natural experiments, quasi-experimental.
Generative	Theory based	Realist evaluation, process tracing, contribution analysis, impact pathways analysis.
Narrative	Participatory	Success case method, most significant change, outcome mapping.
Causal Package	Case based	Within case: analytic induction, network analysis, and process tracing. Across case: qualitative comparison case analysis.
Complex Systems	Systems based	Causal loop diagramming, system dynamics.

A crucial concept that must be clarified when discussing the five causal views is that each causal view answers a different causal question. As such, each view can contribute a different angle of explanation and discovery of the causal story. As Gates and Dyson (2017) state, an evaluator who is investigating causality should construct "relevant and defensible causal arguments" (p. 29). A relevant and defensible causal argument may also include an explanation of "causality at multiple levels" (Gates &

Dyson, 2017, p. 29). Because the causal views align with different causal questions, drawing from more than one view (and thus answering more than one causal question) may provide us with a richer understanding of the full causal story. Table 2. details the questions that that align with each causal view.

Table 2.
Causal Views and Associated Evaluation Questions

Causal View	Evaluation Questions
Successionist	• What effects are statistically significantly associated with this intervention? • Does the intervention work to produce intended effects? • Can we attribute effects to the intervention?
Generative	• What works, how, for whom, and under what circumstances? • How and why does the intervention work?
Narrative	• According to stakeholders, what influence, effects, and/or difference did the intervention make for their lives?
Causal Package	• Is it likely that *(sic)* intervention has made a difference? • How does the intervention work in combination with other interventions or factors to make a difference?
Complex Systems	• How do multiple causal factors and feedback processes affect change in this intervention or situation? • What's working now and how?

Adapted from "Implications of the Changing Conversation About Causality for Evaluators," by Gates, E. and Dyson, L., 2017, *American Journal of Evaluation, Vol. 38*, p. 37.

Understanding that each causal view is aligned with particular causal questions alleviates some of the tension that has come about in what Scriven (2008) describes as the causal wars. Scriven states that "the causal wars are about what is to count as

scientifically impeccable evidence of a causal connection, usually in the context of the evaluation of interventions into human affairs" (p. 11). The argument in the causal wars is between evaluators who claim that the RCT is the only acceptable design to provide evidence of causality and other evaluators who disagree (Scriven, 2008). The causal wars were still active in the late 2010s (Gao et. al, 2017; Gates & Dyson, 2017). Taking a causal pluralism stance may help resolve the wars because within causal pluralism the methodologies aligned with each causal view provide evidence that answers specific evaluation questions related to different dimensions of causality that are not in competition with one another.

Differences and Commonalities in Ways of Thinking About Causality

The causal views share some differences and commonalities. One major area of difference is the level at which an outcome is measured. Within the successionist view, the outcome being measured is at the population level (Gates & Dyson, 2017). Within the generative view, the outcomes are measured at the sub-group level because this view is concerned with what works, for whom, and in what circumstances (Gates & Dyson, 2017). The what works, for whom and in what circumstances guiding question can be answered by conducting sub-group analyses based on groups of characteristics, as is done in realist evaluation (Pawson & Tilley, 1997). However, within the narrative view, the outcomes are measured at the individual level, because an individual human agent has made choices and actions that changed their life (Abell, 2004). As another area of difference, outcomes in the complex systems view are observed at the system level; outcomes are seen as effects that are observed as a system changes (Gates & Dyson,

2017). Within the causal package view, outcomes can be measured at either the population, sub-group, or individual level (Gates & Dyson, 2017).

Another major area of difference between the causal views concerns the concept of context. Within the successionist view, changes in one causal variable within a model are meant to change the outcome variable in a linear fashion, and contextual factors are controlled to make this relationship more evident (Shadish et al., 2002). This line of thinking does not apply in the narrative, causal package, complex system and generative views, which hold that contextual variables are not separable from causal factors (Gates & Dyson, 2017). For example, within the narrative view context is a set of factors that influence whether an intervention is effective (Gates & Dyson, 2017). Similarly, the causal package view holds that social programs are steeped in environments of complexity in which multiple interventions are co-occurring; in these environments, researcher control over what participants experience is limited (Gates & Dyson, 2017). Additionally, theory-based approaches within the generative view hold that the causal mechanism is triggered only if the right context and people are present; thus, context is an important causal agent in these approaches as well (Gates & Dyson, 2017). Finally, the complex system view holds that causal relationships are dependent on the context (Gates & Dyson, 2017). Thus, the role of context is an important point of difference among the causal views.

However, the views also share some commonalities. The generative view and the narrative view both hold that human agency – human actions and motivation – is a major causal factor. In their description of generative causation, Pawson and Tilley (1997) point

out that the generative view focuses on internal as well as external causes behind a change. Thus, a generative view of causation would require thinking about what external inputs made the intervention work *and* thinking about what internal powers or agency inside of participants triggered the intervention to work (Pawson & Tilley, 1997). Within the generative view, it is not an intervention by itself that "works," it is the actions of individuals and the context they are immersed in that makes an intervention work (Pawson & Tilley, 1997). The narrative view aligns with the generative view on this point; humans are key drivers of change in the narrative view as well (Abell, 2004). Human agency might not play as large a role in explaining causation within the causal package, successionist, and complex systems views, which tend to emphasize sub-group, population or system-level dynamics (Gates & Dyson, 2017). These examples of the commonalities and differences between the causal views highlights that practicing from a causal pluralism lens requires an understanding of the assumptions and key ideas of each view.

The Credibility of Story-Based Methodologies

This section synthesizes the literature addressing the credibility of the narrative way of thinking and the story-based causal evaluation methodologies aligned with this way of thinking. Themes that emerged from the literature include: 1) narrative ways of thinking are credible; 2) story-based causal evaluation methodologies reveal unexpected outcomes; 3) they increase participant voice and cultural validity; 4) they are transformative; 5) they focus on successful parts, rather than the successful whole; and, 6) and they include corroboration of evidence. While these themes were derived from

existing literature addressing the topic of credibility in story-based causal methodologies, this literature was sparse and this is the gap in the literature that is addressed through the study.

Narrative Ways of Thinking about Causality are Credible

A number of authors argue that the narrative way of thinking about causality is credible. *Narrativists* (a term coined by Abell, 2004) believe that human agency is the main reason why conditions change (Abell, 2004). Impact comes about through the individual actions of people featured in the story. Abell explains that narrative inquiry can draw out how individual motivations brought about change. Thus, one argument for the credibility of the narrative way of thinking about causality is that it captures the stories of the individuals who are the key drivers of change.

Stories also describe in detail the contexts in which change occurs, providing information about the causal mechanisms that interacted with human behaviors to effect change. Stories are typically a chronological account of how impact came to be (van Wessel, 2018). Thus, they clearly illustrate temporal precedence, a necessary condition to provide evidence of causality (Shadish et al., 2002). Additionally, stories not only illustrate the important contextual variables that were present when a change occurred, but also how those variables (such as setting, participants, and relationships) interacted and weaved together as a holistic whole (Limato et al., 2017). Stories can illustrate the context in which multiple outcomes emerge and can show how outcomes interact; van Wessel (2018) states that oftentimes outcomes co-occur, rather than one set of conditions leading to one outcome.

Furthermore, stories provide context that demonstrates the linkages between the actors and the setting so that the overall cause and effect dynamic is more salient. As Dart and Davis (2003) state, "A good story defines relationships, a sequence of events, cause and effect, and a priority among items – and those elements are likely to be remembered as a complex whole" (p. 141). Stories may also reveal that a causal pattern is complex and non-linear, particularly if a researcher has not found a significant impact with experimental methods but has reason to suspect that a change has occurred (Abell, 2004).

Narrativists do not claim that stories portray an objective, generalizable truth. As van Wessel (2018) states, stories are not "an inferior stand-in for objective evidence" but are a "rich and meaningful source of knowledge in their own right" (p. 415). One way that stories can be rich and meaningful is by ensuring that they adhere to standards of rigor within narrative inquiry (Brinkerhoff, 2003). By detailing the circumstance, environment, and motivations of individual actors, stories can be a powerful way to portray cause and effect for an individual case (Abell, 2004). Even if these individual stories don't represent the widespread experience of a phenomenon, they are still worth knowing (Brinkerhoff, 2003). Narrativists believe that the story that is told is true for that particular context and point in time and is not intended for generalization (Abell, 2004; Brinkerhoff, 2003). Thus, narrativists make the argument that individual stories can reveal important information about causality for the individual case, but the findings from a case should not be generalized to a larger population.

It is important to note one issue of contention between the narrative and successionist ways of thinking when it comes to motivation. Within the narrative way of

thinking, individual motivation is a key driver of change. This assertion is in direct contrast with the random assignment procedure within the RCT (the most well-known design in the successionist way of thinking) that controls for motivation. Part of what can contribute to selection bias in an experimental design with poor random assignment is if particular individuals are motivated to succeed and these individuals are not randomly distributed between control and treatment groups any difference in effect between the groups might not be attributable to the intervention (as motivation may be the explanation) (Shadish et al., 2002). Whereas motivation is considered a threat to the causal argument in the successionist way of thinking, within the narrative way of thinking motivation is a necessary causal agent – it must be present for change to occur. One might argue that for narrativists the framing is: How did the individual's motivation interact with elements of the intervention to produce the observed outcomes? Narrativists might also argue that motivation is not a binary variable that one has or doesn't have; rather, it emerges based on the context and cannot be controlled for. This is a key tension between the successionist way of thinking and the narrative way of thinking.

Unexpected Outcomes

Evaluators argue that the story-based causal evaluation methodologies are credible for contexts in which there is flexibility to explore whether unexpected outcomes may have occurred. Dart and Davies (2003) state that stories collected through most significant change reveal outcomes that are meaningful within the social contexts of the participants. Oftentimes these stories reveal unexpected perspectives and detail outcomes that program staff were not aware were possible (Dart & Davies, 2003). The

methodology allows participants to describe changes that they perceived and interpret their value and significance according to their own value system (Choy & Lidstone, 2013).

Similarly, success case method allows unexpected outcomes to be articulated. Success case method explores areas where an intervention is working and does not seek to answer the question of whether the entire intervention works (Brinkerhoff, 2003). As such, success case method is a strong choice for exploratory evaluation of pilot programs where the outcome may still be unknown (Brinkerhoff, 2003). Because success case method and most significant change explore multiple explanations for why success may have happened, they are particularly helpful in evaluation contexts where the expected outcomes of the intervention are not yet clarified and there is an openness to exploring what those outcomes may be from the perspective of participants.

However, there are some differences between success case method and most significant change when it comes to outcome articulation. Within success case method, the participants in the evaluation study are those who are deemed successful by the organization that has requested the evaluation (Brinkerhoff, 2003). These participants share their story of how they were successful within a particular intervention or other experience (Brinkherhoff, 2003). This process allows participants to articulate unexpected outcomes; however, they are also asked to comment on specific pre-determined outcomes of interest to the organization that requested the evaluation (Brinkerhoff, 2003). This emphasis on learning about pre-determined outcomes of interest may stop participants from sharing information about the other outcomes that

have occurred for them. In contrast, most significant change is more open-ended. Participants are not asked to elaborate on pre-determined outcomes of interest for the organization that requested the evaluation. Rather, they are asked to openly recount a story about how an intervention may have impacted them (Dart & Davies, 2003). This openness creates more opportunity for unexpected outcomes to be shared.

Participant Voice and Multicultural Validity

Evaluators state that the story-based causal evaluation methodologies lift up participant voices and perspectives. Within most significant change, stories are collected not just from program staff but also from program participants, giving participants the opportunity to voice their own stories (Dart & Davies, 2003). The methodology also includes a component where participants rate which stories are most reflective of real impact (Dart & Davies, 2003). Through this process, participants are able to crystallize what impact looks like to them, and documentation of this is shared with other stakeholders (Dart & Davies, 2003). Similarly, success case method also includes a component wherein successful participants are asked to share their story of why something worked and illuminate the mechanisms of why something worked from their own perspective (Brinkerhoff, 2003). In success case method, non-successful participants may also be interviewed to illustrate program failures (Brinkerhoff, 2003), giving those participants who did not have a positive experience with the intervention the opportunity to share their voices as well.

Most significant change also takes steps to provide evidence of multicultural validity by welcoming participants to bring cultural values into the discussion. This

aspect of the methodology may explain why it is a popular choice of methodology among international development professionals (Dart & Davies, 2003). A key facet in strengthening evidence of multicultural validity in a study is including participants' voice in such a way that the participants' culture helps define what is correct, true or trustworthy evidence – these concepts must be centered in the participants' cultural values (Hood et al., 2015). Most significant change includes a component of asking participants to assess which outcomes told within the stories are important to them (Dart & Davies, 2003); this question places their cultural values at the center of determining what is correct and true.

As an illustration of this, Choy & Lidstone (2013) have utilized most significant change and storytelling in their work with Pacific Islander communities. They argue that storytelling illuminates how a participant's cultural values inform the criteria they use to select the outcomes they found meaningful from an intervention (Choy & Lidstone, 2013). Choy and Lidstone argue that stories promote conversation, allow meaning to be conveyed through the completeness of a narrative, are less formal than more conventional evaluation approaches, and are relatively short and easy to pay attention to. These characteristics of storytelling may make it more approachable for a wider range of participants (Choy & Lidstone, 2013), allowing them to participate more fully in the analytic process of selecting the most impactful story within most significant change. By ensuring that the participant's cultural lens is included not only in provision of the stories but also in interpreting the value of stories, most significant change procedures enhance cultural validity, which lends to its credibility for use in cross-cultural contexts.

Between the two methodologies, one could argue that most significant change elevates participant voice more than success case method. Most significant change encourages each group of stakeholders (beneficiaries, program staff, and donors) to articulate which outcomes they see as most valuable across all of the stories (Dart & Davies, 2003). This process necessitates discussion among the stakeholder groups and allows for a more collective, communal rating experience among participants about which stories and outcomes are valuable (Dart & Davies, 2003). In contrast, success case method does not invite the participant to rate which stories are most valuable. However, it must be stated that both methodologies are primarily intended to meet the need of the donor or organizational entity that requested the evaluation. The story or stories thought of as most emblematic of impact within most significant change are ultimately chosen by the intervention's donors (Dart & Davies, 2003). Thus, more weight is placed upon the donor's ratings of the stories. Similarly, within success change method, the organizational leads are the ones who determine which participants they think are most successful in order to gather stories from them (Brinkerhoff, 2003).

Transformation

A number of authors commented on the transformational elements of most significant change. The methodology is considered transformational because it promotes dialogue between beneficiaries, donors, and program implementers about what criteria they use to determine whether an outcome is valuable. Limato et al. (2017) implemented most significant change to evaluate an Indonesian maternal health program. Limato et al. were impressed by the way the methodology promoted dialogue between the program

beneficiaries and health providers about which outcomes were meaningful. Through a deliberative process, the participants arrived at one story that detailed the most significant impact to them – a story that described the benefit that was created by the health practitioners from the perspective of a beneficiary (Limato et al., 2017). The authors state that through this deliberative process participants began to appreciate why people from a different stakeholder group than the one they belong to might value something differently (Limato et al., 2017). As Limato et al. describe: "program implementers and local decision makers recognized that beneficiaries may have a different, but no less important, perspective on the worth of the program" (p. 109). The result of using most significant change was that the perspective among all of the stakeholders was broadened.

In a review of positive thinking evaluation approaches, Stame (2014) highlighted most significant change for its ability to make participants believe that they are important actors within the programs from which they receive services. This aspect of most significant change gives it transformational qualities because it distributes power from the program implementers and donors to the participants by seeking participants' definitions of success. Limato et al. (2017) also discuss how within most significant change the beneficiaries and stakeholders engage in analysis of the raw data of the stories, rather than the evaluator selecting key data to include in a final report. In this way, the methodology is transformational because it transfers the power of selecting what's meaningful from the evaluator to the participants. Thus, most significant change is a credible methodological choice for contexts in which there is a need to disrupt the power dynamics between donors, implementers, beneficiaries, and evaluators.

However, there is also an important critique that most significant change is not transformational. Dinh et al. (2019) critique the step of the methodology in which the donors choose the story that best represents the most significant change. This step of placing the final choice of what constitutes the best story in the hands of those with the most power does not appear to be transformational. However, Dinh et al. also acknowledge that Dart and Davies (the main developers of the methodology) have offered alternatives to counterbalance this power dynamic.

Success case method, in contrast, has less potential to be transformational. Within success case method, the participants are typically not engaged in laying out criteria at the outset for what success is (Brinkerhoff, 2003). Within this methodology, it is more likely the organizational leadership who makes decisions about what success is before seeking out stories from participants who meet the success criteria (Brinkerhoff, 2003). Additionally, success case method does not involve a dialogic process between participants, program implementers, and donors about which outcomes are most meaningful for each group. Furthermore, Brinkerhoff (2003) states that participants' stories should be corroborated with other evidence in a success case method study. One might argue that having a need to corroborate could take power away from the participant (Dinh et al., 2019). However, one possible use of success case method findings is to share participants' success stories widely within an organization (Brinkerhoff, 2003). One could argue that this sharing of success stories does elevate the participants' voice and has potential to transform power dynamics within an organization.

Successful Parts, Rather than the Whole

Some authors state that story-based causal evaluation methodologies focus on which aspects of or for whom an intervention was been successful. The methodologies are not intended to provide evidence as to whether the intervention "worked" for an entire sample. As such, success case method identifies pockets of success within an intervention and uncovers conditions of success for those participants who had the most success (Brinkerhoff, 2003). Non-story-based causal methodologies do not highlight pockets of success in this same way. Medina et al. (2015) experienced this when they utilized success case method for the evaluation of a public health training initiative. They stated that success case method helped identify that the training was quite successful for a small cohort of trainees, even though overall results showed that only one-third of participants retained knowledge from the training. Medina et al. found that the group of individuals that succeeded in applying knowledge from the training had certain levels of preparedness prior to the training (pre-existing knowledge and resources). Had the authors not investigated success stories through the success case method, they would not have discovered that this level of preparedness was necessary for the training to be most effective (Medina et al., 2015).

Success case method can provide evaluators with information about the why, how and for whom a training was successful (Medina et al., 2015). As Medina et al. (2015) state, "Our use of SCM enabled us to gather important information about how a training initiative was being used, in what context, how the training had been leveraged to build additional skills, and what outcomes had been achieved" (p. 131). Similar to realist

evaluation, success case method has the potential to answer the question of what works for whom and in what circumstances (Pawson & Tilley, 1997) through the crafting of success stories. Most significant change also shares this element on focusing on the positive stories to key in on what works for whom (Stame, 2014). Both methodologies encourage the evaluator to elicit stories of failure in addition to stories of success (Brinkerhoff, 2003; Dart & Davies, 2003); thus, the evaluator may also derive findings regarding what doesn't work for whom.

Corroboration

The story-based causal evaluation methodologies include an element of corroborating findings with other data to triangulate the findings. Brinkerhoff (2003) states that the stories collected through success case method should be corroborated through different forms of evidence, such as: visiting the story site, interviewing others within the same story context, and reviewing records. Brinkerhoff describes the evidence building with success case method to be similar to evidence building for a court case; one must gather corroborating evidence to make the argument for what happened. In their application of success case method through a public health training initiative, Medina et al. (2015) also used additional data sources, such as customer satisfaction surveys and knowledge gain assessments to accompany the stories captured through success case method.

Those experienced with most significant change also advocate for corroboration of findings. One way to enhance credibility of the findings is to conduct visits to the sites where stories were collected to confirm that the stories are accurate (Dart & Davies,

2003). Another way is to include participants' stories of negative aspects of an intervention, or things that didn't work, to serve as lessons learned (Dart & Davies, 2003; Limato et al., 2017). If generalization of findings is desired, more participants could be asked to provide stories (Limato et al., 2017). However, Dart and Davies (2003) acknowledge that the method is designed to collect the perspective of those who experienced success; as such, it doesn't capture the average participant experience. But, they state that the average participant experience could be gathered if that was needed to satisfy a study's aims (Dart & Davies, 2003). An additional suggestion for adding credibility to the methodology was to collect stories throughout the intervention to capture outcomes as they emerge (Limato et al., 2017), rather than waiting to the end to collect stories, which could bias participants' perspectives to focus on the most recent aspects of an intervention.

Both methodologies have similar suggestions for corroboration of evidence and both appear to be adaptable for use in combination with other approaches. For example, Dart and Davies (2003) share that most significant change can be combined with thematic analysis and quantization of how often themes emerge across the stories. Similarly, Medina et al. (2015) state that success case method has been combined with the quasi-experimental design of time series in the past, reiterating that these story-based causal methodologies can be combined with other methods that are more quantitatively based. It is important to note that some authors who advocate using story-based causal approaches do not think that the evidence generated from these approaches needs to be

corroborated with other (particularly quantitative) forms of evidence, as this creates the impression that the participant's story on its own is not credible (Dinh et al., 2019).

Summary

The first section of the literature review provided an overview of Gates and Dyson's (2017) causal pluralism framework, which argues that there are different ways of thinking about causality and that all of these causal views are credible. The second and third sections of the literature review synthesized how evaluators justify the credibility of the narrative way of thinking about causality and the story-based causal evaluation methodologies. Five themes arose within the literature, including: 1) story-based causal methodologies are able to unearth unexpected outcomes; 2) story-based causal methodologies place an emphasis on collecting participant voice; 3) story-based causal methodologies are transformational; 4) story-based causal methodologies focus on which parts of an intervention was successful and for whom; and 5) story-based methodologies are intended to be corroborated against additional forms of evidence.

However, there are gaps in the literature. The literature review discussed the ways in which story-based causal methodologies are credible according to evaluators who have published their work in peer-reviewed journals. Only 14 articles were identified through the literature review. These articles capture the perspective of a small number of evaluators who successfully emerged through the peer-review process. The population of evaluators who utilize story-based causal methodologies is much wider than the scholars whose articles were considered through this literature review. Additionally, the articles did not have a primary focus of examining credibility. When asked directly about

credibility, evaluators may have a more in-depth, detailed response than they might if they were merely introducing the methodology within an article. There is a gap in the literature because we have a limited understanding of how evaluators justify the credibility of these methodologies to multiple stakeholders.

CHAPTER III: METHODOLOGY

Chapter three presents the research methodology utilized to answer the main research questions of the study. First, I present the rationale for choosing case study as the research design for this study. Then I explain the participant recruitment and selection methods, as well as the data collection and analysis procedures. Finally, I describe my paradigmatic approach and positionality as the researcher. Table 3. displays key elements of this study design.

Table 3.
Research Matrix

Research Questions	Data Source	Collection Methods	Analysis Procedures
RQ1: What arguments are made by evaluators to justify the credibility of story-based causal methodologies to evaluation stakeholders? Do the arguments that evaluators make to justify the credibility of these	Four evaluators who have implemented a story-based causal evaluation methodology.	One interview with each evaluator to capture their perspectives on credibility. Review of documents and artifacts relevant to describing the case as well as answering the research questions.	A cross-case synthesis approach suggested by Yin (2018) to answer the research questions and sub-questions posed by the study. Thematic qualitative analysis procedures described by Creswell (2016) to guide the coding and theming procedures.

methodologies differ depending on which evaluation stakeholder is the audience for the argument? If so, how do they differ?			
RQ2: From the perspective of evaluators, how do contextual factors influence whether story-based causal methodologies are perceived as credible by evaluation stakeholders?	Four evaluators who have implemented a story-based causal evaluation methodology.	One interview with each evaluator to capture their perspectives on credibility. Another interview with each evaluator to capture the evaluation story. Review of documents and artifacts relevant to describing the case as well as answering the research questions.	A cross-case synthesis approach suggested by Yin (2018) to answer the research questions and sub-questions posed by the study. Thematic qualitative analysis procedures described by Creswell (2016) to guide the coding and theming procedures.

Rationale for Research Design

Case study was an appropriate research design to achieve the purpose of the study because the research questions demanded an in-depth inquiry across several cases in order to examine the central phenomenon of the study (Yin, 2009). The phenomenon that

was explored through this study was: how do evaluators justify the credibility of story-based causal methodologies to multiple stakeholders across different contexts? Examining how this phenomenon occurred across different evaluators' practice enabled me to derive cross-case themes about how evaluators' construct their credibility arguments as well what relational, contextual, and environmental characteristics affected the perceived credibility of the methodologies. Case study was also an appropriate design choice for this study because case study is an effective research design to capture participants' context-revealing stories (Yin, 2009), and the second research question focused on understanding the contexts of story-based causal evaluation methodologies. Finally, Yin also states that the case study should examine a contemporary phenomenon. The phenomenon of exploring credible alternatives to the experimental design for causal research is a contemporary phenomenon (Gates & Dyson, 2017). A four-case design was chosen for this study to provide sufficient evidence to corroborate findings and locate divergent findings to answer the research questions.

Research Questions

The research questions that drove the study were:

RQ1: What arguments are made by evaluators to justify the credibility of story-based causal methodologies to evaluation stakeholders?

- Do the arguments that evaluators make to justify the credibility of these methodologies differ depending on which evaluation stakeholder is the audience for the argument? If so, how do they differ?

RQ2: From the perspective of evaluators, how do contextual factors influence whether story-based causal methodologies are perceived as credible by evaluation stakeholders?

Stakeholders included the following groups: evaluation clients, donors or funders (if not the evaluation clients), program beneficiaries, program implementers, program participants, peer evaluators, and political stakeholders.

Participant Recruitment and Selection

Within case study methodology, the unit of analysis (the case) must be specified (Yin, 2009). My multiple case study design included four case studies of story-based causal evaluations implemented by four different evaluators. Thus, each story-based causal evaluation was a case, resulting in four cases in total. I used several approaches to identify potential participants, including:

- identifying evaluators within the scholarly or grey literature who had conducted most significant change or the success case method;
- identifying evaluators who had presented at American Evaluation Association events on these methodologies;
- asking for referrals from members of American Evaluation Association topical interest groups that are aligned with story-based approaches;
- identifying evaluators by using a search on Google; and,
- asking my professional network of evaluators to recommend individuals who may fit the inclusion criteria.

The inclusion criteria for participants in the study was that they had to have conducted at least one story-based causal methodology. Eighteen potential participants met the inclusion criteria and were contacted and four of these participants agreed to participate in the study.

Data Collection

Within a case study, it is helpful to have multiple forms of data collected to provide enough detail to construct the case and answer the key research questions (Yin, 2009). This case study included the collection of data through interviews, a review of artifacts, and a review of documentation.

Interviews

Four evaluators participated in the interviews. Pertinent information about each participant is included in Table 4., including their pseudonym, race, gender, and a high-level description of the case reviewed for the study. It should be noted that participants were not asked to identify their race or gender; the race and gender information listed below is from what I observed.

Table 4.
Information About Study Participants

Pseudonym	Race	Gender	Evaluation Case
Dan	White	Male	A success case method study to evaluate a continuing medical education program.
Claire	White	Female	A success case method to evaluate a leadership program at a large corporation.

Emily	White	Female	A most significant change study to evaluate the theory of change of a non-governmental development organization working in West Africa.
Lori	White	Female	A most significant change study to evaluate the impact of a fellowship program at a national environmental organization.

I conducted two interviews with each participant. Each interview lasted approximately one hour and was recorded. The interviews took place from January to February 2021 over Zoom. The first interview focused on an occasion when the participant utilized a story-based causal evaluation methodology (see interview protocol in Appendix A). The interview asked the participant to expand upon details about the setting and context, the evaluation questions, the main actors within that setting, the evaluation procedures, and the result of the evaluation. Gathering these specific features of the story (setting, main actors, plot and chronology) enabled me to construct the case as a story in the findings section of the study (Creswell, 2013).

The second interview focused on the credibility justifications used by the participant over the course of their implementation of the story-based causal evaluation (see interview protocol in Appendix B). The interview was open-ended to elicit the kinds of justifications they made during the case to different evaluation stakeholders about the credibility of using story-based methods for causal evaluation. The interview also focused on the environmental, relational, or contextual factors that influenced whether the story-based methodology they utilized was perceived as a credible choice for causal evaluation.

Evaluation Documents and Artifacts

Evaluation documents and artifacts were collected to capture how the evaluators articulated the credibility of story-based causal methodologies. They also provided detail to enrich the thick description of the cases. Any object that was primarily comprised of text I referred to as a document and any object that included visual content I referred to as an artifact. Documents and artifacts in the multiple case study included: publications, slide deck presentations and other presentation content, and evaluation products for clients. A data collection protocol form was utilized for the review of documents and artifacts and can be found in Appendix C. Appendix D details how each data collection protocol question is connected to the research questions of the study; this demonstrates alignment between research questions to sources of data (Anfara et al., 2002; Yin, 2018).

Data Analysis

Yin (2018) recommends structuring the analysis for a case study using a cross-case synthesis approach that aggregates findings across cases. This study utilized the analysis approach of within-case and cross-case analysis (Yin, 2018) in which within-case themes were identified that informed the cross-case themes. In order to aggregate the findings, the data were coded, categorized, and themed. The analysis began with a coding strategy for qualitative analysis that was developed by Creswell (2016). Before coding began, I developed a codebook that included some a priori codes relating to existing theory. This a priori codebook can be found in Appendix E. I began the coding by reviewing all of the data available for one case and labeling each segment of text (Creswell, 2016) with either an a priori code or an emergent code (Creswell, 2013). After

the data for the first case were coded, the codebook was expanded to include all of the emergent codes that resulted from coding. This codebook was then utilized to code the data from the remaining cases. As more emergent codes emerged from the review of data from each new case, the codebook was updated to include those emergent codes and the previously reviewed cases were re-reviewed to apply that new code. Atlas.ti 8 was utilized to code the data. The final codebook is included as Appendix F.

After coding was complete, the analysis of the data continued into the categorization and theming process, first to develop themes for within-case analysis, following with theme development for cross-case analysis. For within-case analysis, the codes were grouped into similar categories, and then these categories were grouped into themes based on the similarities of the content (Creswell, 2016). Within-case themes were derived that answer the research questions from the vantage point of each participant and their case.

For cross-case analysis, a code frequency table was developed that displayed how frequently codes occurred across cases. This table (shown in Appendix G) helped to identify patterns of where codes occurred across cases. Additionally, within-case themes were reviewed to determine whether the data that supported those themes could be categorized together to inform cross-case themes. The cross-case themes answered the research questions from the vantage points of all participants and all of the cases. Additionally, as disconfirming evidence arose within the analysis process, it was drawn upon to strengthen and balance the themes presented in the findings (Yin, 2018).

Appendix H includes a flow chart of the data collection, coding, theming, and writing process.

Ensuring Rigor

The study also included specific procedures to increase rigor. Creswell and Miller (2000) state that the researcher's paradigmatic assumptions should drive which procedures they choose to employ to enhance rigor. Although my paradigm for this study was a pragmatic paradigm, I also embraced elements of the critical paradigm to enhance the rigor of the study. The procedures for enhancing rigor that align with the critical paradigm include collaboration, peer debriefing, and researcher reflexivity, according to Creswell and Miller.

The study included all three of these rigor-enhancing strategies. I wrote reflexive notes during all phases of the research. These reflexive notes recorded the extent to which my lived experience impacted how I interpreted the data (Peshkin, 1988). To ensure collaboration with research participants, I shared their interview transcripts with them as a member checking exercise. I also shared preliminary findings with them and asked for their feedback over email. Their feedback was considered in the overall analysis. In regard to peer debriefing, I asked academic peers to comment on whether the findings were understandable and usable. They were sent high-level bullets summarizing the findings section and were asked to provide feedback over email. Their feedback was also incorporated into the findings.

Researcher Positionality

Within any research study, the researcher should reflect on their positionality, which includes both the research paradigm from which they operate as well as how their own life experience and background may influence their design and implementation choices. Wilson (2008) states that paradigms are comprised of ontology (the nature of reality), epistemology (the researcher's relationship to the world), axiology (the ethics or morals that guide the search for knowledge), and methodology (the process through which knowledge is gained). Table 5. provides a summary of the four paradigm assumptions for the pragmatic paradigm – the paradigm that I most closely align with.

Table 5.
Summary of Pragmatic Paradigm

Paradigm Assumptions	How the Assumption Is Framed Within Pragmatic Paradigm
Ontology	• The researcher embraces both the post-positivist and constructivist points of view; as such, she embraces qualitative and quantitative methods and operates from points in-between (Teddlie & Tashakkori, 2002). • Uncovering results that can be implemented within the specific context is more important than uncovering a singular truth (Mertens & Wilson, 2012).
Epistemology	• Findings are applied to solve a problem or resolve the research question (Teddlie & Tashakkori, 2002)
Axiology	• Values are informed by what is most practical and what works best in applied settings (Teddlie & Tashakkori, 2002)
Methodology	• The research question is more important than the method or the paradigm that informs the method (Teddlie & Tashakkori, 2002)

As a pragmatist, I believe that I can use various tools, including qualitative and quantitative tools, to unearth meaning. I don't believe that my research produces one "truth"; rather, I believes it produces one version of the truth. Additionally, as most of my professional experience has been as an evaluator, one of my primary values is to answer evaluation questions with the best methodology of fit, knowing that the findings only reflect one version of the reality. This value is evident in my topic of choice for this study: I do not believe that experimental studies produce results that reflect the only version of the truth, but I do believe they are credible at producing evidence about one version of the truth. In some cases, experimental studies are the correct choice of methodology to fit the research purpose. But in other cases, other causal methodologies are needed to fit the context. It is this belief that allows me to embrace the view of causal pluralism. My primary interest in this study is in raising the credibility of non-experimental causal designs that may be a better fit for small social programs seeking evidence of causality.

My pragmatic paradigm has influenced the methodology chosen for this study. As a pragmatist, I want the findings to be utilized by the evaluation field; thus, I have presented the chapters in a manner that is readable and accessible to evaluators. Additionally, operating from the pragmatic assumption that the research question should align with the methodology, I chose a case study because it allowed me to explore the context in which evaluators justify the credibility of story-based causal methodologies. Finally, I also interviewed evaluators with applied experience because I wanted to

provide findings that were applicable to other evaluators who intend to use these methodologies.

Researcher Background

I have been a practicing evaluator since 2010, working primarily in social service settings such as governmental agencies and non-profits. Within these settings, the constraints of lack of funds or sufficient sample size to complete an experimental design were common. Additionally, many of the organizations I worked for valued receiving information in a timely way to make decisions. Thus, my work experience in applied settings has led me to operate from a pragmatic paradigm. I place a high value on: 1) choosing a methodology that fits the context; 2) answering the client's questions; and, 3) making decisions that balance rigor with feasibility.

Finally, I must also acknowledge factors about my position in society that may have influenced this research. Throughout my Ph.D. studies, I worked part- or full-time as an evaluator and consider myself an evaluation practitioner. My primary interest is to provide research findings that can be applied by evaluation practitioners. Additionally, I recognize that I have been afforded privilege in society as I am a white, middle-income, cisgender female. I sought to maintain humility throughout the study and regularly reflect on how this privilege might have influenced my research choices and interpretation. I strove to unearth, examine, and address the subjectivity that comes with this privilege through the critical paradigm strategies I implemented to enhance the rigor of this study.

CHAPTER IV: FINDINGS

This study examined the arguments made by four evaluators when justifying the use of story-based causal methodologies. This study also examined how the evaluation context of these evaluators influenced how stakeholders perceived the credibility of these methodologies. This chapter comprises two sections. The first section presents the case story of each evaluator and how they conducted an evaluation using a story-based methodology. Each evaluator's case story is followed by a presentation of the themes that arose from analysis of their case materials relating to the topic of credibility. The four cases stories are told in the order of Dan, Claire, Emily and Lori. The names utilized are pseudonyms to protect the confidentiality of the participants. Additionally, some non-essential details about participants or in their quotes have been changed to protect their confidentiality.

This chapter's second section presents the cross-case themes that emerged across the case stories. These cross-case themes describe the evaluators' justifications for using story-based causal methodologies. These cross-case themes include: 1) type of evidence needed; 2) paradigm as determinant of perceived credibility; 3) truthfulness of accounts; 4) triangulation among data sources; 5) elevating participant voice; 6) unexpected outcomes or mechanisms; and, 7) credibility from the evaluator or organization.

Case Stories and Within-Case Themes

Dan's Story Using Success Case Method

Dan conducted a success case method study to evaluate the success of a continuing medical education (CME) program designed to help practitioners implement new tobacco cessation guidelines. The program was delivered by a national partnership of CME providers who were implementing the new tobacco cessation guidelines, which included encouraging the use of a medication that was proven to be effective in helping people to stop smoking. The purpose of the evaluation was to learn whether the CME program was contributing to practitioners' success in implementing the new guidelines. The evaluation was designed to highlight instances in which a healthcare setting and provider successfully changed their practice in a significant way, understand how that change occurred, and understand what role the educational intervention played in making that change occur.

At the time of the evaluation, Dan had a leadership role in a research unit within a CME department at a university and was also a professor at the university. The work of the unit was funded primarily through grants, often from pharmaceutical companies. At this time in his career (and at the time of his participation in this study), he identified as a qualitative researcher operating from a constructivist mindset. While he respected the use of quantitative methodologies and believed that they had their place, particularly when the research question and context would be best served through a quantitative approach, he found that the case study approach often answered the questions that he had:

> I just felt that while quantitative approaches were very powerful, they could only answer a limited range of questions. That was a decision I made early on in my career as a grad student, to go that route. I wasn't wedded to any one methodology, which appeals to not so much the educator in me as the learner. That's why I liked academia. I would use methodologies that fit the problem of interest, and if I hadn't used it before, then I'd learn it. It just happened that case study was often applicable and I think that's because my questions started to take forms that were influenced by my background doing case studies.

In the last line of that passage, it is apparent that as Dan's experience conducting case studies deepened, his ability to see how the case study could be applicable to different research contexts also expanded. When the opportunity to evaluate the CME program arose, he felt that a case study methodology would be the right choice as the purpose of the evaluation was to deepen understanding rather than prove that the intervention was effective. He was also interested in using success case method in particular for the evaluation because there were no instances in the literature of the methodology being used in medical education. He identified this as an opportunity to address a gap in the literature and inform the field of how this methodology could be used to learn how success occurs in medical education interventions.

The entity funding the evaluation was a pharmaceutical company. This company was committed to supporting evidence that could be used generally by the medical field and would not fund research projects directly tied to their financial interests. To meet the interests of the company, which was interested in research findings with broad

applicability across the CME field, Dan ensured the evaluation "could inform other projects that built upon what we were doing, and projects beyond that that had nothing to do with tobacco cessation." The organizations providing the medical education were another important group of stakeholders in the evaluation. There were three different kinds of medical education interventions being provided, ranging from face-to-face cohort models to variations on online learning models. These education providers were most concerned that the effectiveness of their three different models would be compared. However, Dan allayed their concerns by explaining that the intent of the success case method was to learn about individual contexts, not to compare effectiveness across the models.

Another important group of stakeholders was the group of practitioners implementing the tobacco cessation guidelines. The practitioners had a stake in their patients' health improving. Dan explained that many of the practitioners held a sense of guilt when patients were not well, and that participating in the education program gave them "a chance to address that." The patients, of course, were another important stakeholder group. By participating in the program, they had a chance to improve their health. As Dan said of the intervention, "The number of people who were successful in quitting over baseline was increased pretty substantially." An additional stakeholder group in the evaluation included CME professionals who consumed research to advance their own practice. Finally, researchers within the CME field were also stakeholders. Dan was eager to publish findings that would address the gap in the literature around how

success case method could be implemented to learn about the success of CME interventions.

Dan felt that a convergence of factors made the timing right for this kind of evaluation. The new tobacco cessation medication that was part of the intervention was shown to be effective at helping people quit smoking, and it had been a long time since a new drug for tobacco cessation had entered the field. Additionally, the leader of his research office was considered a “visionary” with widespread recognition; this encouraged the participation of different medical education providers - who are typically in competition with one another - to participate in the evaluation. Because these providers had experience working with Dan and the leader of his research office, the evaluation began with a degree of trust already built. As he said, among “the three partners that had three competing approaches to doing the clinical improvement component, there was enough trust. They were willing to let us go ahead, do our data collection, and feel confident that they weren't going to pay the price.”

With the funding in place and the stakeholders on board, Dan assembled a small team to conduct the evaluation. Within success case method, one of the first steps is to identify participants who are deemed to be successful by certain criteria. These are the participants whose stories are gathered to help uncover the factors behind their success and how the intervention may have contributed to that success. In Dan’s case, there were measurable clinical outcomes within the patient records that could be utilized to determine instances of successful implementation of the intervention. He and his evaluation team used these data to identify the successful practitioner cases to examine.

They also set up an expert panel to review their success case selections and validate those selections.

To inform their choice of data collection and analysis procedures, the team utilized Stake's (1995) case study guide and Brinkerhoff's (2003) success case method guide. Dan considered his use of Stake's guide as an important "guidepost" to demonstrate to stakeholders that the study was credible. As he said, "we weren't just out there winging it. We had a highly-regarded, well-documented methodology." Procedures included interviews of the practitioners using a semi-structured survey and peer interviews at each site. Data from these interviews were utilized to craft a case description for each site. The team's analysis also revealed within-case and cross-case findings.

The findings from Dan's evaluation study included surprising insights for the field. They called into question some of the pervasive thinking within CME at the time, specifically the thought that one practitioner alone could successfully implement a practice change. As he said,

> Part of what we were doing was showing the mistake in logic behind a lot of the CME. For example, we showed that even with a relatively simple guideline to implement, it oftentimes took a team effort. The whole clinic had to become involved, which caused them to question the idea that continuing education should have as its target audience individual physicians.

His findings led to the recommendation that CME as a field should think more critically about how to integrate a new intervention into the practices of an entire clinic. The

findings also highlighted cases in which integrating the new tobacco cessation guidelines led to patient progress in quitting tobacco. Dan and his team shared the findings back to their stakeholders and to the CME research field through publications and at professional association meetings.

Themes within Dan's Case

Dan's case materials revealed two major themes in regard to how he justified using success case method when examining causality, as well as how his evaluation context influenced stakeholders' perception of the credibility of the methodology. These two themes were: 1) trustworthiness and 2) learning instead of proving.

Trustworthiness. In his success case method study, Dan relied upon research procedures designed to ensure validity of findings to enhance the credibility of his work. He used a well-known case study methodological guide - Stake's (1995) guide - to inform how his team conduced the case study, in addition to using Brinkerhoff's (2003) success case method guide to frame the study. The validity procedures included: diversifying the cases that were studied by including practitioners with different specialties, triangulating interview data with evidence of the intervention's implementation from health records, asking probing questions during interviews, and sharing the draft report back with participants to ensure the findings were accurate. Additionally, the presence of cross-case themes contributed to the trustworthiness of the team's findings, because despite there being great diversity in the cases, there were commonalities of experience.

Dan believed in the importance of having specific procedures in place to improve the trustworthiness of the evaluation. In the grant proposal for the study, he purposely included details regarding procedures he would use to ensure the findings would be trustworthy. When he was writing up the evaluation results for a journal article, he was mindful that he would need to make a strong case for how the procedures contributed to the validity of the findings. Depending on who the stakeholder was, he might also tailor his presentation of the findings to make the strongest case for trustworthiness that would appeal to that stakeholder. He described presenting the findings to different stakeholder groups this way:

> It's a bit like a diamond, if you will, it has multiple facets and the diamond is the rationale, the justification for different parts of the study. All of them would be present in one way or another in presenting the study to a stakeholder while they're reading the initial proposal or in the final report. We might emphasize a different facet of the diamond, give more attention to it, depending on who the stakeholder is.

As an example, when Dan's team presented their results to an audience they knew was more receptive to quantitative data, they placed more emphasis on how they generated evidence from quantitative data. His experience with success case method seems to suggest that including multiple validity procedures in the methods may help the methodology appear more credible to evaluation stakeholders. In particular, triangulating the qualitative data with a quantitative source may make the findings appear more credible for certain stakeholders.

Learning Instead of Proving. In discussing the credibility of success case method, Dan emphasized the distinction between what the methodology is best suited to do versus what it is not suited to do. He felt that success case method is best suited to contexts in which the objective is to learn about, not to assess, the effectiveness of an intervention. As he said, "The aim of it was not so much to demonstrate or prove anything as it was to get a better understanding. It was really learning." Part of this learning process can also include exploring which factors within an intervention may contribute to change, rather than needing to "prove" that certain factors caused change. As he expressed, "We were learning, in the cases of these successes, what actually happened, what part of the education contributed, in particular, among all the other contributors." He also explained that researchers gather multiple forms of evidence from multiple sources in a case study as part of what he calls a "process of developing a better understanding." He believed that case studies can highlight factors that contributed to change, but are best suited to enhance learning rather than prove that an intervention worked.

Dan also recognized that the value in using a methodology like success case method is that it can be utilized to enhance learning across participants. He explained that methods like success case method help practitioners share knowledge about how they are approaching similar work. He noted that when practitioners share their stories, it sometimes highlights when they are out of step with their peers' practice. He believes that there are "multiple forms of knowledge, including practical knowledge"; thus, when

practitioners share knowledge about what they have learned from their practice, that is a valuable resource.

Additionally, success case method was an amenable methodology choice for the stakeholders in Dan's evaluation because the methodology does not include a comparison of effectiveness across programs. Knowing that their success (or non-success) would not be compared with others' was an important factor in easing practitioners' concerns about participating. He allayed the practitioners' concerns by explaining that success case method is meant to explore factors of success within each specific context and is not meant to compare similar interventions across sites. His frequent communication about what success case method is meant and not meant to do encouraged practitioners to have more comfort participating in the study. As he stated, the team was "very explicit about ways that we thought that the findings could be used or what conclusions could be drawn from them, and what conclusions should not be drawn."

Claire's Story Using Success Case Method

Claire conducted a success case method study to evaluate the success of an employee leadership training program for a large food processing corporation. The goal of the program was to train leadership to make faster decisions, be more innovative, and to experiment with their products. As she explained,

> They wanted to get leadership to make fast decisions, so increased agility in the way that people would make decisions. People would be better connected, so that they could work better across different enterprises within the organization, so that they started experimenting, like trying out new things in different sub-units

> of the corporation to come up with more innovation. Ultimately, they want to innovate and rejuvenate the organization and get it less stiff and flexible so they could actually increase market growth and profits.

The purpose of the success case method evaluation was to demonstrate whether these outcomes for employees - increased agility in decision making and increased experimentation - had occurred. In particular, the success case method was utilized to gather stories from employees who were performing the best on these outcomes and understand how the employee training contributed to their performance.

At the time of the evaluation, Claire was in a leadership position at an evaluation organization that completed about four evaluation projects a year for different clients. The organization had a well-known founder, and did not need to advertise because clients would contact them in order to work with the founder. She was also a professor at a university. She described her evaluation work as "between academia and consulting" because she brought an academic lens and academic rigor to the work, but also needed to balance that with the demands of the clients she worked for. As she explained, sometimes the founder would say to her, "You are way too academic – stop overthinking this and be practical." However, she also found that being positioned between academia and consulting was beneficial. She appreciated the ability to conduct success case method evaluations while also being a professor, because she could bring real-world experience in evaluation into the classroom. As she explained, "Those experiences can be valued at least by some of the graduate students."

Claire's training in evaluation also influenced her evaluation approach at the time. Having earned a Ph.D. in evaluation, she was able to identify the right approach for each specific context. She explained that her academic background and her skill in navigating which evaluation approach to use made her clients perceive her as credible. It equipped her with the ability to offer a suite of evaluation options. As she said,

> There's other ways to think about evaluation and evaluation is more than one method. Just because you know how to do a survey doesn't make you an evaluator.
>
> Having good evaluation training and being able to talk about different methods, the advantages, and disadvantages, that different approaches to evaluation could bring is important, I think.

Claire's evaluation approach at the time (and at the time of her participation in this study) was that findings should be utilized, and one way to ensure utilization is to provide findings to the client quickly. She found that conducting evaluation methodologies like success case method, in which findings are produced after 8-12 weeks, was more useful than doing long-term experimental studies, because oftentimes the findings from those studies come years later than when the information was needed for programmatic decision making. As she said, "If you want to do useful evaluation it is needed now."

Additionally, Claire believed that another way to ensure utilization of findings was to provide evidence of how the program is working for the best performers. She emphasized that the advantage of using the success case method over an experimental

trial is that one learns from the experience of the "outliers" – those who experienced extra-ordinary success or non-success. As she said,

> What the success case method assumes is that you really learn the most from the outliers on the top and the bottom, because then you can move them to the next level. What do we know if we have a statistically significant outcome from a randomized controlled trial? We know that the average moved, but we don't know anything about what's the best the training did. And what are the barriers for those who didn't perform at all? We just know what mediocracy means, that's all.

She elaborated that success case method is a unique evaluation methodology because it examines how the best performers become the best:

> That's, I think, where the success case method becomes exciting. If you learn from the best, so why are you the best in this? Why are you so excelling in this while the majority is average? What are the factors that contribute to that? Once you know that you can leverage that knowledge to improve your organization.

In Claire's view, one should be curious about what makes the best performers the best because that helps one learn how to improve performance for everyone.

The evaluation also had key stakeholders. The corporation's management was a key stakeholder because they funded the leadership training for employees. They wanted the training program to be effective because they trained thousands of employees. If they could uncover why it wasn't effective, they could change elements of the training to make it more effective. Additionally, the management wanted to improve the training to

improve the overall health of the organization. As Claire said, “The impact is downstream. So better work units, better organizational culture, better impacts for corporate, so it goes down the stream.” The employees of the corporation were also a key stakeholder, as the findings were meant to improve the quality of their training, which in turn was meant to improve how they worked with those on their team. Finally, the human resources department was also a key stakeholder because they wanted to learn more about the effectiveness of the training and to utilize the success stories to market the training to other employees.

Claire and her team began the evaluation by identifying high and low performers by reviewing data on training dosage and training completion. Their procedures also included a survey and interviews. Claire described the process of conducting the evaluation as participatory because staff from the corporation was engaged in the evaluation process. As she said of success case method more generally, “The success case method is a sequential mixed methods approach, a utilization-focused approach and value driven approach because you engage stakeholders from the beginning on in all steps of the evaluation and in the data interpretation.” At different stages in the evaluation, she presented findings back to the corporation: when the survey findings were ready, when they heard insightful information from the interviews, and in a final report. She stated that the corporation utilized the findings to improve the training; as one example, they immediately addressed a learning activity that didn’t work as well as expected and tried a new approach in the next cohort.

Themes Within Claire's Case

In reviewing Claire's case materials, two major themes emerged relating to how she justified using success case method when examining causality, as well as how her evaluation context influenced stakeholders' perception of the credibility of the methodology. These two themes included: 1) choosing a methodology that produces the evidence needed; and, 2) position on "quant-qual" debate determines position on credibility.

Choosing a Methodology that Produces the Evidence Needed. Claire believes that the purpose of evaluation is to improve programs. As she says,

> "Why do we evaluate? That's the question. What's the purpose of it? To publish papers that are published in some journal that no one will ever use, or to produce some findings for a group of people that's going to change the world?"

She tends to find methodologies like success case method more helpful in improving programs than experimental methodologies, which are often designed to examine whether an intervention had an effect for the average participant. She is interested in knowing the story behind the individual at the tail of the distribution who experienced a large effect; she is less interested in knowing the story of the average individual. As she says of experimental studies that tend to report the effect size for the average individual, "Is that effect - even if it's a meaningful effect – is it the best we can do? Or is it the average we can do?" According to her, the exciting part of implementing a success case method study is the examination of stories from those who experienced a large effect

from the intervention. Then, the organization can learn from the experience of the best performers in order to improve the performance of other staff members.

Claire remarked that when she conducts a success case method evaluation, she generalizes the findings from the sample included in her study to the rest of the population who received the intervention, but she does not necessarily generalize the findings to other training interventions. In general, she thinks there is an inherent conflict between the intent behind generalization and the drive to improve programs. As she explained,

> Generalizability assumes that interventions don't change…it assumes implementation fidelity and assumes exactly the same intervention, which is completely against anything I believe an evaluation should be doing. Evaluation should force you to change the program.

According to Claire, evaluation should produce findings that are translated into actions that improve the way a program is run; because staff should be constantly improving the program, the intervention should not remain static. For her, *improving* a program though evaluation rather than *proving* that a program had an effect is much more interesting; as she said, "That's what I want to do in my life as an evaluator. I want to improve programs that help people that care about helping people rather than proving programs."

She also stated that evaluators should choose a methodology based on two primary considerations: 1) the methodology should directly relate to the evaluation question; and, 2) the methodology should directly relate to the level of rigor that is required to produce the evidence that is needed. As she explained, the most important

consideration in selecting a methodology is determining which methodology best answers the evaluation question. One should consider all of the available evaluation designs. The second most important consideration is how much risk is associated with how the findings will be used.

As Claire elaborated, if the findings from an evaluation will be used to pilot a medical intervention, then the findings must meet a certain standard of evidence because the risk of them not being correct could cause serious harm. For these types of scenarios, particularly medical scenarios, an experimental study might be the right choice. But in other scenarios, such as a training evaluation, the risk of the findings being incorrect is not likely to cause serious harm; rather, the risk would more likely be one of wasted time. As she clarified, in reference to a training evaluation, "The question becomes, what counts as evidence and what's your standard for credibility? We're not developing drugs or surgeries. We are providing evidence that training works and leads to impact."

Thus, according to Claire, the kind of evidence and the level or rigor needed depends upon the context. She stated that for success case method, the standard of evidence is whether or not the evidence passes a test of reasonable doubt. Ultimately, striking the right balance between rigor and feasibility is related to the amount of effort needed to produce the level of evidence that is needed. She likened it to the amount of effort one would put into selecting an apple:

> Our bar for evidence is beyond reasonable doubt. Do you have a good reason for questioning why the results would not be true? Then you can go into research paradigms and really dig into what is and is not, what does count for evidence.

> For me, it's always, for what purpose? If I think evaluation, you go to the grocery store and you have a pile of apples, and you pick an apple. You evaluate the apples in the store, and you pick an apple. Are you going home and wondering whether you picked the best apple?

As she is illustrating, if one's purpose is to select an apple, one conducts the sufficient amount of evaluation to collect the sufficient amount of evidence to make that decision; the same kind of reasoning would apply when selecting an evaluation methodology.

To further illustrate the point that the level of rigor needed depends upon the context, Claire provided the example of a diagnosis made by a doctor. When one visits a doctor, the doctor makes an assessment of the medical concern, often using qualitative data, and then makes a diagnosis. She argued that in this scenario, we as a society widely accept that the diagnosis is credible. She also provided the example of a court case. In a court case, evidence is brought together, including qualitative evidence, and a determination is made as to whether the case argument passes reasonable doubt. This type of determination is also widely viewed as credible in our society. Thus, according to her, contextual factors are key in determining whether the type of evidence provided is credible.

Position on "Quant-Qual" Debate Determines Position on Credibility. For Claire, whether a person views success case method as credible often comes down to where that person falls in the "quant-qual" debate. In her experience teaching others how to do success case method, she observed that some students believed that evidence was not credible unless it was quantitative evidence. In her experience, students who aligned

with a quantitative approach – typically those interested in what she described as "measurement" – would often question the credibility of the evidence produced through success case method. She reflected that viewing only quantitative evidence as credible is connected to issues of power and privilege. As she said, "we know that mixed methods, more qualitative oriented methods, work better and are more embraced by people of less privilege, of more diverse people." She went on to explain, "Hardcore methods come with privilege, and understanding of those methods comes with privilege because knowledge is power."

Claire expressed that another group of stakeholders who seem to reject success case method as credible for examining causality is political staff or governmental employees. As she said, "Politics right now is all about evidence-based practice. Everyone wants an experiment or a quasi-experiment. They wouldn't hire you for a success case study… (they) want to prove something works or doesn't work" (parenthesis mine). Thus, in her assessment, whether or not success case method is perceived as credible for examining causality depends upon the paradigm of the stakeholders.

Emily's Story Using Most Significant Change

Emily conducted a most significant change study to evaluate whether outcomes within a non-governmental organization's (NGO) theory of change were achieved. At the time, she was the lead internal evaluator placed within this NGO. The NGO worked with West African communities on community-led development projects. Her task was to utilize most significant change to evaluate whether the NGO was achieving the long-term outcomes it sought to achieve in some of the communities it had been working with for a

few years. Specifically, she was evaluating whether the long-term outcomes evident in the communities aligned with the outcomes the NGO sought to achieve in their theory of change.

Emily held certain values about how international development work should be done, and those values influenced how she approached evaluation work. As she said,

> My theory of the most sustainable and ethical way to do international development is around centering community, not even just participation, but leadership and agency and ownership. I think that definitely bleeds into how I am an evaluator.

As this quote illustrates, for her it was crucial to use a participatory approach when doing evaluation. Additionally, the NGO's community-led approach to their development projects influenced their choice of using most significant change for evaluation because the methodology centers community voices.

When Emily joined the NGO as an internal evaluator, the NGO had been conducting most significant change studies for a few years. The NGO would partner with a community over the course of a few years, completing different community projects, and would evaluate each project using most significant change. Thus, Emily's evaluation team, comprised of NGO staff, was well-versed in using most significant change and could help the evaluation run smoothly. As she explained, "They had, two, three years of MSC experience — some staff had been doing it since the beginning." However, this evaluation differed from the ones that came before, as this evaluation reviewed impact

from different projects to assess whether the outcomes outlined in the NGO's theory of change were being achieved, rather than reviewing the impact of an individual project.

In the communities where the projects were being implemented, oral storytelling was an established tradition. For this reason, Emily thought that most significant change was an "easy fit" for the NGO's evaluation work, as storytelling is a key component of the methodology. Additionally, the NGO preferred an evaluation methodology that wasn't "extractive." She explained that an extractive evaluation would be one in which the findings are not shared back with the community members who participated in the project or in the study. She thought that the NGO had a unique stance in terms of the kinds of methods they viewed as extractive. As she said,

> I got checked too by my bosses, "That's too extractive." Even within my own like, "Oh, I'm super hippie-dippie participatory methods are the best" mindset, I was still being told "that's too much." The culture of the organization, it's very different than a lot of larger NGOs that push randomized controlled trials and all of those things.

Both the NGO and Emily appreciated most significant change because it is not extractive. In their implementation of most significant change, they "prioritize(d) closing the loop and opening up space for other feedback" (parenthesis mine). According to Emily, the community-led values of the NGO aligned with the choice of most significant change.

A key stakeholder in the evaluation was the NGO that employed Emily to conduct this internal evaluation. Their stake in the evaluation was to learn whether they were creating change in the key areas of their theory of change. Another key stakeholder group

was the group of local project volunteers who helped to design, implement and monitor the development projects. They wanted the evaluation to show they had succeeded, but they also wanted to see an honest depiction of how well the NGO played their role in the project. As Emily described,

> They were really involved in our programs, so I imagine that they would want to show the program in a good light. At the same time, by the time you got to the end of a few years of working with a group of volunteers, like anecdotally from program staff, they would say people are just being more honest — now they're holding the NGO to account.

As for the other community members benefitting from the projects that weren't the volunteers, their interest in the evaluation was making sure that their opinions were heard.

To collect data for the evaluation, Emily's team held focus groups in the local language. The main question her team asked was, "What's the most significant change in your life since the beginning of the program?" They distilled individual stories from these focus groups and then joined with other staff at the NGO to score the stories using a rubric to determine which story best captured the most significant change. After the top stories were selected, the team conducted full interviews with the community members who had shared them. The interviews were conducted by staff members who then wrote the narrative into most significant change stories. Emily also undertook a thematic analysis across the individual stories.

Because presenting the findings back to the community was so important to the NGO and to Emily, findings were shared in a community meeting. During this meeting, the community members whose stories were selected to represent the most significant change read their story out loud. Additionally, Emily's evaluation team displayed photos to further convey findings from the evaluation. As she described,

> We also used pictures to represent the areas of the theory of change. For gender equity, we would have a picture of someone in that community like a dad with his daughter tied on his back, which is normally a female thing but we would show that picture around and then talk about the different themes that we had found.

The team also posted the stories and photos in a community public space so that community members who could not attend the meeting would still be aware of the findings.

Emily believed that the findings influenced how the NGO conducted their work going forward, particularly because the evaluation demonstrated which components of the theory of change appeared to be working as intended and which components did not. As she said, "I do think that the things that came out, and the conversations I had with the program managers did help to shape how they implemented their programs moving forward."

Themes Within Emily's Case

In reviewing Emily's case materials, two major themes emerged about how she justified using most significant change when examining causality, as well as how her evaluation context influenced stakeholders' perception of the credibility of the

methodology. These two themes were: 1) power and privilege and 2) values and credibility.

Power and Privilege. For Emily, implementing most significant change as a staff member of an international NGO working with West African communities highlighted issues of power and privilege in evaluation. She was aware of how her presence as a "white foreigner" might have influenced the study. She recalled that when she left her position and local West African staff took over her responsibilities, the staff let her know that certain pressures to please her were gone. As she said, *"*I definitely got some anecdotal feedback about how it was so much better because people could ask questions, and they didn't feel that they had to behave a certain way." While in that position, she also had a sense that she might be missing something because of her own lens, or that people might withhold information from her because of her status as a foreigner working for an NGO that funded projects in the community. As she said, *"*I think the most pertinent example of power and privilege came out in the staff training and the fact that I was the one analyzing all of the data, so what was being lost in translation, and what were people not telling me as far as what was going on in the field." Ultimately, she believes that the more that evaluation projects can be implemented by locals, the more the dynamic of the foreign evaluator's power influencing the study can be addressed. As she said, "It's so important for people from the country to be in charge of these types of things because no matter what you say, you just come with power and privilege."

The long-standing presence of NGOs in the region also created a dynamic in which community members might mask the truth in favor of telling NGO staff what they

think they want to hear. Telling NGOs what they want to hear was a mechanism for survival; as Emily said, “people were trained through all of these years to tell people what they wanted to hear, so they could get things that they needed, desperately needed.” One strategy the team implemented to counteract this phenomenon was to encourage participants to elaborate on their stories through probing techniques. The resulting conversations increased Emily’s confidence in the data. As she said, “I think actually, with qualitative, you’re more likely to get at the truth if you have the space to probe with people and to have a conversation.” While the reality of participants distrusting or wanting to please the NGO staff was still there, this dynamic was lessened through the probing technique.

Additionally, Emily sought to address power and privilege by being thoughtful about who from her team conducted the focus groups and interviews. Being a white evaluator working in a West African country that had been colonized by white Europeans, she was aware of how her presence influenced how community members interacted with her. She was aware that community members might be distrustful of her; as she said, “My physical presence, there’s nothing I can do about it. There are centuries of oppression and colonization and histories that people have with other people that look like me that I cannot combat.” Additionally, she did not speak the local language. For these reasons, data collection for the evaluation was completed by NGO staff who were also members of the same ethnic group that was engaged in the community projects. As a result, they were able to follow cultural protocols and customs in order to connect with community members and collect the data.

Values and Credibility. Emily stated that some of the methods present in other evaluation studies, such as collecting data at structured timepoints and establishing a comparison group, were not feasible in the context where she conducted her evaluation. She expressed that structuring data collection and reporting around rigid timelines can be in conflict with the needs of the community in which the evaluator is working. As she remarked, "People are going to bring up concerns when they feel comfortable to do so, when you've built trust and relationships, and it doesn't always line up with evaluation timelines and program timelines." She noted that this is particularly true when implementing participatory methodologies, because working with the community to gather their feedback and ensure that findings are shared back to them takes time. Additionally, establishing a comparison group raises ethical considerations for her. In her evaluation, she and her team consciously did not engage with any communities where they could not provide services, because this would be too extractive and did not align with their community-led model.

Emily thinks that most significant change is well suited to the complex conditions of doing evaluation in an international development context in West Africa. She thinks the methodology is an effective way to deeply understand the context of an intervention. As she expressed, "If you're really interested in learning about your program, and you have the bandwidth to dive into some of the complexities of what international development is, then MSC is a great fit." She also stated that practitioners within the community-led development field are open to the idea that most significant change can produce credible evidence of impact. As she noted, within this field, "There's just an

acknowledgment that every community is so unique and has so many contextual factors and that …. There's no pressure to generalize." However, she also expressed that there is sometimes tension between choosing a methodology that is widely seen as credible and choosing a methodology that is best suited to the organization she is working with. For her, this means needing to strike a balance between using post-positivist methods (typically quantitative in nature) and using more participatory, collaborative methods (typically qualitative in nature).

Lori's Story Using Most Significant Change

Lori conducted a most significant change study to evaluate the outcomes of a conservation leadership program operated by a national environmental organization. The program engaged with early career employees (whom they referred to as "fellows") to train them to become leaders in environmental conservation work. The program also included a component in which each fellow implemented a conservation project within their community (these communities were located across the country). The national environmental organization sought Lori out to conduct an evaluation to determine the impact of the program on the fellows and on the community members where the conservation work occurred.

At the time of the evaluation, Lori was working for a small evaluation firm. She was the lead evaluator on the fellowship program evaluation, and decided to implement a most significant change evaluation because it could produce the kind of evidence that the national environmental organization needed – evidence of the impact of their program across multiple sites. She was also drawn to most significant change because she wanted

to train the fellows to conduct part of the evaluation themselves and she thought it was a fairly easy methodology to understand.

Lori liked to use a participatory approach in her evaluation practice. As she described it, “I always like to develop tools with the clients I work with, get their input, make sure it’s going to be something that works for them.” Her practice also included evaluation capacity building; she frequently worked with organizations to build their evaluation skills, teaching non-evaluators how to do evaluation. Another approach that she brought was the utilization-focused approach, in which evaluation is conducted in a way to ensure that the findings will be used. Finally, she also shared that at the time she “was already starting to pay attention to culturally responsive evaluation” as well.

One key stakeholder in the evaluation was the national environmental organization who commissioned the study. At the time of the evaluation, the organization had recently shifted its focus to conservation because of the urgency to address climate change. Lori reflected on how this urgency led to the creation of the fellowship program:

> Also, I think this may be a sociopolitical — I don’t know exactly what to call it — but climate change and the urgency of climate change. Really, I know that has pushed the environmental organization to focus more on conservation. That’s just part of the general context of the creation of this program and the importance of the program to them.

The fellowship program was meaningful to the organization because it was part of their strategy to address climate change. As Lori explained, “They’re hoping to really use this program to help some of their young leaders who were working in the organization

throughout the country to have more of a focus on conservation." Additionally, because it had a leadership component, the fellowship program fit into the larger organizational culture of seeking to grow and retain employees. As she remarked, the project was focused on "young leaders" and the organization generally wanted all of their employees to "do good work and do good things and work up the ranks."

The organization's leadership hoped to utilize the findings to help assess whether this shift to focus more on conservation was "the right fit for the organization." If the findings showed impact, that would influence their decision of whether to continue operating the program. There was also the possibility that the funder would discontinue funding; thus, the organization was deciding whether the impact of the program warranted funding it through other means. Lori was aware of how that dynamic influenced the most significant change study. As she said,

> The financial situation of the funder potentially taking away money definitely impacted what they were looking for. I find that almost whenever I do evaluation, but in particular, in high stakes like that, I hear the clients say, "I'd like to prove that this program is effective." I always say, "Well, as the evaluator, I would say you'd like to investigate the extent to which this program is effective because maybe it is and maybe it's not effective in every way. That's also good to know." There was definitely that financial stakes that were pushing it.

Her statement reflects her belief that it is rare that a program is 100% effective; rather, there might be parts of the program that are effective and others that need improvement.

The organization's leadership wanted to know what parts of the program they might be able to improve. As Lori said, "They really cared about how to make the program better and stronger and what was working well and what wasn't working well." The funder of the fellowship program, which was a large, private-sector company, was also interested in seeing data regarding what could be changed to improve the program. As for the fellows, their stake in the evaluation was that they had an interest in knowing what outcomes resulted from their projects, from the perspective of community members.

Because the fellows were located across the county, Lori and her evaluation partner conducted most of the evaluation work remotely. As this was a participatory evaluation, she and her partner trained the fellows in how to conduct most significant change interviews and the fellows conducted the interviews. The evaluation team then reviewed the interviews for cross-interview themes and brought those themes to an in-person discussion with the fellows. At that meeting, the evaluation team facilitated a discussion to identify which stories captured the most significant change resulting from the fellowship as a whole. Conducting the evaluation in this way met Lori's goal of building the evaluation capacity of the fellows to the point where they could use most significant change entirely by themselves in the future.

When asked whether the environmental organization utilized the findings, Lori was confident that they did. She had an opportunity to travel to the organization's office and present the results in front of leadership. She shared that, "They were really interested in the results." Additionally, she felt that having the fellows collect the impact

stories broadened their understanding of how the fellowship as a whole had an impact. As she said of the fellows,

> We got to have that in-person meeting with them where they got to look through and really think about the most significant change stories that they heard and thinking about it across — not just for their particular project. I think that it had an impact on them. I don't know if they continue to use that approach. I'm not sure exactly how it continued, but I do think that they learned something in that process.

Themes Within Lori's Case

In reviewing Lori's case materials, three major themes emerged regarding how she justified using most significant change when examining causality, as well as how her evaluation context influenced stakeholders' perception of the credibility of the methodology. These three themes included: 1) the methodology highlights factors that contribute to change; 2) the methodology has multi-cultural applicability; and, 3) the methodology can be implemented in a participatory way.

The Methodology Highlights Factors that Contribute to Change. Lori was drawn to most significant change because she believed it could provide evidence of what factors may have contributed to change in a person's life. For her, generally the right question to ask is how something contributed to change rather than trying to attribute change to something. As she said, "The first time I heard that phrase 'contribution, not attribution,' it was like this aha moment." She expressed that it can be difficult to track and control all of the things that might have caused change to happen, so it is more

appealing to think about understanding how a factor contributed to change rather than attributing change to a factor. As she expressed, “I love this idea that we can look and see if a program contributes to a particular outcome, but it won’t be the only thing that has led to that outcome in someone’s life.”

In contrast, thinking about causality as if there is a clear linear connection between one causal factor and one effect does not align with how Lori sees the world. She gave the example of the logic model to illustrate this. She stated that logic models assume that actions and reactions occur in a tidy, linear fashion; but in her view, real life does not progress that way. Referring to logic models, she remarked “It’s like this will lead to this will lead to this will lead to this, and that’s how you get to that. Life isn’t like that, and also there’s lots of other things happening in the lives of these participants that might also contribute to that.” Her sentiment is that context can influence whether change occurs, context is ever shifting, and people cannot be extracted from their context. As she remarked, “We don’t want to extract people from everything else in their lives and just put them in this little box and have them go through this program and see if that leads to this thing.” Thus, approaches within causal evaluation that require controlling for the effect of context and isolating the effect of a single causal factor are not credible to her.

Lori explained that trying to understand the causal contribution of a program involves considering the impact of that program within an ever-shifting context. As she says,

> We want to know “Is our program, given all the other things that are happening in your life, likely to lead to that outcome?” It feels like a more holistic way of

> looking at things, and that really appeals to my way of thinking about the world, which isn't that we want to isolate and get at the perfect thing because even assuming that one could do that, which I don't think you could, but even assuming that one could do that, it's not replicable in life because people are experiencing other things.

As she elaborated, examining causality from a lens of contribution rather than attribution involves thinking about the impact of personal motivation as well as any situational or structural conditions within a person's life that may hinder or facilitate change. She had relinquished the idea that attributing change to one single factor was even possible. As she remarked, "I don't think that there's ever one thing that leads to all the rest of the things in our lives." She believes that the right question to ask in causal evaluation is not about attribution, but about contribution, "because we know there are other things that also contribute to that outcome."

According to Lori, most significant change can help uncover what factors people think contributed to change in their lives. She explained that through the process of recounting their own story, people are able to think meta-cognitively: they analyze their life choices and the events that led to an outcome. Something about recounting their story helps individuals articulate "how experiencing something led to an aha moment" or "how this particular experience made an impact on them and led them to do other things." Thus, according to her, the kind of evidence that most significant change can produce is evidence regarding what individuals think contributed to change in their lives. Additionally, she stated that the methodology can highlight which aspects of the program

may not be contributing to change in people's lives when these aspects do not show up in their accounts.

The Methodology has Multi-cultural Applicability. Lori had conducted most significant change studies in settings where she was not from the same community of the program staff or participants. In some of these settings, she did not speak the same language and was not of the same nationality as the staff and participants. She stated that in these multi-cultural settings the methodology worked well because it elevated and valued the cultural lens of the program staff. In a rural international evaluation she conducted, she partnered with program staff to help identify and recruit participants by drawing on their existing relationships. She said this was particularly helpful because "If we had done that ourselves, that would have taken months and months and months of our time to do that." Furthermore, she said that when doing a most significant change study, the evaluator can also partner with program staff by asking them to conduct interviews or translate participants' stories. Additionally, within the methodology there is a step when program staff rank which stories are most illustrative of the program's impact. As she explained, the evaluator does not derive what success looks like across the stories through thematic analysis; rather, it is program staff who make decisions about what impact looks like and which stories best illustrate how the program creates impact.

Finally, Lori also expressed that she thinks the central evaluation question that drives the methodology translates well across different languages and cultures. The central question in most significant change is, "What is the most significant change?" About this question, she said, "It does translate well. Literally translates in other

languages well. Culturally, it works for the interviewees to understand what they're being asked and for the people who were doing the interviews to know what to ask." It is the simplicity of this central question that makes it easier to apply it across different cultural contexts.

The Methodology Can be Implemented in a Participatory Way. Lori had implemented most significant change at least twice in a participatory way, partnering with program staff. She expressed that the methodology is not hard to understand; as a result, it can easily be implemented by program staff who do not have evaluation training. As she said, "The question is 'What is the most significant change?' The person who is listening or who is answering that question can answer that question. The person who's asking the question - it seems like a fairly simple, straightforward thing to ask." Additionally, the process of gathering and re-telling a story is easy to understand because it is similar to journalism, a widely understood medium. As she said, "You get into that journalistic approach of the what, when, where, why, how, so that there are kind of ways to follow up on that." Thus, her opinion is that the approachability of the question and the familiarity of storytelling make this methodology easy to implement in a participatory way.

Cross-Case Findings

The next section presents the cross-case themes that emerged from a review of the case materials from all four evaluators. These cross-case themes present the commonalities found in the arguments that the evaluators made to justify using story-based causal methodologies. These themes included: 1) type of evidence needed; 2)

paradigm as determinant of perceived credibility; 3) truthfulness of accounts; 4) triangulation among data sources; 5) elevating participant voice; 6) unexpected outcomes or mechanisms; and, 7) experience matters.

Theme 1: Type of Evidence Needed

Using a story-based causal methodology is a credible choice when the methodology can produce the evidence needed to satisfy the purpose of the evaluation. The evaluators described the kinds of contexts in which story-based methodologies are a credible choice. These contexts include: 1) learning rather than proving; and, 2) contributing factors.

Learning Rather Than Proving. The evaluators shared that story-based causal methodologies are a credible choice when the purpose of the evaluation is to learn about the program in order to improve it. In Dan's case, the aim of the study was to learn about the success that practitioners were experiencing when implementing the tobacco cessation guidelines; the aim was not to prove anything. Emily said something similar about her experience. As she said, "The goal was not to publish, the goal was really just for internal programmatic learning." Similarly, Claire stated that the purpose of a success case method study and the purpose of evaluation generally is to improve a program by learning about what's working. Thus, these evaluators expressed that the story-based causal methodologies are a strong choice when the purpose of an evaluation is to learn in order to improve the program.

Contributing Factors. The evaluators expressed that story-based causal methodologies are a credible choice when stakeholders want to know what factors

contributed to change. Both Dan and Lori framed the story-based causal methodologies as ways to uncover evidence of what contributed to change. As Dan said, “We were learning, in the cases of these successes, what actually happened, what part of the education contributed, in particular, among all the other contributors.” He believes that evaluations should not only produce evidence of attribution, but also of contribution. For him, success case method helps answer the question, “When change occurs, how did the educational intervention contribute?” He expressed that he believes that attribution-type studies should be combined with contribution-type studies to fully understand whether and how an intervention worked. As he reflected, “It’s those two things put together that give you a much fuller understanding of the impacts of these interventions and gives you much better guidance on how to improve it in the future.”

Similarly, Lori stated that most significant change is well suited to uncover contributors to change. In particular, she thinks the methodology produces evidence of how a program contributed to changes in a participant’s life, from the participant’s perspective. For example, one of Lori’s participants shared that while participating in the program she expanded her awareness of species migration because she connected these ecological concepts to her own story of migration. Species migration is impacted by conservation efforts. Thus, this participant experiencing an increased awareness of species migration meant that they experienced the program impact of increased awareness of the need for conservation. Lori provided this participant’s story:

> This one woman talked about - she had immigrated from Mexico with her family to the United States. One of the things they started to notice from planting things

> was they planted a butterfly garden, and they noticed the monarchs coming. She told this really beautiful story of how her nuclear family that was doing the work was really getting so much out of being together and doing this thing that was so different than what they had done. Also, that she felt like her story was mirrored in the story of the monarch and the migration of the monarch butterfly, which goes down to Mexico and all the way up to Canada and has this story of moving. The way she said was just so powerful, how it made her feel. Not just like, "Oh, it's fun to be in nature and feel more connected nature," but in this whole other ecosystem piece of feeling connected to other animals who had a similar story.

This example demonstrates how most significant change can uncover these causal connections between how an intervention interacts with an individual's personal context to create change, as told from the participant's perspective. Claire provided a similar observation that participants can see the connections between an intervention and a change in their lives and that their accounts are credible. As she said,

> If it's a success case, you can share it too. Then people are proud to share. Then they have the evidence right there. Did your sales improve? If the person says, "It's because I'm applying this element from this training," why would there be any doubt? Why would there be any question about credibility?

Thus, Lori and Claire hold a similar understanding that story-based causal methodologies can provide evidence of causal contribution by highlighting what factors caused change from the participants' perspectives.

Theme 2: Paradigm as Determinant of Perceived Credibility

The evaluators shared the observation that whether or not a stakeholder perceives story-based causal methodologies as credible for examining causality depends upon the stakeholder's paradigm. Lori described some stakeholders as "qualitative skeptics"; she said that there were qualitative skeptics among all types of stakeholders "who just don't believe in qualitative at all." When speaking to qualitative skeptics about the credibility of most significant change, she would begin with a larger conversation about why qualitative evidence is valid before explaining why the particular methodology is valid.

Claire also encountered qualitative skeptics when discussing success case method with stakeholders; however, she perceived that the skeptics tended to be academics or peer evaluators. She had a counter-argument to these stakeholders; she argued that there is widespread recognition that observational evidence is credible – and observational evidence is often qualitative. She provided an example of how observation is a powerful form of evidence of causality, in particular when one is observing physical reactions: "You are at a billiard table, someone takes the stick and hits the white cue ball and the black cue ball, and the black cue ball goes in. What other evidence do you need?" In this example, it would be hard for the observer to argue that another causal explanation was at play.

She also made the argument that investigators rely on qualitative evidence when building a case against someone thought to have committed a crime; in general, people perceive these cases as credible. As she said, "What do investigators do in a criminal investigation? They look at footprints. It's not quantitative data, it's qualitative data. You

still trust those people, why do you trust them?" Thus, Claire argued that since observational evidence is perceived as credible, and observational evidence is often qualitative evidence, it is possible for qualitative evidence of causality to be viewed as credible.

Similar to Claire, Dan perceived that the stakeholder's paradigm influenced whether they perceived results from success case method to be credible. When presenting information about the evaluation or the findings, he and his team would present the full case for trustworthiness, but highlight certain validity procedures depending on the audience and the assumed paradigm of that audience. For example, when a group of experts with more of a positivist leaning were the audience for the findings presentation, the team highlighted the quantitative step of the study which occurred when they selected the successful case stories.

Finally, Claire and Emily remarked that non-academic stakeholders, such as donors, staff or beneficiaries, tended to be less skeptical of the methodologies. Donors tended to accept that the findings in the evaluation report were credible, in Emily's experience. Similarly, Claire said, "Donors, funders, well they fund you to do this, to produce that kind of evidence because they trust this kind of evidence." Claire also stated that the participants she worked with also trusted this kind of evidence, namely because they were the ones recounting the stories.

Theme 3: Truthfulness of Accounts

The evaluators believe that participants can generally be trusted to tell the truth when they recount their personal stories. Claire's belief that people tell the truth was

unqualified. She believed that in the success case method setting, the participants have no motivation or reason to lie to the evaluator. She also argued that because participation is voluntary, they are not being coerced to participate and thus are more likely to be truthful. In her experience with the methodology, people were just as likely to speak about positive outcomes from a training as they were to voice negative experiences with supervisors. As she stated about her experience with the methodology, "People tell you the truth in general." However, she qualified her statements by saying that it is crucial that an external evaluator implement the methodology as this gives participants more confidence that their stories will remain confidential.

Dan and Emily also believe people generally can be trusted to tell the truth; however, both evaluators relied upon triangulating participants' stories with other forms of data to have more confidence that the success and impact stories were real. Dan said that he and his team were aware that participants may be tempted to frame their stories in a "positive light." This made the team's steps to triangulate the interview data with medical records or other sources of data all the more important. He also shared that when the team identified cross-case themes from different sites that gave them more confidence that the participants weren't fabricating or falsifying their success or impact. In Emily's case, the communities she was working with had a history of telling international NGOs what they wanted to hear in order to receive needed support; this dynamic challenged her certainty that participants were always telling the truth. However, through the process of validating the individual stories by speaking to participants multiple times and

conducting peer interviews, the team built more confidence that the individual stories were trustworthy.

Similarly, Lori was concerned about participants wanting to please her with their answers. However, because participants told her about outcomes that she was not expecting to hear, she felt more confident that they were not trying to please her. As she said of the outcomes the team heard, “They weren’t things that anyone was looking for.” Additionally, she witnessed participants reflecting on their experience and the different factors that motivated their actions; this meta-cognitive processing led her to believe that the participants’ stories were credible.

Her academic training encouraged her to question people’s ability to understand how context influences them. As she said, “People are not always the most reliable reporters and observers of our own lives…we don’t always recognize the things that might be working on us.” However, after conducting the most significant change study, she recognized that this degree of reflection, of understanding how context influences our choices and actions, was occurring for participants. As Lori said of participants, “Some revealed that they were really doing some deep thinking. That meta-cognitive piece that I wasn’t sure they were going to be able to accomplish showed up.” Thus, one of her key reflections was that when participants themselves draw connections between factors in their lives and changes that occurred, it makes their stories more believable.

Theme 4: Triangulation Among Data Sources

The evaluators all shared the belief that story-based causal methodologies should include a component of data triangulation to assure confidence in the credibility of the

findings. All four evaluators utilized procedures to corroborate the stories collected from participants. These procedures entailed triangulating the data collected from the stories with other sources of data. Sometimes, the evaluators triangulated the qualitative data collected through the participant stories with quantitative administrative or survey data. In Dan's case, he triangulated interview data with health records detailing how an intervention was implemented. Being able to review these records helped ease any concerns that participants were providing socially desirable answers. As Dan said, "We could see artifacts that gave us confidence that they weren't just yessing us." Additionally, within his evaluation, the identification of successful sites was based upon clinical quantitative data, which added credibility as it combined qualitative and quantitative methods together in the study. As he said, "That study would never have the credibility we did without such strong evidence that cases were successful."

Claire also triangulated participant stories by using existing administrative data to confirm elements of the stories. As she said about her experiences implementing success case method to evaluate trainings, "Sometimes we get organizational data for triangulation. We can look at completion, depth and breadth of completion of courses, and amount of participation." Similarly, Emily conducted a quantitative survey of members in the community where she conducted the most significant change study to have additional evidence about how change may have occurred. Thus, Dan, Claire and Emily utilized quantitative data, either available through existing administrative data or through surveys, to help support findings drawn from the qualitative data.

Additionally, the evaluators utilized qualitative data collected from staff, peers, or other participants to triangulate the qualitative data collected in the participants' stories. In Claire's case, she corroborated stories by hearing about success in one individual's story and then asking other individuals whether they experienced something similar. As she framed it, "We do the success case interviews and sometimes we say, 'Oh, we heard from someone this and this is happening. Have you seen anything like that also?'" Emily also utilized qualitative data to corroborate the qualitative stories, first by asking peers whether they experienced an impact story in the same way as the individual who told it, and second by asking individuals at community meetings whether findings from her most significant change study resonated. Finally, Lori also utilized qualitative data to corroborate the qualitative impact stories by triangulating the fellows' perspective with the participants' perspective (both perspectives were gathered through interviews).

Theme 5: Elevating Participant Voice

The evaluators had a shared understanding that story-based causal methodologies include procedures that elevate participant voice. One way the methodologies elevate participant voice is to give the participant power in directing the conversation. Dan shared that because the interviews they conducted with success case method were semi-structured and open-ended, it allowed the participant to take them on a journey and direct the conversation, to some extent. As he said, "We followed where they wanted to go to some degree. We let them make decisions about what elements should be included in this story." This approach helped elevate participants' voices by giving them some power to share what they felt were important factors for success. Similarly, Emily shared that most

significant change leaves it up to participants to decide and articulate how an intervention changed their lives. Because the central evaluation question in the methodology is so broad, it gives a participant the space to articulate what part of an intervention was most important to them. As she said:

> I think really just sitting down and asking, what's the most significant change that has happened in your life since, in this case, when this NGO came to your community? It makes it broad. It's not asking about how are you directly benefiting? It's not asking how have your schools changed, how has your health changed? It's really leaving it up to them to decide.

For both Emily and Dan, the methodologies encouraged participants to share what was most important for them.

Claire and Lori also expressed that the story-based causal methodologies elevate participant voice. Claire stated that success case method elevates participant voice because it centers the participant and their experience. Lori also remarked that most significant change elevates participant voice, but not any more than other qualitative methodologies. The one difference between most significant change and other qualitative methodologies, she pointed out, is that within most significant change one focuses on the full story and context of the individual participant, while in most other qualitative methodologies the full story of each participant isn't told.

Theme 6: Unexpected Outcomes or Mechanisms

The evaluators said that findings from the story-based causal methodologies revealed unexpected outcomes or unexpected mechanisms leading to outcomes that

occurred. This aspect of the story-based methodologies justifies their use for causal examination, because other causal methodologies cannot uncover these unexpected outcomes in quite the same way. For Lori, every time she implemented most significant change, it brought up outcomes she was not expecting to hear. The evaluation of the fellows project brought to light several outcomes that neither she nor the fellows anticipated would have resulted from conservation awareness projects, such as an increased sense of empowerment and an increased sense of connection to family and community. Similarly, Claire voiced that sometimes when conducting a success case method evaluation of training effectiveness, she is surprised that the training is effective for certain groups, such as those who have been working for many years and have experienced many trainings. For her, success case method can reveal unexpected outcomes such as these because you are providing space for participants to drive the conversation. As she said, "We learn those kinds of things because you just let people talk and follow up and it may be surprising sometimes."

In Emily's case, most significant change did not reveal any unexpected positive outcomes; however, the methodology did reveal unexpected potentially negative outcomes. For example, the evaluation team observed that new power dynamics were being introduced between the community volunteers implementing the project and local leadership. That was not their intention; so, that was a risk that the team realized they needed to mitigate.

In addition to revealing unexpected outcomes, the story-based causal methodologies reveal unexpected mechanisms that lead to change. When Dan

implemented success change method he was able to uncover mechanisms that actually prevented change from happening. He learned that even straightforward guidance on a medical intervention can be hard to implement because changing people's day-to-day behaviors is difficult. As he said, "What our studies show, and this was a surprise, I think, was how challenging it was for them to implement even simple changes."

Similarly, Claire stated that success case method uncovers unexpected mechanisms for how an intervention produced change. She explained that in instances when one region sees successful outcomes from a training but the other doesn't, success case method can help uncover why that is through the individual interviews. As she said, the reason why a training was successful in a particular region "sometimes comes as a shocker." The participants often reveal contextual or cultural factors specific to that region that contributed to change that the evaluation team was not aware of, according to her. In this way, success case method can reveal the unexpected mechanisms behind the success of a training.

Theme 7: Credibility from the Evaluator or Organization

While the evaluators had encountered skepticism that the story-based causal methodologies were credible in other settings, none of them had to justify that the methodologies were credible to the particular stakeholders involved in their case stories. Their stakeholders thought their studies were credible for reasons beyond the choice of methodology, such as: reputation of the organization that they worked for, their academic credentials, or their ability to proactively address any credibility concerns.

Claire did not typically have to justify using success case method to stakeholders because the organization she worked for was sought out to conduct the methodology. Referring to this dynamic, she said, “The credibility of the method is established before you even start out.” Additionally, the clients she worked with tended to review the credentials of the evaluator; in her case, it helped that she had a PhD in evaluation. As she explained, this credential also helped her relay to stakeholders the advantages and disadvantages of different evaluation approaches, which enhanced her credibility in their eyes.

Emily did not encounter any stakeholders challenging the credibility of most significant change because the organization had been using the methodology for a while. Interestingly, she experienced the inverse of incredulity among stakeholders; she had to caution stakeholders that the findings from most significant change could not be taken as evidence of causal inference in the same way that evidence from a randomized controlled trial might be taken. In both the cases of Emily and Claire, it appeared that the credibility of the organization and/or of the evaluator mattered in terms of whether stakeholders perceived the methodology as credible.

Dan also did not experience any stakeholders saying that success case method was not credible. However, during the proposal process, he took steps to proactively explain why the methodology was credible. As he explained, “We proposed the study in the initial grant, of course, we addressed the question of the suitability of the success case method and how trustworthy any results coming out of that might be.” He also made sure that his research team was well aware of which procedures contributed to the validity of

the findings. As he said, “Working with my research team, I had to be very explicit about what elements of our process were oriented around building a good case for trustworthiness.” He also proactively addressed how procedures within the study contributed to validity when presenting findings to partners in the project or within the academic community. These actions were crucial in heading off any criticism around the credibility of the study. As he said, “It wasn’t like anyone stood up and posed a direct challenge to our methodology, but we were very mindful all along of what we needed to do to strengthen the case.”

Summary

This chapter included two sections. The first section described each evaluator’s implementation of a success case method or a most significant change study and presented within-case themes. The second section presented the cross-case themes, which included: 1) type of evidence needed; 2) paradigm as determinant of perceived credibility; 3) truthfulness of accounts; 4) triangulation among data sources; 5) elevating participant voice; 6) unexpected outcomes or mechanisms; and, 7) credibility from the evaluator or organization. The next chapter discusses how the themes answer the study’s research questions and how the themes relate to and expand upon the existing literature.

CHAPTER V: DISCUSSION

The purpose of this study was to examine how evaluators justify using story-based causal methodologies to answer causal questions in evaluation. The methodology chosen to examine this topic was multiple case study because that methodology is well-suited for exploring cross-case themes. This chapter discusses how the study's findings answer the two main research questions: 1) What arguments are made by evaluators to justify the credibility of story-based causal methodologies to evaluation stakeholders?; and, 2) From the perspective of evaluators, how do contextual factors influence whether story based causal methodologies are perceived as credible by evaluation stakeholders? The chapter also situates the findings within existing literature pertaining to the topic of credibility as it relates to story-based causal methodologies. Finally, this chapter presents a practitioner's guide for evaluators to use in contexts in which they need to justify their choice of using story-based methodologies in causal examination.

Question 1: What arguments are made by evaluators to justify the credibility of story-based causal methodologies to evaluation stakeholders?

Story-Based Causal Methodologies Answer a Specific Causal Question. One argument that the evaluators made to justify using story-based causal methodologies when examining causality was that these methodologies can credibly answer a specific type of causal question, which is: What are participants' perspectives of how an

intervention changed their lives? Lori believes that most significant change effectively answers this question. As presented in Chapter IV, in Lori's within-case theme of The Methodology Highlights Factors that Contribute to Change, she witnessed participants having "aha" moments in which they illustrated how a causal connection might be at play between the intervention they received and the outcomes they experienced. This was particularly salient in the example of the woman who observed the butterfly's migration pattern and likened it to her own migration and connection to a larger ecosystem. As discussed in Chapter IV within the cross-case theme of Type of Evidence Needed, Claire also expressed that success case method can effectively capture participants' perspectives of how change occurred for them.

Lori and Claire's observations resonate with existing literature about the kinds of causal questions that story-based causal methodologies are best equipped to answer. Gates & Dyson (2017) discuss how the narrative way of thinking about causality (with which story-based methodologies are aligned) has a central driving question, which is: "According to stakeholders, what influence, effects, and/or difference did the intervention make for their lives?" (p. 37). Lori and Claire stated that they do indeed utilize story-based methodologies to examine patterns of causality from the perspectives of participants in the study.

Unexpected Outcomes Reduce Evaluator Bias. A second argument that the evaluators made to justify using story-based causal methodologies when examining causality was that these methodologies reveal unexpected outcomes and mechanisms that the evaluator might otherwise be unaware of. The evaluation process is subject to the bias

of the evaluator. Throughout all stages of an evaluation, evaluators make decisions about the procedures of evaluation design, data collection, analysis and writing. Through these decisions, the evaluator places boundaries on what kind of information is and is not gathered or represented. Because the evaluator makes these decisions from their own lens and paradigm, they will necessarily be excluding a wider range of information and interpretations.

However, with story-based causal methodologies, this influence from the evaluator is greatly lessened because participants recount their story and are given more latitude to describe what outcomes occurred for them and how those outcomes occurred. As discussed in Chapter IV under the cross-case theme Unexpected Outcomes or Mechanisms, Lori shared that every time she implemented a most significant change evaluation, participants described outcomes that she was not aware she should be asking about. For example, she did not expect that the conservation project she was evaluating would produce increased sense of connection within a family. As discussed in that same cross-case theme, Claire also stated that in her general experience, success case method gives participants latitude to drive the conversation. Sometimes, participants even revealed unexpected explanations for why they were seeing success. Dan also concurred that his experience with success case method allowed him to uncover unanticipated reasons for why the intervention was not successful in some cases. Story-based causal methodologies allow participants to express their own accounts of how success or impact occurred; in doing so, they break through the boundaries that an evaluator might have placed on the knowledge gained.

This dynamic of increasing participant voice to lessen the potential bias of the evaluator accords with existing literature about multi-cultural validity. La France et al. (2015) assert that the construct of validity (similar to the construct of credibility) is culturally-bound and culturally-defined; as such, all cultures do not share one universal concept of what is valid. When evaluators make decisions about what questions to ask in the data collection phase or what data to include in the analysis stage, they are applying their own cultural lens to make decisions about what counts as credible evidence. Within story-based causal methodologies, evaluators treat participants' accounts of what occurred for them as credible evidence. In so doing, evaluators broaden the understanding of what outcomes occurred from an intervention and how an intervention works in practice.

Having the mindset that participants' accounts are credible increases the overall multi-cultural validity of the evaluation, particularly in those settings where the evaluator is located in a different culture than the participants. Viewing participant accounts as credible evidence may be particularly important when the evaluator holds more power and privilege than the community partners she is working with. As explained in LaFrance et al. (2015),

> Historically, validity has been situated within the social history and culture of dominant groups, such that the legitimizing function of validity… reflects and reinforces that history and power, with negative consequences for persons in nondominant groups (p. 50).

Utilizing story-based causal methodologies to showcase participants' stories and to lessen the evaluator's decision making around whether stories are credible is one way to

increase the multi-cultural validity of a causal evaluation because "multiple means of argument and validation" (LaFrance et al., 2015, p. 57) are examined.

Participants Can Be Trusted to Provide True Accounts. A third argument that the evaluators made to justify using story-based causal methodologies when examining causality was that participants can be trusted to provide a true account of their own lives. Since story-based causal methodologies rely so heavily upon participant stories to elucidate instances of success or impact, believing that participants can credibly relay their accounts is foundational. As discussed in Chapter IV under the cross-case theme Truthfulness of Accounts, Claire stated that participants in a success case method study have no reason to lie, because their stories are being collected by an external evaluator who will keep their identities confidential. Additionally, in her experience implementing the methodology she perceived an equal likelihood of participants sharing negative as well as positive feedback. As discussed in that same section, Lori was convinced that participants were not simply telling her what she expected to hear because she heard so many surprising stories of how outcomes occurred. Additionally, observing participants depict how the intervention influenced their choices and actions convinced her that participants were meta-cognitively examining how the intervention and other factors influenced them. For her, observing participants use a meta-cognitive approach to re-tell their stories added to the stories' credibility.

These findings align with existing literature describing how story-based causal methodologies are rooted in the approach of narrative inquiry, which holds that participants are capable of truthfully and comprehensively recounting their stories. Abell

(2004) states that writing about one person's story and how their actions led to results is "unimpeachable" (p. 296) evidence that change occurred for that person in that way; the caveat is that one person's causal story is not generalizable. Additionally, van Wessel (2018) argues that stories are a credible form of evidence, remarking that an individual's story can clarify their response to an intervention through how they describe their attitudes, understanding of something, or their behaviors. van Wessel (2018) also argues that stories are the only form of data that can draw connections between how outcomes build upon other outcomes to contribute to change over the course of time.

However, in the social sciences it is well known that participants may sometimes say what they think the researcher wants to hear. There is also the dynamic of participants being embarrassed or ashamed of the truth; thus, they may mask the truth when telling their story. This dynamic of hiding the truth from a researcher is sometimes termed "social desirability" (Fowler, 2014, p. 94). Indeed, Emily discussed in her within-case theme Power and Privilege that the communities she worked with had a tendency to provide socially desirable answers, driven by a need to receive resources from NGOs. However, her team was able to counteract this dynamic by asking follow up questions. While social desirability is an often-encountered dynamic, the evaluators maintained that participants are capable of telling truthful accounts and were more likely than not to do so. They also conceded that particular conditions increase the likelihood that participants tell the truth, including having an external evaluator who will keep identities confidential and including qualitative probing. Additionally, as discussed in the next section, all the evaluators included procedures to triangulate their data sources, indicating that even

though participant accounts can be trusted, gathering additional perspectives is a crucial step to building a more comprehensive picture of how a story unfolded.

Story-Based Causal Methodologies Include Triangulation of Data. Another argument made by the evaluators to justify using story-based causal methodologies when examining causality was that these methodologies include procedures to triangulate the data. When multiple sources of data support the same conclusions, it increases the credibility of the findings. Triangulation entails gathering multiple sources of data and cross-checking understandings gleaned from one source with another; within qualitative research, triangulation is a recommended procedure to enhance the credibility of a study (Anfara et al., 2002; Tracy, 2010).

As discussed in Chapter IV under the cross-case theme Triangulation Among Data Sources, all four of the evaluators included some component of triangulation in their studies. Dan triangulated data from interviews with health record data that demonstrated how the intervention was implemented. Similarly, Claire compared existing administrative data with story elements to corroborate the stories. Emily implemented a survey of members in the community to supplement the stories she gathered. Finally, Lori triangulated data from participant interviews with data from the fellows' interviews to build a deeper understanding of the program's achievements.

Given that triangulation is a well-known procedure for enhancing the credibility of findings from a qualitative study, it is not surprising that all four evaluators employed it. Additionally, Dart and Davies (2003) and Brinkerhoff (2003) recommend employing triangulation to strengthen the credibility of findings in most significant change and

success case method, respectively. Within the literature, there are also examples of story-based causal methodologies that employ triangulation to corroborate findings (Limato et al., 2017; Medina et al., 2015).

Sub-Question 1: Do the arguments that evaluators make to justify the credibility of these methodologies differ depending on which evaluation stakeholder is the audience for the argument? The evaluators in the study shared that it wasn't typically stakeholder type that made a difference in whether a stakeholder thought the methodology was credible; rather, it was the stakeholder's paradigm that seemed to matter. As Lori discussed in Chapter IV under the cross-case theme Paradigm as Determinant of Perceived Credibility, there were "qualitative skeptics" across all stakeholder types (donor, beneficiary, program staff, etc.) when she implemented most significant change. As discussed in the same section, Claire also encountered qualitative skeptics when she promoted using success case method.

Skepticism of using qualitative methods is long-standing. There are those who believe that the positivist research paradigm is the only one that illuminates the truth. These individuals argue that quantitative methods are superior to qualitative methods for social science (Gao et al., 2017). However, this view that the positivist research paradigm is the most credible paradigm has been challenged for the last 30 years during what has been termed the paradigm wars. Gao et al. (2017) state that the paradigm wars were fought around whether one could use experimental designs, borne from the natural sciences, within the social sciences and still have valid results. Positivists argued for the primacy of experimental designs to arrive at generalizable truth that could be transported

to other contexts, while constructivists (proponents of qualitative methodologies) argued that truth is subjective and cannot be divorced from context (Gao et al., 2017). Some have stated that the paradigm wars reached a resolution when mixed methods emerged as a way to honor the value of both qualitative and quantitative methodologies (Christie & Fleischer, 2010).

However, others have argued that perhaps the energy from the paradigm wars was merely transferred into what is referred to as the causal wars (Scriven, as cited by Gao et al., 2017). Scriven (2008) defined the term causal wars, stating that "the causal wars are about what is to count as scientifically impeccable evidence of a causal connection, usually in the context of the evaluation of interventions into human affairs" (p. 11). In essence, the causal wars are fought over the same argument that was behind the paradigm wars - an argument over whether different ways of approaching research are equally credible.

Thus, when Lori, Dan, and Emily made arguments that story-based causal methodologies are credible for examining causal questions, they also needed to establish the argument that qualitative methods in general are credible. As discussed in Chapter IV in the cross-case theme Paradigm as Determinant of Perceived Credibility, when Lori spoke about using most significant change and why it was a credible methodology, she'd start with a general conversation about why qualitative research methods are credible. As discussed in that same section, Claire's main argument when addressing qualitative skeptics was to draw on a line of argumentation established by Scriven (2008); namely, that when a person observes factor X leading to factor Y (such as a billiard ball hitting

another, which then goes into the pocket), that person's observation is credible evidence of causality. This is Scriven's key argument about why observational evidence is credible. As he states, "Almost all of the causal claims made in the real world that are beyond reasonable doubt are based on observation or direct inference from observation" (p. 20). Arguing from a different angle, Dan took a more holistic approach when addressing audiences that might be more skeptical of qualitative evaluation; he took care to present the full case for credibility, highlighting both quantitative procedures that were used and qualitative procedures that were used.

Question 2: From the perspective of evaluators, how do contextual factors influence whether story-based causal methodologies are perceived as credible by evaluation stakeholders?

Stakeholders in Learning Contexts May Be More Likely to Find Story-Based Causal Methodologies Credible. When the context of the evaluation is to learn about how well a program is working and not necessarily to prove that a program works, story-based causal methodologies are more likely to be perceived as credible. As discussed in Chapter IV in the cross-case theme Type of Evidence Needed, in Dan's opinion, success case method is best suited to evaluation contexts in which there is a curiosity to learn which factors in an intervention contributed to change. As he explained, success case method can reveal details about what actually occurred over the course of an intervention. This level of detail is not typically provided in other causal evaluation approaches; thus, success case method can provide a more detailed examination of the factors that might contribute to change and how those factors operate within specific contexts. He also

emphasized that because success case method is not designed to compare effectiveness across programs, it is particularly well suited for low-stakes learning contexts wherein participants located at different sites can focus on learning from one another's stories and not be apprehensive about being compared.

As discussed in Chapter IV in the cross-case theme Type of Evidence Needed and in the within-case theme Meeting the Evidence Need, Claire also expressed that success case method is well-suited to learning contexts. She shared that success case method is designed to highlight the accounts of the best performers within an intervention so that others who are average performers or below can learn from their experience. This highlights one way in which story-based causal methodologies have a unique advantage over experimental causal evaluation approaches, as experimental approaches tend to focus on the performance of the average person and whether an intervention made a difference for them. As discussed in those same sections, Claire and Emily said that the goal of using story-based causal methodologies is not to prove that an intervention worked to publish results in an academic setting; rather, the goal is to utilize the findings to learn about the program and hopefully improve it. Thus, with story-based causal methodologies, the goal is not to produce evidence of the standard that would be needed to demonstrate that causality occurred for an academic audience. Rather, the goal is to produce evidence of the standard needed to learn about what's working for those who are seeing success and transfer those lessons learned to improve the program.

These findings resonate with existing literature describing how story-based causal methodologies can effectively produce credible evidence in a learning context. About

success case method, Brinkerhoff (2003) states, “The SCM is a useful approach whenever there is an interest in assessing and learning about how well a program is working” (p. 191). He writes that success case method findings can be applied with a learning mindset. Findings can be used to illuminate aspects of a mature program that are working well in addition to aspects that need improvement. Or, they can be used to demonstrate which elements of a pilot program worked as intended and which elements need further refinement.

Similarly, Dart and Davies (2003) share that most significant change works well in learning environments, as the core purpose of the methodology “is to improve the program by focusing the direction of the work towards explicitly valued directions” (p. 140). By inquiring into the values held by stakeholders in the evaluation, most significant change provides findings that can be implemented to steer the intervention more toward what stakeholders’ value about the program over time. Additionally, most significant change, similar to success case method, is designed to highlight extreme cases of success in order to learn from that success and alter programming to encourage more of those successes (Dart & Davies, 2003).

Stakeholders in Low-Risk Contexts May Be More Likely to Find Story-Based Causal Methodologies Credible. When the context of the evaluation is a low-risk environment, story-based causal methodologies are more likely to be perceived as credible. Randomized controlled trials are a standard when evaluating causality within a medical intervention. Within a medical intervention study, such as a drug trial, the consequences of incorrectly determining that a drug was effective are high. Likewise, the

consequences of incorrectly measuring the impact of side effects are high. In these evaluation contexts, utilizing a randomized controlled trial to examine causality is the correct choice because the standard of evidence must be one in which all plausible alternative causal explanations are addressed. But, as discussed in Chapter IV in Claire's within-case theme Choosing a Methodology That Produces the Evidence Needed, there are many other evaluation contexts in which the risk of incorrectly measuring the results of an intervention is low. She explained that training contexts are low-risk evaluation contexts. If the training is not as effective as the evaluation said it was, typically nobody will be seriously harmed. Thus, in evaluation contexts in which the risk is low, story-based causal methodologies are more likely to be perceived as credible because the risks of an incorrect finding from the evaluation harming anyone are low.

As discussed in that same section in Chapter IV, Claire thinks that a study should have the level of rigor needed for the intended use of the evaluation while also meeting the conditions for feasibility. The resonating argument from her, tying together themes from across her case materials, is that success case method is a highly credible methodology in contexts where: 1) the evidence needed is evidence about how an intervention is working for the best performers; 2) the risk of the findings being "wrong" and harming humans is low; and, 3) the evidence is needed quickly to make decisions to improve a program (within 8-12 weeks).

This finding that different standards of evidence are needed for different risk contexts resonates with existing literature from Grob (2017) regarding adequate evidence. Grob makes the argument to evaluators that "evaluation results will be viewed through

their client's mind-set, not their own" (p. 126). Thus, while many evaluators rely upon methodological procedures to justify the credibility of their choice of methodology, perhaps an equally important consideration is understanding what makes a methodology credible through the eyes of evaluation stakeholders. When the evaluation stakeholders want evidence about what is working for their best performers and when they need that evidence quickly, story-based causal methodologies, and particularly success case method, are a credible choice.

Stakeholders in Multi-Cultural Contexts May Be More Likely to Find Story-Based Causal Methodologies Credible. When the context is multi-cultural, and the goal is to address dynamics of power and privilege through participatory evaluation, story-based causal methodologies are more likely to be perceived as credible. There are dynamics of power and privilege in multi-cultural evaluation, particularly when the evaluator is not from the same cultural or racial background as the participants. As discussed in Chapter IV in Emily's within-case theme Power and Privilege, she noted that her presence as a white foreigner working as an evaluator in West Africa influenced her most significant change evaluation. As the only person analyzing the data, she worried she was missing something. She also had concerns that her presence at data collection events might influence what participants shared.

One way that evaluators address these dynamics is the participatory evaluation approach, which entails partnering with program staff or program participants to design and implement an evaluation study. Emily engaged in participatory evaluation when she implemented the most significant change evaluation. She was an internal evaluator and

she partnered with staff from her NGO who were also members of the same ethnic group as the participants who took part in the study. This was helpful, as these staff members spoke the same language as participants and knew how to follow cultural protocols when collecting the data. Relying on these staff, she limited her influence on the participants by limiting her presence at data collection events and also enhanced the credibility of the study by ensuring that translation was more accurate by utilizing native speakers for translation.

As discussed in Chapter IV in Lori's within-case themes The Methodology Has Multi-cultural Applicability and The Methodology Can Be Implemented in a Participatory Way, Lori also had experience implementing most significant change in multi-cultural settings, and used a participatory approach as well. In situations when she was not of the same cultural background as the program staff or participants, she partnered with program staff to help with recruiting participants, conducting interviews, or translating stories. Partnering with local program staff in this way enhances the credibility of most significant change because translation is more accurate and participants are less apprehensive to participate because they have existing relationships with program staff.

This finding resonates with existing literature addressing how participatory methods add credibility to evaluation work in multi-cultural settings. Kirkhart (2013) argues that "validity may be examined through different lenses and in the context of different applications" (p. 134). Kirkhart understands that there are five major perspectives that form the foundation of arguments for whether an evaluation has some

degree of multicultural validity. These perspectives include methodological, relational, experiential, theoretical, and consequential:

> Methodological justifications of multicultural validity direct attention to the choices of epistemology and method (design, tools and procedures). Relational justifications include relationships among evaluation participants and places. Experiential justifications approach validity from the perspective of the life experiences of program participants or other stakeholders. Invoking theoretical justifications of multicultural validity leads to scrutiny of theoretical foundations. Consequential justifications examine the impacts or sequelae of evaluation to support validity claims. Validity arguments employ multiple justifications, and these justifications interact and build upon (or oppose) one another; they are not independent (p. 135).

Applying Kirkhart's framework, because story-based causal methodologies can easily be paired with a participatory approach in which staff implement pieces of the evaluation, one can make the methodological justification that story-based causal methodologies enhance the multi-cultural validity of an evaluation through methodology. In addition, as most significant change in particular draws upon program staff expertise to help with the valuing of impact stories, one can also make the relational justification that most significant change enhances the multi-cultural validity of an evaluation by drawing on the knowledge that staff have through relationships to participants. Finally, because story-based causal methodologies center participant voice, one can make the experiential

justification that the methodologies increase multi-cultural validity because they value the experience of the individual participant.

Stakeholders Who Value Centering Participant Voice May Find Story-Based Causal Methodologies More Credible. As discussed in Chapter IV within the cross-case theme Elevating Participant Voice, all four evaluators stated that the story-based causal methodologies elevate participant voice in the evaluation process. When the organization commissioning the evaluation values elevating participant voice, then the story-based causal methodologies are more likely to be perceived as credible. As presented in Emily's case story in Chapter IV, she worked for an NGO that had a community-led approach to international development in the communities where they worked in West Africa. The NGO valued centering community members' voices in how development work should be done. This value mapped directly onto the values inherent in most significant change, which is that the evaluation stakeholders' accounts should be the central form of data collected (Dart & Davies, 2003). Additionally, the NGO steered clear of evaluation methodologies that they perceived as "extractive," which would include any evaluations that collect data from participants but do not report the findings back to participants. Most significant change includes a step in which evaluation stakeholders read the impact stories and rank which of the stories is most illustrative of impact to them; thus, they are made aware of the findings and also help participate in the analysis stage (Dart & Davies, 2003). Engaging participants in this way is not extractive; because the methodology is not extractive, it aligned with the NGO's values. Additionally, success case method also has the potential to not be extractive, as one

potential use of success stories is to share them across an organization to encourage uptake of successful behaviors (Brinkerhoff, 2003).

Contributions of the Study

Prior to this study, there was no comprehensive literature review of justifications that evaluators make when using story-based methodologies in causal inquiry. Thus, the first major contribution of this study was to provide a comprehensive literature review in Chapter III that detailed existing justifications. For the most part, this study substantiated themes within the existing literature around how evaluators justify using story-based causal methodologies when examining causality. In particular, findings from the study resonated with existing themes that story-based causal methodologies are credible because they unearth unexpected outcomes, elevate participant voice, and include triangulation procedures. Thus, another contribution of this study was to demonstrate that the experiences of four evaluators who have conducted story-based causal methodologies resonate with themes in the existing literature. Additionally, the findings served to highlight and specify particular contexts in which stakeholders might be more likely to perceive story-based causal methodologies as credible. This level of specification detailed in Chapter IV and Chapter V did not previously exist in the literature and is a unique contribution.

The purpose of this study was also to provide a practitioner's guide for evaluators to help them navigate conversations and contexts in which they need to justify using story-based methodologies for causal examination. The practitioner's guide that follows (Table 6.) was developed to integrate the justifications found in the literature with the

justifications voiced by the evaluators in this study concerning why story-based causal methodologies are credible for use in causal examination. Having these justifications easily accessible is helpful for evaluators when responding to challenges to credibility by evaluation stakeholders. This guide also details evaluation contexts that might be more amenable to accepting story-based causal methodologies as credible, which can be helpful for evaluators when assessing whether these methodologies are the right choice. The practitioner's guide is another contribution that this study has made to the literature.

Table 6.
Practitioner's Guide

Arguments that evaluators can make to establish credibility of story-based causal methodologies	When implementing an evaluation examining causality, there are different causal questions that evaluators can ask. Story-based causal methodologies are best suited to answer a specific causal question, which is: What are participants' perspectives of how and why an intervention changed their lives?
	When the participant shares their story in a story-based causal methodology, they often reveal unexpected outcomes from the intervention. They also may reveal unexpected mechanisms for how a casual factor, interacting with factors within the context, brought about an effect. This facet of story-based causal methodologies reduces potential evaluator bias and also increases multi-cultural validity.
	More often than not, participants can be trusted to provide their accounts truthfully in story-based causal methodologies. When stories are collected through an external evaluator, that enhances the likelihood that participants accurately present their stories. Additionally, qualitative probing techniques can be employed to ensure participants are providing truthful accounts.
	Story-based causal methodologies typically include procedures to triangulate the data. In addition to collecting participant stories, evaluators can also collect peer interviews, conduct site visits, review administrative records or administer surveys to provide further evidence to support that an impact or success occurred.

Contexts in which stakeholders are more likely to perceive story-based causal methodologies as credible	Story-based causal methodologies are useful in scenarios when the goal is to learn about how the program worked for individuals. Story-based methodologies may be less helpful in scenarios when the goal is to prove that an intervention worked for the average participant.
	Story-based causal methodologies are a strong choice when the context is a low-risk context. If there might be serious harm resulting from incorrect findings, then story-based causal methodologies are not the correct choice.
	Story-based causal methodologies are well-suited to multi-cultural contexts because they can be easily implemented using a participatory approach in which evaluators partner with program staff to design and implement the study.
	Story-based causal methodologies are useful in scenarios when the organization requesting the evaluation wishes to elevate and center participant voice as a primary source of data.

Another contribution of this study was to demonstrate a novel use of the case study methodology. The case study methodology has never been applied before to answer the research questions from this study relating to examining the credibility of using story-based methodologies for causal examination. This study demonstrated that one can effectively use a narrative methodology (case study) to examine methodological elements of other narrative methodologies (success case method and most significant change).

Limitations

This study has some important limitations. The first limitation was that there were no opportunities to observe participants conducting story-based causal methodologies. The hope was to include observational data as an additional source of data to integrate

into the case study in addition to the interview, document, and artifact data. Unfortunately, the COVID-19 pandemic contributed to the inability to collect observational data in person; the data collection period coincided directly with the pandemic, and in-person observation was not considered safe. Additionally, the pandemic may have limited virtual opportunities for observation, as virtual opportunities to observe were also less common as evaluation work was more generally interrupted due to the virus. The lack of observational data limited my ability to provide rich descriptions of specific evaluation contexts in which story-based causal methodologies were being implemented.

The second limitation of the study was that the only perspective gathered through interviews was that of each evaluator who participated in the study. The multiple case study design might have benefited from gathering stakeholder interviews to more directly understand their perceptions of credibility. Unfortunately, the complexity of conducting a multiple case study of that magnitude exceeded the time and resources available for this study. I attempted to account for this limitation by including other sources of data in addition to the evaluator interviews, including documents and artifacts that described the evaluations that served as the cases. The third limitation of the study was that there was a lack of racial diversity among the participants as all four participants were white. As such, the study findings do not include the perspectives of non-white evaluators. This is an important limitation, as it cannot be assumed that the experience of white evaluators is the same as the experience of non-white evaluators when justifying a methodology choice.

Future Research

In the first chapter of this study, I identified Gates & Dyson's (2017) work as influential in determining the study's purpose. Gates and Dyson invite researchers to study how evaluators defend their choice of methodology when evaluating causality. This study attempts to capture the arguments that evaluators make when using story-based causal methodologies for causal inquiry. However, there is still work to be done exploring the justifications that evaluators make when approaching causal inquiry from other less well-known viewpoints, including the causal package, generative, and complex systems causal viewpoints.

Additionally, further research to capture how evaluators construct arguments to defend using story-based causal methodologies for causal inquiry could be beneficial. For example, it might be worth conducting a phenomenological study using a larger sample size but asking a similar research question to the first one utilized in this study, which was "What arguments are made by evaluators to justify the credibility of these story-based causal methodologies to evaluation stakeholders?" Using a phenomenological approach might reveal a more concentrated explanation of evaluators' justifications, highlighting the deeper essence of the arguments. There might also be the potential for a follow-up quantitative survey to determine whether there is agreement or disagreement with the justifications presented in this study among a wider group of evaluators. Furthermore, another area for future research would be to explore how braiding different approaches to causal inquiry increases the overall validity of findings.

Finally, another opportunity for future research would be to explore how power and privilege intersect with the determination of credibility. LaFrance et al. (2015) challenge us to ask who is served by existing definitions of credibility in evaluation. Who is left out in determining which evaluation methodologies are credible? Who decides what is legitimate? How are current definitions of what is credible in evaluation being used to hold up existing power structures within the evaluation and research world? These questions could be explored further.

Summary and Conclusion

The central problem addressed by this study was the challenge that evaluators encounter when attempting to establish evidence of causality for small social programs. While some small social programs may be able to utilize experimental or quasi-experimental designs, many face challenges related to high cost, small sample sizes, or a concern that withholding services is unethical. Evaluators may also determine that experimental or quasi-experimental designs are not the correct choice for small social programs because of the complex nature of those programs, or because the evidence need is more targeted toward knowing the "how" or "why" a program worked instead of "whether" it worked. However, alternative evaluation designs to the experimental design (such as story-based causal methodologies) exist for small social programs. The challenge for evaluators is that the evidence supporting their use in causal examination is not well developed. This central problem was presented in the first chapter.

The second chapter of this study synthesized existing literature describing evaluators' justifications for using story-based causal evaluation methodologies to

conduct causal evaluation. This literature review introduced the Gates & Dyson (2017) framework that describes multiple ways of thinking about causality that are equally valid. The literature review also explored why narrative inquiry is a credible research lens, including the key argument that there is value in more deeply understanding an individual's perspective, motivation, and context. Additional key themes emerged from the literature review, including: 1) story-based causal methodologies reveal unexpected outcomes; 2) story-based causal methodologies emphasize participant voice; 3) story-based causal methodologies have the capacity to be transformational; 4) story-based causal methodologies focus on for whom an intervention was successful and which parts of the intervention made it successful; and 5) story-based causal methodologies are designed to include procedures that enhance credibility by corroborating data sources.

The third chapter of this study provided an overview of the research methodology. The research methodology included a multiple case study focusing on four cases of evaluators who implemented a story-based causal methodology. The design included interviews with each evaluator, as well as collection of documents and artifacts from each evaluator. Analysis included a cross-case synthesis approach, and within and cross-case themes were derived. The fourth chapter presented the following cross-case themes: 1) type of evidence needed; 2) paradigm as determinant of perceived credibility; 3) truthfulness of accounts; 4) triangulation among data sources; 5) elevating participant voice; 6) unexpected outcomes or mechanisms; and, 7) credibility from the evaluator or organization.

The fifth chapter of this study discussed the connection between the themes derived from the multiple case study and themes within existing literature regarding story-based causal methodologies and credibility in causal inquiry. Combining the themes derived from the study with existing literature yielded a practitioner's guide that evaluators can use when seeking to justify their choice of using story-based causal methodologies for causal evaluation to multiple stakeholders. This practitioner's guide directly answers Gates & Dyson's (2017) call to the evaluation community to more deeply understand "how evaluators can justify the causal approach taken to multiple audiences" (p. 43). This study provided evaluators with a line of argumentation to defend using story-based causal methodologies with small social programs that have few options of causal inquiry designs that fit their circumstance. This study also attempted to raise awareness about the dynamic nature of credibility and how beliefs on what is credible can shift depending on stakeholders and context. It is critical that researchers continue to unpack the implications of discussions of credibility on causal inquiry within social programs.

BIBLIOGRAPHY

Abell, P. (2004). Narrative explanation: An alternative to variable-centered explanation? *Annual Review of Sociology, 30*(1), 287-310.

Anfara, V. A., Brown, K. M., & Mangione, T. L. (2002). Qualitative analysis on stage: Making the research process more public. *Educational Researcher, 31*(7), 28-38.

Azzam, T., & Christie, C. (2007). Using public databases to study relative program impact. *The Canadian Journal of Program Evaluation, 22*(2), 57-68.

Boruch, R. (2007). Encouraging the flight of error: Ethical standards, evidence standards, and randomized trials. *New Directions for Evaluation, 2007*(113), 55-73.

Brinkerhoff, R. O. (2003). *The success case method: Find out quickly what's working and what's not* (1st ed.). Berrett-Koehler.

Corporation for National Community Service. (n.d.) *Budgeting for evaluation.* https://www.nationalservice.gov/sites/default/files/resource/Budgeting%20for%20Evaluation_090914st10.17.pdf

Chatterji, M. (2007). Grades of evidence: Variability in quality of findings in effectiveness studies of complex field interventions. *American Journal of Evaluation, 28*(3), 239-255.

Christie, C. A., & Fleischer, D. N. (2010). Insight into evaluation practice: A content analysis of designs and methods used in evaluation studies published in North American evaluation-focused journals. *American Journal of Evaluation, 31*(3), 326-346.

Choy, S., & Lidstone, J. (2013). Evaluating leadership development using the most significant change technique. *Studies in Educational Evaluation, 39*(4), 218-224.

Creswell, J. (2013). *Qualitative inquiry and research design: Choosing among five approaches*. SAGE Publications.

Creswell, J. (2016). *30 essential skills for the qualitative researcher*. SAGE Publications.

Creswell, J. and Miller, D. (2000). Determining Validity in Qualitative Inquiry. *Theory Into Practice, 39*(3), 124-130.

Creswell, J. and Plano Clark, V. (2017). *Designing and conducting mixed methods research*. (3rd ed.). SAGE Publications.

Dart, J., & Davies, R. (2003). A dialogical, story-based evaluation tool: The most significant change technique. *American Journal of Evaluation, 24*(2), 137-155.

Dinh, K., Worth, H., & Haire, B. (2019). Buddhist evaluation: Applying a buddhist world view to the most significant change technique. *Evaluation, 25*(4), 477-495.

Fowler, F.J. (2014). Survey Research Methods (5th ed.). SAGE Publications.

Gao, X., Shen, J., & Krenn, H. Y. (2017). Using WWC sanctioned rigorous methods to develop comparison groups for evaluation. *Evaluation and Program Planning, 65*, 148-155.

Gates, E., & Dyson, L. (2017). Implications of the changing conversation about causality for evaluators. *American Journal of Evaluation, 38*(1), 29-46.

Glass, G., & Hopkins, K. (1996). *Statistical methods in education and psychology*. (3rd ed.). Pearson.

Gliner, J., Morgan, G., & Leech, N. (2016). *Research methods in applied settings: An*

integrated approach to design and analysis. (3rd ed.). Routledge.

Grob, G.F. (2017). Evaluation practice: Proof, truth, client relationships, and professional growth. *American Journal of Evaluation, 93*(1), 123-132.

Hawk, M. (2015). The girlfriends project: Evaluating a promising community-based intervention from a bottom-up perspective. *American Journal of Evaluation, 36*(2), 179-190.

Hawkins, A. J. (2016). Realist evaluation and randomised controlled trials for testing program theory in complex social systems. *Evaluation, 22*(3), 270-285.

Hood, S., Hopson, K., Kirkhart, K. (2015). Culturally responsive evaluation: Theory, practice and future implications. In H.P. Hatry, K. E. Newcomer, and J.S. Wholey (Eds.), *Handbook of practical program evaluation* (4th ed., pp. 281-318). Jossey-Bass.

Johnson, R. (2017). Dialectical pluralism: A metaparadigm whose time has come. *Journal of Mixed Methods Research, 11*(2), 156-173.

Kirkhart, K.E. (2013). Advancing considerations of culture and validity: Honoring the key evaluation checklist. In S.I. Donaldson (Ed.), *The future of evaluation in society: A tribute to Michael Scriven* (pp. 129-159). Information Age Publishing.

LaFrance, J., Kirkhart, K.E., and Nichols, R. (2015). Cultural views of validity. In S. Hood, R. Hopson, & H. Frierson (Eds.), *Continuing the journey to reposition culture and cultural context in evaluation theory and practice* (pp. 49-71). Information Age Publishing.

Lehmann, E. R. (2015). What if 'what works' doesn't? *Evaluation, 21*(2), 167-172.

Limato, R., Ahmed, R., Magdalena, A., Nasir, S., & Kotvojs, F. (2017). Use of most significant change (MSC) technique to evaluate health promotion training of maternal community health workers in Cianjur District, Indonesia. *Evaluation and Program Planning, 66,* 102-110.

Linden, A., Trochim, W. M. K., & Adams, J. L. (2006). Evaluating program effectiveness using the regression point displacement design. *Evaluation & the Health Professions, 29*(4), 407-423.

Mason, W. A., Cogua-Lopez, J., Fleming, C. B., & Scheier, L. M. (2018). Challenges facing evidence-based prevention: Incorporating an abductive theory of method. *Evaluation & the Health Professions, 41*(2), 155-182.

Marchal, B., van Belle, S., van Olmen, J., Hoerée, T., & Kegels, G. (2012). Is realist evaluation keeping its promise? A review of published empirical studies in the field of health systems research. *Evaluation, 18*(2), 192-212.

Medina, L., Acosta-Pérez, E., Velez, C., Martínez, G., Rivera, M., Sardiñas, L., & Pattatucci, A. (2015). Training and capacity building evaluation: Maximizing resources and results with success case method. *Evaluation and Program Planning, 52,* 126-132.

Mertens, D., & Wilson, A. (2012). *Program evaluation theory and practice: A comprehensive guide.* Guilford Press.

Patton, M. (2015). *Qualitative research and evaluation methods* (4th ed.). SAGE Publications.

Pawson, R. & Tilley, N. (1997). *Realistic evaluation.* SAGE Publications.

Peshkin, A. (1988). In search of subjectivity - one's own. *Educational Researcher, 17*(7), 17-21.

Rossi, P.H., Lipsey, M.W., & Henry, G.T. (2019). *Evaluation: A systematic approach* (8th ed.). SAGE Publications.

Scriven, M. (2008). A summative evaluation of RCT methodology: An alternative approach to causal research. *Journal of MultiDisciplinary Evaluation, 5*(9), 11-24.

Shadish, W., Cook, T., & Campbell, D. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Houghton Mifflin Company.

Stame, N. (2014). Positive thinking approaches to evaluation and program perspectives. *The Canadian Journal of Program Evaluation, 29*(2) 67-86.

Stake, R.E. (1995). *The art of case study research.* SAGE Publications.

Teddlie, C. & Tashakkori, A. (2003). Major issues and controversies in the use of mixed methods in the social and behavioral sciences. In C. Teddlie & A. Tashakkori (Eds.), *Handbook of mixed methods in social & behavioral research* (pp. 3-50). SAGE Publications.

Tracy, S.J. (2010). Qualitative quality: Eight “big-tent” criteria for excellent qualitative research. *Qualitative Inquiry, 16*(10), 837-851.

van Wessel, M. (2018). Narrative assessment: A new approach to evaluation of advocacy for development. *Evaluation, 24*(4), 400-418.

Vingilis, E., & Pederson, L. (2001). Using the right tools to answer the right questions:

The importance of evaluative research techniques for health services evaluation research in the 21st century. *The Canadian Journal of Program Evaluation, 16*(2), 1-26.

Westhorp, G. (2008). *Development of realist evaluation models and methods for use in small-scale community based settings* (Publication No. 10290761) [Doctoral dissertation, Nottingham Trent University]. ProQuest Dissertations and Theses Global.

Wilson, S. (2008). *Research is ceremony: Indigenous research*. Fernwood Publishing.

Yarbrough, D.B., Shula, L.M., Hopson, R.K., Caruthers, F.A. (2010). *The Program Evaluation Standards: A guide for evaluators and evaluation users* (3rd ed.). Corwin Press.

Yin, R. (2009). *Case study research: Design and methods* (4th ed.). SAGE Publications.

Yin, R. (2018). Case study research and applications (6th ed.). SAGE Publications.

APPENDICES

Appendix A: Interview 1 Protocol - Listen to the Story

Researcher note: The participant will be prompted ahead of time to remember a story-based causal evaluation that they completed. Upon beginning the interview, the researcher will state: Please think back to a time when you conducted a (success case method or most significant change) evaluation. When you have one in mind and are ready to begin, please let me know.

This first section will focus on some of the high-level details about the evaluation.

1. To begin, please give me a high-level overview of what the evaluation was.
    a. What were you trying to demonstrate or prove?
2. Where was the evaluation located, and when was it conducted?
3. What methodology did you use in the evaluation (success case method or most significant change)?
    a. What were your data collection, analysis, and presentation procedures?
4. Why did you select (success case method/most significant chance) as your method of choice?
5. What were your findings?

The next section will focus on some details about you and your organization at the time of the evaluation.

6. What organization did you work for at the time of this evaluation?
    a. Was it a research institute, an evaluation firm, etc.?
7. What was your role in the organization?

8. What was your role in the evaluation project (the project lead, research associate, etc.)?
9. At this time in your career, did you ascribe to any theories or approaches about how evaluation should be done?

The next series of questions will establish more details about the stakeholders of the evaluation.

10. Often in evaluation we talk about evaluation stakeholders – those who have a stake in the information produced from the evaluation. I'm going to ask you about each level of stakeholder and will ask you to help me understand their stake in the evaluation:
    a. Donor or funder. Who was paying for the evaluation?
        i. What was their stake in the evaluation?
    b. Program implementer. What was the organization that implemented the services or program being evaluated? How would you characterize the organization in terms of size?
        i. What was their stake in the evaluation?
    c. Program participants: Who were you collecting data from?
        i. What was their stake in the evaluation?
    d. Program beneficiaries. Who was the program intending to serve?
        i. What was their stake in the evaluation?

e. Peer evaluators/research community. Were peer evaluators or the research community involved in the evaluation in any way? Did you share knowledge about this evaluation with the evaluation/research community?
    i. What was their stake in the evaluation?
f. Political stakeholders. Were there any members of the political or government community that were interested in your findings?
    i. What was their stake in the evaluation?

The next series of questions will establish more details about the context of the evaluation. When I say the word "context" what I'm referring to are contextual or environmental factors such as: physical location of the program, organization that you worked for, culture and diversity, values and beliefs, history and tradition, power and privilege, and political context.

11. Now that you have all of those factors in your mind, can you describe how any of these may have played a role in the evaluation?

The next series of questions will focus on major moments of the evaluation story.

12. As you moved through the evaluation, did you encounter any conflicts or challenges?
    a. How did you resolve those conflicts or challenges?
13. Did the client utilize the findings? If so, how?

Appendix B: Interview 2 Protocol - Clarify the Justifications

Researcher note: The focus of this interview will be on the arguments you might make to justify this choice of methodology for use in causal evaluation. The focus of this interview will also be on which factors about the evaluation context may influence perceived credibility of these methods among evaluation stakeholders.

Now we will think back to the evaluation story that you shared in the first interview. Let's take a moment to remember that story. When you are ready, please let me know.

1. Were there moments during the evaluation when you had to justify that (success case method or most significant change) was a credible methodology for demonstrating that an intervention made a difference?
    a. What kinds of justifications or arguments did you make?
2. Did your justifications change depending on who the evaluation stakeholder was? As a reminder, evaluation stakeholders can include:
    a. Donor or funder
    b. Program implementers
    c. Program participants
    d. Program beneficiaries
    e. Peer evaluators/research community
    f. Political stakeholders

Next, I'm going to ask a few questions about components of the (success case method or most significant change) methodology that may relate to credibility. You will see in these

questions that I am aiming to ground your responses in examples from the evaluation story.

2. As you know, the (success case method or most significant change) methodology gathers individual's accounts of how an intervention impacted their lives. Based on your experience implementing this methodology for this evaluation, did you trust the individual to provide their own account of how an intervention changed their lives? If so, why?
3. Did this methodology uncover evidence of outcomes that the program staff or the evaluator did not know had occurred? If so, how?
4. *(If the participant used the methodology in multi-cultural settings)* Based on your experience during this evaluation, was the methodology a credible choice for use in a multi-cultural setting? If so, why?
5. Did the methodology elevate participant voice? If so, how?
6. Did the methodology address issues of power and privilege in evaluation? If so, how?
7. Did you generalize the findings that were produced from this methodology? Why or why not?
8. Did you corroborate the evidence produced from the stories with other forms of evidence? If so, why?

Now let's shift to speak about the context of your evaluation story and perceptions of credibility. Similar to the previous interview, when I say the word "context" what I'm referring to are contextual or environmental factors such as: physical location of the

program, organization that you worked for, culture and diversity, values and beliefs, history and tradition, power and privilege, and political context.

9. Now that you have all of those factors in your mind, can you speak to how any of these may have had a role in whether or not stakeholders perceived your evaluation as credible?
10. Are there any other contextual or environmental factors that may have influenced whether stakeholders perceived your evaluation as credible?

Appendix C: Document and Artifact Review Protocol

Appendix C Document and Artifact Review Protocol	
Protocol Question	Response
Write down descriptive details of this document/artifact: object title, high level overview of content, purpose of object, audience of object.	
How does this object describe key elements of the story of the evaluation (setting, characters, plot, chronology)?	
How does this object describe details of the evaluation context or setting?	
How does this object justify the use of story-based causal approaches?	

Appendix D: Alignment of Data Collection Protocols and Research Questions

Appendix D Alignment of Data Collection Protocols and Research Questions		
Protocol	Question	Research Question Aligned With
Interview 1	To begin, please give me a high-level overview of what the evaluation was.	RQ2
Interview 1	What were you trying to demonstrate or prove?	RQ2
Interview 1	Where was the evaluation located, and when was it conducted?	RQ2
Interview 1	What methodology did you use in the evaluation (success case method or most significant change)?	RQ2
Interview 1	What were your data collection, analysis, and presentation procedures?	RQ2
Interview 1	What were your findings?	RQ2
Interview 1	What organization did you work for at the time of this evaluation?	RQ2
Interview 1	Was it a research institute, an evaluation firm, etc.?	RQ2
Interview 1	What was your role in the organization?	RQ2
Interview 1	What was your role in the evaluation project (the project lead, research associate, etc.)?	RQ2
Interview 1	At this time in your career, did you ascribe to any theories or approaches about how evaluation should be done?	RQ2, RQ1
Interview 1	Often in evaluation we talk about evaluation stakeholders – those who have a stake in the information produced from the evaluation. I'm going to ask you about each level of stakeholder and will	RQ2

	ask you to help me understand their stake in the evaluation. • Donor or funder • Program implementers • Program participants • Program beneficiaries • Peer evaluators/research community • Political stakeholders	
Interview 2	Were there moments during the evaluation when you had to justify that (success case method or most significant change) was a credible methodology for demonstrating that an intervention made a difference?	RQ1
Interview 2	What kinds of justifications or arguments did you make?	RQ1
Interview 2	Did your justifications change depending on who the evaluation stakeholder was? As a reminder, evaluation stakeholders can include: • Donor or funder • Program implementers • Program participants • Program beneficiaries • Peer evaluators/research community • Political stakeholders	RQ1
Interview 2	As you know, the (success case method or most significant change) methodology gathers individual's accounts of how an intervention impacted their lives. Based on your experience implementing this methodology for this evaluation, did you trust the individual to provide their own account of how an intervention changed their lives? If so, why?	RQ1
Interview 2	Did this methodology uncover evidence of outcomes that the program staff or the evaluator did not know had occurred? If so, how?	RQ1
Interview 2	*(If the participant used the methodology in multi-cultural settings)* Based on your experience during	RQ1

	this evaluation, was the methodology a credible choice for use in a multi-cultural setting? If so, why?	
Interview 2	Did the methodology elevate participant voice? If so, how?	RQ1
Interview 2	Did the methodology address issues of power and privilege in evaluation? If so, how?	RQ1
Interview 2	Did you generalize the findings that were produced from this methodology? Why or why not?	RQ1
Interview 2	Did you corroborate the evidence produced from the stories with other forms of evidence? If so, why?	RQ1
Interview 2	*Now let's shift to speak about the context of your evaluation story and perceptions of credibility. Similar to the previous interview, when I say the word "context" what I'm referring to are contextual or environmental factors such as: physical location of the program, organization that you worked for, culture and diversity, values and beliefs, history and tradition, power and privilege, and political context.* Now that you have all of those factors in your mind, can you speak to how any of these may have had a role in whether or not stakeholders perceived your evaluation as credible?	RQ2
Interview 2	Are there any other contextual or environmental factors that may have influenced whether stakeholders perceived your evaluation as credible?	RQ2
Documents or Artifacts	How does this object describe key elements of the story of the evaluation (setting, characters, plot, chronology)?	RQ2
Documents or Artifacts	How does this object describe details of the evaluation context or setting?	RQ2
Documents or Artifacts	How does this object justify the use of story-based causal approaches?	RQ1

Appendix E: A Priori Codebook

Appendix E A Priori Code Codebook	
Code	Explanation of Code
Corroboration	Any text referring to how these methodologies require rigorous corroboration with other forms of data in addition to the story, increasing their credibility for use in causal research.
Narrative/credible	Any text referring to narrative ways of thinking, or story-based ways of thinking, as being credible for causal research.
Participant voice/cultural validity	Any text referring to how these methodologies raise different cultural perspectives and place value on what outcomes participants find meaningful from their own cultural perspective.
Successful parts v. whole	Any text referring to how these methodologies have the ability to focus on what works for whom (as opposed to what works for the average participant).
Transformation	Any text referring to how these methodologies challenge existing power and privilege structures, and thus reveal alternative explanations for causality that might not have been revealed with other causal methodologies.
Unexpected outcomes	Any text referring to how the story-based methodologies can reveal to the evaluator different outcomes than what could have been revealed with other causal methodologies.

Appendix F: Final Codebook

Appendix F Final Codebook	
Code	Explanation of Code
Capacity building with the method	Any text referring to the evaluator teaching their client to do success case method or most significant change themselves.
Context & credibility - culture and diversity	Any text referring to how the context of culture and diversity in the evaluation influenced perceptions of credibility.
Context & credibility - external evaluator	Any text referring to how the context of being an external evaluator in the evaluation influenced perceptions of credibility.
Context & credibility - history and tradition	Any text referring to how the context of history and tradition in the evaluation influenced perceptions of credibility.
Context & credibility - organization	Any text referring to how the organization's reputation influenced the perceived credibility of the methodology among stakeholders.
Context & credibility - power and privilege	Any text referring to how the context of power and privilege in the evaluation influenced perceptions of credibility.
Context & credibility - relatable/accessible	Any text referring to how credibility is increased when data is collected or presented through stories, as stories are relatable.
Context & credibility - evaluator background	Any text referring to how the context of evaluator background in the evaluation influenced perceptions of credibility.
Contribution vs. attribution	Any text referring to the difference between looking at contribution and attribution in causality.
Corroboration	Any text referring to how these methodologies require rigorous corroboration with other forms of

	data in addition to the story, increasing their credibility for use in causal research.
Credibility - strengthened with other methods	Any text referring to how combining most significant change or success case method with other methodologies or methods increased the credibility of the approach. Any text referring to instances in which validity procedures outside of corroboration were utilized to enhance credibility.
Cultural validity	Any text referring to how these methodologies raise different cultural perspectives and place value on what outcomes participants find meaningful from their own cultural perspective.
Definition of impact	Any text referring to how the evaluator defines what impact means.
Narrative/credible	Any text referring to narrative ways of thinking, or story-based ways of thinking, as being credible for causal research.
No challenge to credibility	Any text referring to the evaluator not having a problem with stakeholders challenging the credibility of success case method or most significant change.
Participant voice	Any text referring to how the method elevates participant voice.
Purpose of evaluation	Any text referring to the reason why evaluators do evaluation – the underlying purpose.
Right evidence for context	Any text referring to evidence needing to fit the context it is produced for; this relates to what the methodology can and can't do.
Stakeholder perception of credibility – non-technical stakeholders	Any text referring to what credibility looks like to non-technical stakeholders, such as participants or beneficiaries or staff without experience in evaluation.

Stakeholder perception of credibility - paradigm	Any text referring to how perception of credibility shifts depending on the stakeholder's paradigm.
Stakeholder perception of credibility - use	Any text referring to stakeholders' willingness to participate depending on use of the findings.
Successful parts v. whole	Any text referring to how these methodologies have the ability to focus on what works for whom (as opposed to what works for the average participant).
Transformation	Any text referring to how these methodologies challenge existing power and privilege structures, and thus reveal alternative explanations for causality that might not have been revealed with other causal methodologies.
Unexpected outcomes	Any text referring to how the story-based methodologies can reveal to the evaluator different outcomes than what could have been revealed with other causal methodologies.

Appendix G: Code Frequency Table

Code	Case 1	Case 2	Case 3	Case 4	Total times code was applied
Capacity building with the method	0	1	0	4	5
Context & credibility - culture and diversity	0	1	1	0	2
Context & credibility - external evaluator	0	1	0	0	1
Context & credibility - history and tradition	0	1	4	1	6
Context & credibility - organization	1	1	4	2	8
Context & credibility - power and privilege	1	2	2	0	5
Context & credibility - relatable/accessible	0	0	1	1	2
Context & credibility - evaluator background	0	1	0	0	1
Contribution vs. attribution	1	0	1	2	4
Corroboration	2	1	3	1	7
Credibility - strengthened with other methods	4	0	3	0	7
Cultural validity	1	1	1	2	5
Definition of impact	0	2	0	0	2
Narrative/credible	2	3	3	1	9
No challenge to credibility	1	1	1	2	5
Participant voice	1	2	4	2	9
Purpose of evaluation	1	6	0	0	7
Right evidence for context	4	12	3	2	21
Stakeholder perception of credibility – non-technical stakeholders	0	0	2	2	4
Stakeholder perception of credibility - paradigm	3	3	2	1	9

Code	Case 1	Case 2	Case 3	Case 4	Total times code was applied
Stakeholder perception of credibility - use	1	3	2	2	8
Successful parts v. whole	0	2	0	0	2
Transformation	0	1	3	1	5
Unexpected outcomes	2	1	2	3	8

Appendix H: Data Collection, Analysis and Writing Process