2022

Application of an Organizational Evaluation Capacity Assessment in a Multinational NGO: A Case Study to Support Applied Practice

Ryan James Smyth
University of Denver

Application of an Organizational Evaluation Capacity Assessment in a Multinational NGO: A Case Study to Support Applied Practice

Abstract

As evaluation capacity building (ECB) has rapidly emerged as a practice in human service organizations and as a field of academic inquiry, attention has focused on methods of evaluation capacity building while assessment of organizational evaluation capacity (EC) has lagged behind. To examine the practice of organizational evaluation capacity assessment, this dissertation presents two separate but related studies. In sub-study 1, I present a qualitative evidence synthesis of the research theorizing organizational evaluation capacity models. In sub-study 2, I support the implementation of one of the tools from the evidence-synthesis at a multinational human service organization. I use a concurrent mixed methods instrumental case study to describe how the sample organization implements an evaluation capacity assessment survey, interprets the results, and determines the next course of action in their evaluation capacity building initiatives. In the conclusion, I discuss the two sub-studies and use the lessons and observations from the case study to theorize an application framework for organizational evaluation capacity assessments.

Document Type

Dissertation

Degree Name

Ph.D.

Department

Quantitative Research Methods

First Advisor

Robyn Thomas Pitts

Second Advisor

P. Bruce Uhrmacher

Third Advisor

Kathy Green

Keywords

Evaluation capacity building, Human service organizations, Non-profit, Organizational evaluation capacity assessment, Organizational measurement, Program evaluation

Subject Categories

Quantitative, Qualitative, Comparative, and Historical Methodologies | Statistical Methodology

Application of an Organizational Evaluation Capacity Assessment in a Multinational NGO: A Case Study to Support Applied Practice

_____

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

Ryan James Smyth

August 2022

Advisor: Robyn Thomas Pitts

Author: Ryan James Smyth
Title: Application of an Organizational Evaluation Capacity Assessment in a Multinational NGO: A Case Study to Support Applied Practice
Advisor: Robyn Thomas Pitts
Degree Date: August 2022

Abstract

As evaluation capacity building (ECB) has rapidly emerged as a practice in human service organizations and as a field of academic inquiry, attention has focused on methods of evaluation capacity building while assessment of organizational evaluation capacity (EC) has lagged behind. To examine the practice of organizational evaluation capacity assessment, this dissertation presents two separate but related studies. In sub-study 1, I present a qualitative evidence synthesis of the research theorizing organizational evaluation capacity models. In sub-study 2, I support the implementation of one of the tools from the evidence-synthesis at a multinational human service organization. I use a concurrent mixed methods instrumental case study to describe how the sample organization implements an evaluation capacity assessment survey, interprets the results, and determines the next course of action in their evaluation capacity building initiatives. In the conclusion, I discuss the two sub-studies and use the lessons and observations from the case study to theorize an application framework for organizational evaluation capacity assessments.

Acknowledgements

I am profoundly grateful to my committee members: Dr. Kathy Green, Dr. Duan Zhang, Dr. Bruce Uhrmacher, and Dr. Robyn Thomas Pitts. All of you are excellent professors who taught me skills I rely on not only in this mixed-methods research but my career; each of you have made me a better evaluation practitioner and professional.

I am indebted to my incredible community who have continually held me up during my time in the program. I am so grateful I had Brandon Dent and Kelly Smyth-Dent with me in Denver, who love me and my family as well as anyone possibly could. I doubt I would have ever had academic goals without the inspiration of Dr. John Penniman, who's friendship has shaped so many aspects of my life and worldview. And to Randy Skeen for the life-long bond that has taught me to always have big dreams, continually inspires me to grow, and remains a bottomless well of love and support.

All the opportunities I have had in my life I owe to my parents. They always encouraged educational pursuits, even when I wasn't interested in them, and have unconditionally supported my interests, goals, and decisions without hesitation. I aspire to be as unselfish and giving to my children as they have been to me.

Lastly, I dedicate my dissertation to my wife, Jessica Miller. There is no one I admire more; her abilities as a wife, mother, professional, and friend have no equal. She has given me the three greatest gifts of my life in our three boys, Jameson, Cody, and Maddox, and I feel endlessly proud to be her husband. She is the reason I can believe in myself, as her devotion to our partnership provides the majority of strength and resiliency I can summon. Like every other accomplishment in my life, the completion of my dissertation would not have been possible without her.

Table of Contents

List of Tables

List of Figures

Chapter One: Introduction

Over the last few decades, human service organizations have faced steadily growing demand from numerous audiences and stakeholders to increase their use of evaluation to demonstrate program efficacy and accountability (Lynch-Cerullo & Cooney, 2011; Cheng & King, 2017). In addition to the external push to meet funder accountability reporting, internal incentives to increase program evaluation efforts exist like the desire to improve organizational learning, respond to new funding opportunities, support strategic planning, and increase beneficiary impact (Preskill & Boyle, 2008; Labin et al., 2012; Despard, 2016). The resulting factors have normalized program evaluation as a necessary function in the human service landscape, with donors asking for more data and reporting, and organizations placing a higher priority on organizational learning cultures (Carman, 2011; Mitchell & Berlan, 2018). However, this cultural shift created the need to acquire new skills and practices, as human service organizations' accountability metrics and analysis previously focused on financial health, program expense ratios, and funding controls (Mitchell & Berlan, 2018).

Despite the practice of program evaluation now normalized in human service organizations' practice, organizations with limited resources struggle with evaluation and reporting to their funders (Carman, 2011). The lack of resources and capacity, both from financial and human capital, results in organizations not acquiring the skills and systems

to properly embed program evaluation best practices in their organization (Preskill & Boyle, 2008; Labin et al., 2012; Cheng and King, 2017). The increased demand along with the struggle to scale up program evaluation practice has led to the rapid emergence of the field of evaluation capacity building (Cousins, Goh, Clark, & Lee, 2004; Preskill & Boyle, 2008). For this dissertation, I adopt a definition of evaluation capacity building found in Labin et al. (2012): "an intentional process aimed at increasing the individual motivation, knowledge, skills, and to enhance an organization's ability to conduct and use evaluation." Practical examples of evaluation capacity building processes include training activities, involvement in evaluation, and technical assistance/coaching/support to increase the skills, knowledge, practices, and culture of organizations' individual staff and collective abilities.

Description of the Problem

Preskill (2014), a leading evaluation capacity building scholar who has mapped the field throughout its exponential growth over the last two decades, suggests "there is a good deal of agreement about the construct, goals, objectives, contextual variables, challenges, and opportunities for building evaluation capacity within organizations." However, despite the advancement, she called the need to evaluate evaluation capacity building activities the "elephant in the room," stating the need to focus on ensuring that our evaluation capacity building efforts make a difference through reaching the right people, increase organizational learning, and to investigate its influence. Other scholars agree the evidence for the impact of evaluation capacity building is minimal (Despard, 2016); while attention has focused on methods of evaluation capacity building, the study

of the assessment of organizational evaluation capacity has lagged behind (Nielsen et al., 2011; Fierro & Christie, 2016; Cheng & King, 2017).

It should be self-evident that the same systematic practices applied to investigating program impact should apply to investigating evaluation capacity building initiatives. An organization can apply the same methods and activities to plan and/or measure the success of evaluation capacity building initiatives. The assessment of organizational evaluation capacity prior to undertaking evaluation capacity building initiatives is meaningful for multiple reasons. First, whether explicit or implicit, perspectives on what constitutes evaluation capacity inevitably shape evaluation capacity building initiatives (Naccarella et al, 2007), and therefore stakeholders should agree upon what encompasses organizational evaluation capacity before planning activities. Further, measurement of organization evaluation capacity would assist in highlighting dimensions or areas in need of more concentrated focus. Lastly, the ability to find a baseline of evaluation capacity, and then repeat measurement in the future to reveal change, would serve to evaluate the results of evaluation capacity building activities (Preskill, 2014).

In response to the identified research gap on the assessment of organizational evaluation capacity, a small number of studies have presented theories and models of organizational evaluation capacity. However, to date there has not been a synthesis of their overlapping characteristics, limitations, and the level of evidence for their use. There is also an absence of guidance for organizations to implement the models and their accompanying tools for applied practice, as the research to date has focused on model/tool creation and validation. Nakaima and Sridharan (2017) suggest that the field

could benefit from "better stories of the dynamics of organizational capacity building from specific case studies." In other words, we need more research to bridge the gap between academic inquiry and applied use in the field.

To address these gaps, in this dissertation, I will present and integrate two related sub-studies. In sub-study 1, I undertake a qualitative evidence synthesis of the research theorizing organizational evaluation capacity models. Subsequently, I utilize one of the models and tools from the qualitative evidence synthesis and support its application at a multinational human service organization, to describe the experience from the perspective of the organization's staff. I use a mixed methods case study, interviewing monitoring, evaluation, accountability and learning (MEAL) specialists at the organization, as they implement an organizational evaluation capacity survey, interpret the results, and determine the next course of action in their evaluation capacity building initiatives. Finally, in the conclusion, I integrate the two sub-studies to theorize an application framework for implementation of evaluation capacity assessments.

Research Questions

My overarching goal in this dissertation is to support human service organizations in the implementation of organizational evaluation capacity assessments, to inform their evaluation capacity building plans. Sub-study details a qualitative evidence synthesis seeking to: (1) synthesize the extent of research theorizing organizational evaluation capacity models; (2) detail dimension commonality across EC models; (3) examine the extent the models have undergone tests of validity; and (4) identify possibilities for future research to expand the evidence base.

Sub-study 2 will examine how a human service organization applies an organizational evaluation capacity assessment tool to support capacity building goals, using a concordant mixed methods single instrumental case study. The study aims to detail the experiences in planning the process, implementing the tool, and interpreting the results, and determining future capacity building activities. There are three primary research questions for the concurrent mixed-methods single instrumental case study: (1) what considerations are necessary to implement an evaluation capacity assessment? (2) How do the evaluation experts in the organization interpret the results? And (3), how does the organization use the results to make decisions about investing in evaluation capacity building initiatives?

Significance of the Studies

The growing number of evaluation capacity models provide the opportunity to systematize a review of their overlapping characteristics and find opportunities for further research. No research to date has compiled all the existing models, chronologically analyzed their development and influences, analyzed the extent of dimension commonality, and reviewed the levels of validity the models have investigated. The results of sub-study 1 offer a meaningful contribution to the academic literature by providing data and commentary on the evidence base, as well as provide practitioners clear comparisons to find the right models and tools for their own use.

Another significant gap in the research is the absence of guidance for organizations to implement organizational evaluation capacity assessments for applied practice. Similarly, there is not a descriptive case of an organization undertaking the

process within the academic literature. Sub-study 2 is one of the first cases to describe the experience of a human service organization's application of a previously developed assessment of evaluation capacity and detail the intended use of their findings. The case details a meaningful example of the value and challenges of using an organizational evaluation capacity assessment tool, how the process can be imitated, and provides other organizations data to benchmark their own assessment's results. My hope is the studies inspire future literature creation to guide and direct organizations on the process, challenges, and benefits of using the tools validated in the academic community.

Lastly, to address the absence of guidance for organizations to use organizational evaluation capacity assessment tools, the conclusion integrates the data from studies 1 and 2 to create an application framework. The framework provides direct guidance for organizations to create and manage a process for similar assessments, to inform their evaluation capacity building initiatives and strategically invest in their organization's growth.

Ethical Considerations

My research did not work with vulnerable groups and did not ask questions of personal vulnerability in the survey. Accordingly, I received expediated approval from the University of Denver Institutional Review Board (IRB) before collecting data. Throughout the dissertation I protect the anonymity and confidentiality of the sample organization and research participants, to reduce or eliminate any possibility of harm from participating in this study. I obtained consent verbally from all participants in the case study before collecting any data and sent participants a document with a description

of the research study, informing them of their rights, and sharing the protocols in place to keep their responses anonymous and protect all data. The anonymity of the staff responses was critical to receive valid data and to avoid any type of negative consequences for the sample organization. At the completion of the analysis, I discussed interpretations of the assessment process with the participating organization to ensure approval of their portrayal in the narrative.

Researcher Positionality

Creswell (2007) states qualitative researchers should report their values and biases to "position themselves" in a study. I choose this topic because I am professionally engaged in the field of evaluation, have great interest in organizational evaluation capacity building, and have been employed in human service organizations. Accordingly, I attempted to take a reflexive approach to identify my biases and how they influence my interpretations of the data. Before undertaking the research, I followed suggestions from Holmes (2020) to identify and develop my positionality through locating myself about the subject; locating myself about the participants; and locating myself about the research context.

Locating myself around the subject, I have previously worked for multi-national NGOs for over 10 years. I have served in program management and evaluation contexts and implemented evaluation capacity building activities within an organization. I started this research with strong opinions about what has worked and why some initiatives have failed. I also have strong opinions about the nature of human service organizations, especially those that are non-profit. I attempted to be thoughtful about guarding against

drawing conclusions informed by my previous experience rather than the data, especially in the case study. This was most important when I was discussing the survey results with the case study organization's staff, attempting to collect their interpretations.

To locate myself about the participants, primarily in the case study, organizational staff viewed me as both a researcher but also at times as a "consultant" supporting their work. The distinction matters as my research goals connect to, but are different from, the direct goals of the organization. I tried to stress my role as a researcher with each participant, to fully inform or remind them of my role to create trustworthiness in the findings.

Lastly, with respect to the research, in my doctoral program I have had more training in quantitative methods than qualitative methods. I try to remain self-reflective on a bias towards quantitative ideas of validity and post-positivistic philosophies based on the emphasis of my academic training.

Organization of the Dissertation

My dissertation consists of six chapters, with two results chapters. In Chapter 1, I introduce the research questions, ethical considerations, and describe my positionality. In Chapter 2, I provide a literature review that overviews the research pertaining to program evaluation at human service organizations and evaluation capacity building. In Chapter 3, I detail the methodology used for the two sub-studies, describing the protocols for the qualitative evidence synthesis and concurrent mixed methods for the instrumental case study. In Chapters 4 and 5, I detail the results from the qualitative evidence synthesis and concurrent mixed methods instrumental case study, respectively. Lastly, in Chapter 6, I

conclude the dissertation with a discussion for each of the sub-studies, explicate an application framework for organizational evaluation capacity assessments, and make suggestions for future research.

Chapter Two: Literature Review

In this literature review, I seek to describe the challenges human service organizations face to evaluate their programs; provide a more detailed description of evaluation capacity building's goals and practices; and briefly introduce the literature on evaluation capacity assessments. The qualitative evidence synthesis in this dissertation (Chapter 4) adds to the literature review, as it attempts to systematically identify and explore models of organizational evaluation capacity in detail.

Program Evaluation in Human Service Organizations

Scholars have a diversity of definitions for program evaluation, but broadly define it as the process of assessing value of a program or policy. In one of the earliest models of evaluation, Stufflebeam defined evaluation in his model as, "the systematic process of delineating, obtaining, reporting, and applying descriptive and judgmental information about an object's value" (Alkin, 2013). In an early text on evaluation methods, Weiss (1998) defined evaluation with respect to program standards and goals, stating it is "the systematic assessment of the operation and/or outcomes of a program or policy, compared to a set of explicit or implicit standards, as a means of contributing to the improvement of the program or policy." Around the same time, Mark, Henry and Julnes (2000) defined evaluation with a focus on description, stating evaluations is "…systematic inquiry that describes and explains the policies and programs' operations,

effects, justifications, and social implications." The strength of the three definitions together illustrate program evaluation is systematic, assesses value of a program or policy, and describes its effects.

With those definitions and characteristics in mind, program evaluation allows human service organizations to understand how effective their programs are at meeting the social needs they set out to improve. In addition, it allows donors of those programs to feel secure their investments are making a difference, and help organizations tell stories about the public good they provide. Beyond improving and describing their programs, research has demonstrated increasing program evaluation practice helps organizations improve organizational learning, diversify funding opportunities, and undertake strategic planning (Preskill & Boyle, 2008; Labin et al., 2012; Despard, 2016). Additionally, there have been external pushes to increase evaluation practice, as human service organizations have faced steadily growing demand from numerous audiences and stakeholders to demonstrate program efficacy and accountability (Lynch-Cerullo & Cooney, 2011; Cheng & King, 2017). For all these reasons, human service originations have normalized program evaluation as a necessary function.

Carman and Fredricks (2008) undertook a study to understand the proportion of nonprofits conducting evaluations, the types of evaluation activities they perform, and how they are using the information gained from their evaluation practices. The sample of 189 organizations that responded to their survey found 90% did at least some evaluation and almost half made a concerted effort to evaluate most of their programs. Only 5% reported they did not evaluate any of their programs and organizational activities. The

survey found that the most prominent activities were reporting activities (94%), regulatory activities like audits (86%), and performance reviews (80%). However, well over half the organization reported formal program monitoring (69%) and formal program evaluations (55%). The most frequently reported use of evaluation practices was to help make changes in existing programs (93%), reporting to the board (82%) and to establish program goals or targets (75%). At least two-thirds of the organizations reported using evaluation data for strategic planning purposes (69%), decisions about staffing (68%), to help develop new programs (68%), and to report to funders (67%).

Organizational Facilitators and Obstacles to Evaluation

In a similar survey study, Mitchell and Berlan (2016) sought out to understand what public reporting charities perceived as catalysts and obstacles to their program evaluation practice. Based on their literature review, they hypothesized possible factors could include mandates from funders, internal and external requirements, the desire to understand and improve program effectiveness, organizational culture, information availability, special interest groups, leadership, and management support and commitment. The results of their survey found organizations perceived the strongest driving factors for program evaluation to be the desire to understand or improve program effectiveness, legitimacy, funding availability, clarity or specificity of goals, and organizational culture. They noted that, surprisingly, requirements from funders as significantly less important than anticipated.

In a 2020 study looking at organizations' evaluation capacity and their accountability motivations, Bryan, Robichau, and L'Esperance confirmed the modest

impact compliance or funder demands have on facilitating evaluation, finding no association with "upward accountability" and internal capacity to do evaluation in organizations. However, when they found "lateral accountability" present, referring to being accountable to the organization itself, by members of the staff, board, and volunteers, there was significant correlation with staff competencies in evaluation and organizational learning climate. The study also found a correlation with "downward accountability," prioritizing beneficiaries and clients by ensuring quality program implementation, with a strong learning climate. The authors suggest organizations where managers emphasize internal accountability practices value inquiry and application of learning in a substantive way, rather than just reporting to a funder.

Mitchell and Berlan (2016) used their survey data to create a predictive model weighing catalysts for evaluation practice with an outcome of evaluation rigor. They found that evaluation appears to be most rigorous when (1) evaluation is a priority; (2) a supportive organizational culture exists; (3) management requires evaluation rather than funders; (4) measurement is not difficult; (5) evaluation is not primarily motivated by personal interest; and (6) evaluation is likely to reveal success. Similar to the findings from Bryan, Robichau, and L'Esperance (2020), it appears internal or "lateral" motivation for program evaluation leads to more rigorous evaluation practice, higher staff competencies, and stronger learning cultures.

There are common obstacles and challenges organizations face even when aspiring to rigorous evaluation beyond the absences of the catalysts that facilitate evaluation in organizations. In their survey of 179 nonprofit organizations, Carman and

Fredricks (2010) reported the most common challenge identified was not having enough time, identified by 68% of the organizations. The next three biggest obstacles were cited by about 50% of the organizations: not enough funding (51%), not enough evaluation expertise (50%), and not enough staff (49%). Some additional challenges on their survey that have different themes from funding and staff included the cost of technical assistance (36%), data collection issues (35%), data management issues (21%), confidentiality issues (21%), and lack of support from the board (16%). One issue absent in their survey, but often cited as an obstacle for many organizations, is employee turnover (Preskill & Boyle, 2008). Turnover is a particularly challenging issue because it can mitigate efforts to build capacity through staff training, if they staff who receive the training end up leaving the organization. Carman and Fredricks found no significant differences between the different types of nonprofit organizations and the types of obstacles they cited.

Organizational Leadership

Organizational leadership is an important enough factor in organizational development of evaluation practice to warrant its own section (Preskill & Boyle, 2008; Labin et al., 2012; Preskill, 2014). Labin et al. (2012) suggested at the organizational level, leadership, culture, and resources are highly related; however, the study found leadership was the least frequently targeted organizational factor and the least frequently reported organizational outcome when building evaluation capacity. Preskill (2014) agreed with the finding stating the field had not "paid enough attention to the role senior

leaders play in organizations, and how they influence, shape, and sustain an evaluation and learning culture."

In 2008, Preskill and Boyle's Multidisciplinary Model of Evaluation Capacity Building proposed the extent to which the organization's leadership values learning and evaluation, and creates a culture of inquiry, significantly affects an organization's ability to build evaluation capacity or if practice becomes sustained. In 2020, Wade and Kallemeyn performed an interview study to understand if organizations that undergo an evaluation capacity building intervention sustain evaluation practice and confirmed leadership was a primary support for sustainability. They identified four sources of leadership for evaluation: (1) Board involvement, (2) supportive Executive Directors, who provided resources for evaluation, (3) Executive Directors who were fully engaged in evaluation, and (4) staff champions in leadership roles. In a different interview study regarding evaluation policies in organizations, Al Hudib & Cousins (2021) confirmed importance of leadership stating it:

> Serves to link the capacity to do evaluation and the capacity to use evaluations. Leadership emerges as a variable critical for supporting or leveraging the capacity to plan and implement evaluation effectively, on the one hand, while being responsible for integrating evaluation into decision-making processes, on the other.

Initiatives to increase evaluation capacity risk effectiveness without focusing on leadership in an organization. A study by Lawrenz et al. (2016) found that staff who had trained in evaluation practice desired to implement the lessons but were unable to because they did not have direct control over their time. They suggested the results underscored the critical importance of organizational understanding and appreciation of

the value of evaluation, as the staff given capacity building training did not often control the institutional resources needed to conduct evaluations. As noted in the preceding section, this finding correlates with the two most common cited obstacles to evaluation practice in organizations: not having enough time and funding.

Preskill (2014) provides five suggestions for how leadership can support evaluative thinking and practice. First, leadership should understand how strategy and evaluation interconnect. This can include the development and support of organizational evaluation policies and frameworks which I discuss in the subsequent section. Second, leadership must provide adequate resources for evaluation, which would mitigate the main obstacles cited in evaluation practice of not having enough time, staff, and funding. Put another way, Carman and Fredericks (2010) stated that organizations should view time, funding, and other resources for evaluation as the cost of doing business. Third, leadership should be "active consumers of evaluation information," demonstrating the link between the capacity to do and use evaluation as noted by Al Hudib and Cousins (2021). Fourth, leadership can build their board's understanding of and support for evaluation, connecting on of the main leadership stakeholders noted by Lawrenz et al. (2016). And lastly, Preskill suggests leadership promote evaluation as a means for ongoing organizational learning.

Organizational Evaluation Policies

Al Hudib and Cousins (2021) illustrate how organizations have evaluation policies to govern how they conduct evaluations, to demonstrate how they intend to be accountable to stakeholders, to support organizational learning, and to use evidence for

decision-making. Trochim (2009) defines an evaluation policy as "any rule or principle that a group or organization uses to guide its decisions and actions when doing evaluation" and believes they can even be unwritten and implicit norms that guide evaluations in an organization. Trochim provides common categories of policies to include evaluation goals; evaluation participation; evaluation capacity building; evaluation management, evaluation roles; evaluation process and methods; evaluation use; and evaluation quality.

Evaluation policies can impact organizational evaluation capacity and effectiveness in multiple ways. First, an evaluation policy can be a communication tool within an organization and to its stakeholders, helping to clarify beliefs and expectations about evaluation (Preskill & Boyle, 2008; Trochim, 2009). Al Hudib and Cousins (2021) underscore the meaningful potential to influence evaluation practice, to include when to evaluate outcomes, how to evaluate, who evaluates and their roles, and to dictate resources. Additionally, Trochim (2009) suggests written evaluation policies can make evaluation a more transparent and democratic endeavor, engendering participation and dialogue.

Evaluation Capacity Building

Over the last few decades, human service organizations have faced steadily growing demand from external audiences to increase their use of evaluation to demonstrate program efficacy and accountability (Lynch-Cerullo & Cooney, 2011; Cheng & King, 2017). In addition to the external push to meet funder accountability reporting, internal incentives to increase program evaluation efforts exist like the desire

to improve organizational learning, respond to new funding opportunities, support strategic planning, and increase beneficiary impact (Preskill & Boyle, 2008; Labin et al., 2012; Despard, 2016). The resulting factors have normalized program evaluation as a necessary function in the human service landscape, with donors asking for more data and reporting, and organizations placing a higher priority on organizational learning cultures (Carman, 2011; Mitchell & Berlan, 2018). However, the cultural shift created the need to acquire new skills and practices, as human service organizations' accountability metrics and analysis previously focused on financial health, program expense ratios, and funding controls (Mitchell & Berlan, 2018).

Despite the practice of program evaluation now normalized in human service organizations' practice, organizations with limited resources struggle with evaluation and reporting to their funders (Carman, 2011). The lack of resources and capacity, both from financial and human capital, result in organizations not acquiring the skills and systems to properly embed program evaluation best practices in their organization (Preskill & Boyle, 2008; Labin et al., 2012; Cheng and King, 2017). The increased demand along with the struggle to scale up program evaluation practice has led to the rapid emergence of the field of evaluation capacity building (Cousins, Goh, Clark, & Lee, 2004; Preskill & Boyle, 2008).

Multiple definitions of evaluation capacity building exist in the literature but there is consensus that it is a complex and contextual process to improve program evaluation outcomes and embed evaluation best practices into an organization (Preskill & Boyle, 2008; Cousins et al., 2008; Labin et al., 2012; Cheng and King, 2017). Nielsen et al.

(2011) argue the diverse definitions are rooted in different understandings of the purpose of evaluation (e.g., evaluation as a management tool, as a research tool for understanding interventions, and/or for accountability). One of the most commonly used definitions, and the I adopt in this dissertation, is from the systematic synthesis undertaken by Labin et al. (2012): an intentional process aimed at increasing the individual motivation, knowledge, skills, and to enhance an organization's ability to conduct and use evaluation. Although the field has rapidly developed over the past two decades, codifying important concepts and activities, scholars believe there is more to investigate including the processes to foster and sustain evaluation capacity building initiatives in different settings, its outputs, and ultimate outcomes (Preskill, 2014; King, 2020).

Evaluation Capacity Building Outcomes and Activities

In 2014, Labin attempted to use multiple evaluation capacity building measurement tools and map their concepts onto a common framework to help define activities and outcomes for evaluation capacity building. Labin's Integrated Evaluation Capacity Building Model suggested common outcomes of evaluating capacity building to include individual and organization level impact. For the individual level, Labin suggested evaluation capacity building can change attitudes toward evaluation and increase individual expertise, knowledge, skills, and behavior. At the organizational level, evaluation capacity building can help leadership adopt a mindset of evaluative thinking; build organizational culture; develop processes, polices and best practices to support evaluation and its use; and mainstream evaluation capacity through integration into roles, planning, and monitoring.

Evaluation capacity building processes use diverse approaches and activities to achieve those outcomes. Preskill and Boyle (2008) categorized 10 strategies used for evaluation capacity building: internship; written materials; technology; meetings' appreciative inquiry; communities of practice; training; involvement in an evaluation process; technical assistance' and coaching and mentorship. In their synthesis of evaluation capacity building studies, Labin et al. (2012) found that the most frequently used activities are training; involvement in evaluation; and technical assistance, coaching, or support. Multiple studies have shown indirect evaluation capacity building, like involving staff in an evaluation, is an impactful way to increase individual knowledge, skills, and develop an appreciation for evaluation (Bourgeois et al., 2008; Labin et al., 2012; Cousins & Bourgeois, 2014). Along with diverse strategies and activities, Labin et al. (2012) found that initiatives deliver evaluation capacity building through different mechanisms, from in-person meetings, remote connection and web-based mechanisms, and through written materials such as evaluation manuals.

Preskill and Boyle (2008) provide guidance on the planning of evaluation capacity building activities and approaches, stating organizational resources, staff roles and characteristics, current evaluation practices, and desired learning objectives and expected outcomes should dictate the processes. In their study on evaluation policy and its influence on organizational evaluation capacity, Al Hudib and Cousins (2021) strongly affirmed the role of sociopolitical, cultural, and economic contexts in an organization, calling it the most influential consideration shaping an evaluation capacity building process. They state context is so critical because it helps "explain how the interaction

between hierarchies, systems, structures, and people influences evaluation processes and use within organizations." Accordingly, every evaluation capacity building initiative should start with an inventory or assessment of current organizational context, which I investigate further in a subsequent section on evaluation capacity building assessments in the qualitative evidence synthesis (Chapter 4).

Facilitators and Challenges to Success in Evaluation Capacity Building

One of the foundations of evaluation capacity building is the importance of participatory processes (Labin et al., 2012). The types of activities suggested in the preceding section are evidence of participation's centrality, like involvement in evaluations, mentorship, coaching, and communities of practice. Labin et al. (2012) found experiential learning activities paired with training and technical assistance were associated with successful achievement of individual knowledge and behavioral outcomes. The study also found that technical assistance was critical in sustaining organizational changes from evaluation capacity building initiatives, suggesting organizations should therefore consider the need for ongoing resources even beyond initial capacity building activities.

Building on the concept of sustained changes, Wade and Kallemeyn (2020) performed an interview study to understand if organizations that undergo an evaluation capacity building intervention sustain evaluation practice and how it develops over time. Similar to previous studies, their findings suggested two of the most critical factors were leadership and dedicated resources. Additionally, the study affirmed the experiential and participatory findings from Labin et al. (2012), stating that practicing and using

evaluation were highly impactful for learning, but opportunities only existed when organizations provided resources and encouraged staff spend time on evaluations.

Preskill and Boyle (2008) stress the need for evaluation capacity building to target the "cultural" level of an organization. Evidence of impact at the "cultural" level include leaders communicating the value of evaluation, evaluation's visibility in the organization's strategy documents, and evaluation practices embedded in the organization's policies and systems. Targeting the "cultural" or organizational level for impact can also help mitigate the challenges from staff turnover, as individual level capacity building activities can be mitigated if the staff trained leave the organization.

Another major challenge to successful implementation of evaluation capacity building initiatives is the lack of metrics for assessment and measuring progress in capacity building efforts (Preskill 2014, Despard, 2016, Nakaima & Sridharan, 2020). Preskill (2014) notes that initiatives rarely involve a contract for evaluating their results and therefore evidence of the effectiveness of different designs is inadequate. Nakaima and Sridharan (2020) suggest we need metrics at the organizational level for concepts like organizational skills, culture, commitment, and knowledge. These metrics would help initiatives choose approaches and strategies, as well as measure change over time. Additionally, along with the metrics, Nakaima & Sridharan (2020) believe there is a need for better stories of the dynamics of organizational capacity building using case studies, to more richly illustrate what works in practice and why. The mixed methods case study in Chapter 5 seeks to respond to this need.

Organizational Evaluation Capacity Assessments

In 2014, Preskill suggested "there is a good deal of agreement about the construct, goals, objectives, contextual variables, challenges, and opportunities for building evaluation capacity within organizations." However, despite the advancement, she called the need to evaluate evaluation capacity building activities the "elephant in the room," stating the need to focus on ensuring that our evaluation capacity building efforts make a difference through reaching the right people, increase organizational learning, and are able to be investigated for their influence. Other scholars agree the evidence for the impact of evaluation capacity building is minimal (Despard, 2016); while attention has focused on methods of evaluation capacity building, the study of the assessment of organizational evaluation capacity has lagged behind (Nielsen et al., 2011; Fierro & Christie, 2016; Cheng & King, 2017).

It should be self-evident that the same systematic practices applied to investigating program impact should apply to the impact of evaluation capacity building initiatives. Moreover, the assessment of organizational evaluation capacity prior to undertaking evaluation capacity building initiatives is meaningful for multiple reasons. First, whether explicit or implicit, perspectives on what constitutes evaluation capacity inevitably shape evaluation capacity building initiatives (Naccarella et al, 2007), and therefore stakeholders should agree upon what encompasses organizational evaluation before planning activities. Further, measurement of organization evaluation capacity would assist in highlighting dimensions or resources in need of more concentrated focus. Lastly, the ability to find a baseline of evaluation capacity, and then repeat the

measurement in the future to reveal change would serve to evaluate the results of evaluation capacity building activities (Preskill, 2014).

Evaluation capacity and evaluation capacity building are related but different concepts. As previously noted, evaluation capacity building is "an intentional process aimed at increasing the individual motivation, knowledge, skills, and to enhance an organization's ability to conduct and use evaluation" (Labin et al., 2012). Evaluation capacity is the outcome of evaluation capacity building (Cheng and King 2017). Literature on evaluation capacity agrees it is a multidimensional construct, reflecting the definition and practice of evaluation capacity building, focusing on dimensions like the ability to do and use evaluation, organizational learning, and the skills of individuals within the organization.

Over time, the models of organizational evaluation capacity in academic literature have expanded from conceptual models based on expert reflections (e.g., Stufflebeam, 2002; McDonald, Rogers, & Kefford, 2003; Cousins, Goh, Clark, & Lee, 2004; Naccarella et al., 2007; Volkov & King, 2007), to structural models relating the dimensions to one another (Bourgeois & Cousins, 2008; Taylor-Powell & Boyd, 2008; Preskill & Boyle, 2008), and finally to quantitively developed models using surveys to identify dimensions and measure relationships (e.g., Nielsen, Lemire, & Skov, 2011; Taylor-Ritzler et al. 2013; Cousins et al., 2014; Bourgeois, I., Whynot, & Thériault, 2015; Cheng & King, 2017; Gagnon et al., 2018). While case studies have helped demonstrate qualitative content validity, only a few models have tested statistical analyses for construct validity (Nielsen et al., 2011; Taylor-Ritzler et al., 2013; Gagnon et

al., 2018), and questions concerning reliability, transferability, and overall outcomes remain (Cheng & King, 2017*;* El Hassar, Poth, Gokiert & Bulut, 2021). This dissertation explores the models in detail in its qualitative evidence synthesis on theories of organizational evaluation capacity (Chapter 4).

Organizational context plays a critical factor in evaluation capacity building strategy selection and overall success (Preskill & Boyle, 2008; Al Hudib and Cousins, 2021). Examples include, at minimum, organizational resources available, staff roles and characteristics, current evaluation practices, and desired learning objectives and expected outcomes. To understand the current context, El Hassar, Poth, Gokiert and Bulut (2021) suggest using an assessment tool to inform the organization's conceptualization of evaluation capacity and inform development of capacity building initiatives. This dissertation explores using one of the models from research, contextualizing to a multinational human services organization, and presents the results in a mixed methods case study (Chapter 5).

Chapter Three: Methods

The overarching goal of this dissertation, comprised of two related sub-studies, is to support human service organizations in the implementation of organizational evaluation capacity assessments to inform their evaluation capacity building efforts. In sub-study 1, I will present a qualitative evidence synthesis seeking to: (1) synthesize the extent of research theorizing organizational evaluation capacity models; (2) detail dimension commonality across EC models; (3) examine the extent the models have undergone tests of validity; and (4) identify possibilities for future research to expand the evidence base.

In sub-study 2, I will examine how a human service organization applies an organizational evaluation capacity assessment tool to support evaluation capacity building goals, using a concurrent mixed methods single instrumental case study. The study aims to detail the experience of planning the process, implementing the tool, interpreting the results, and determining future capacity building activities. There are three primary research questions for the case study: (1) what considerations are necessary to implement an evaluation capacity assessment? (2) How do the evaluation experts in the organization interpret the results? And (3), how does the organization use the results to make decisions about investing in evaluation capacity building initiatives?

Lastly, in the conclusion, I will use an inductive research approach, borrowing methods from the building theory from case study method, to integrate the data and findings from sub-studies 1 and 2, to theorize an application framework for organizational evaluation capacity assessments.

Qualitative Evidence Synthesis

Sub-study 1 is a qualitative evidence synthesis of organizational evaluation capacity models. A qualitative evidence synthesis is a method for collecting and analyzing qualitative studies to uncover themes or constructs across selected publications (Grant & Booth, 2018), similar to a meta-analysis or systematic review. Traditionally, a meta-analysis combines studies that have tested the same hypothesis or implemented the same intervention, to aggregate the results and use statistical methods to make claims about effect size. Scholars distinguish a systematic review by its systematic approach to identifying studies, appraising their quality, and summarizing the evidence (Khan, Kunz, Kleijnen, & Antes, 2003). Generally, systematic reviews have been associated with meta-analysis and quantitative methods. However, reviews including qualitative studies can follow similar protocols using the same "replicable, rigorous, and transparent methodology and presentation" (Siddaway, Wood, & Hedges, 2019).

There results of a qualitative evidence synthesis can be the development of a new theory, a summation of research to date with an overarching narrative, or a wider generalization than studies can make on their own (Grant & Booth, 2018). However, there is not consensus on the exact methods of a qualitative evidence synthesis, demonstrated by the diversity of terminology for similar reviews. Grant and Booth

(2018) suggest qualitative systematic review and qualitative evidence synthesis are used synonymously. Siddaway, Wood, & Hedges (2019) suggest the terms meta-synthesis, qualitative meta-analysis, and meta-ethnography can be used when reviews integrate qualitative research. Similar to Grant and Booth's outcomes for qualitative evidence synthesis, Siddaway, Wood, & Hedges suggest qualitative meta-synthesis "locate themes, concepts or theories that provide novel or more powerful explanations for the phenomenon under review."

This study adopts the term qualitative evidence synthesis for a few reasons. First, as Grant and Booth (2018) point out, the terms meta-synthesis and meta-ethnography are lacking because combining the terms meta and synthesis is tautological, and meta-ethnography is misleading as reviews of qualitative research are not bound by the ethnographic method. Second, for this study, I used methods that closely resemble the methods of a traditional systematic review but with minor adaptations, like the absence of methodological quality assessments. Accordingly, as Grant and Booth propose qualitative systematic review and qualitative evidence synthesis are used synonymously, the term qualitative evidence synthesis feels most appropriate. Lastly, the Cochrane Collaboration's handbook uses the term qualitative evidence synthesis leading Grant and Booth (2018) to suggest the term is moving towards greater consensus.

Despite the method title, this study expected to collect a few articles that use quantitative methods to create or test their models of organizational evaluation capacity. However, it is highly unlikely the models will be similar enough to allow for any type of statistical meta-analysis. More importantly, this study is interested in summarizing the

research to date with an overarching narrative and finding commonalities across the diverse models. Accordingly, even when reviewing the articles with quantitative methods, this study, its research questions, and its methods have a clear qualitative focus.

Although debated, common practice of the qualitative evidence synthesis method is the use of the systematic review process for a methodological base. This study borrowed its methods from a five-step process to complete a systematic review found in Khan, Kunz, Kleijnen, & Antes (2003), with minor adaptions to account for the qualitative evidence base. The five steps are to (1) define the question, (2) identify relevant publications, (3) assess the studies for eligibility, (4) summarize the evidence, and (5) interpret the results. I detail each of the steps in the subsequent paragraphs.

Defining the Question

Khan, Kunz, Kleijnen, and Antes (2003) state the researcher should create the research question through free-form query. For this study, my initial question is: what is the extent of research theorizing organizational evaluation capacity models? With the initial free form question identified, the authors state reviewers should pose a more structured and explicit statement to include population, intervention, outcome, and study design. Accordingly, I seek to understand how human service organizations attempting to build evaluation capacity (population), can apply organizational evaluation capacity models (intervention), to support planning of capacity building initiatives (outcome), through a qualitative evidence synthesis of the academic and grey literature (study design).

Building off my explicit research statement, the aims of the study are to: (1) synthesize the extent of research theorizing organizational evaluation capacity models; (2) detail dimension commonality across EC models; (3) examine the extent the models have undergone tests of validity; and (4) identify possibilities for future research to expand the evidence base.

Search Strategy

According to Khan, Kunz, Kleijnen, and Antes (2003), the next step is to identify all relevant publications. My search strategy had two components to attempt an exhaustive search: an academic literature review and supplementation through a grey literature search. Including unpublished studies or tools produced by organizations outside of academic publishing is critical to minimize bias and maximize the representativeness of the sample (Higgins et al., 2020). Evaluation capacity models from grey literature are particularly important as assessment tools developed by practitioners and organizations are likely to reside in their private files or organizational websites.

I used multiple search strategies suggested by the Cochrane Collaboration's handbook's chapter on qualitative evidence. First, I conducted a topic-based search, in the academic literature, retrieving possible studies through initial reviews of titles and abstracts. I leveraged electronic databases with a focus on ProQuest, SAGE Research Methods, SAGE Premier, and Science Direct, targeting peer-reviewed articles from 2000-2021. Primary search keywords included evaluation capacity assessment, evaluation capacity measurement, and evaluation capacity building. Additionally, I augmented the keywords using: organizational, nonprofit, and human service. I used

Boolean search operators (AND, OR, etc.) and truncation symbols to account for all words starting with a particular combination of letters (organization$ to include organization, organizations, organizational, etc.). In addition to the aggregate databases, I individually searched the following evaluation journals: American Journal of Evaluation, Canadian Journal of Program Evaluation, New Directions for Evaluation, and Evaluation and Program Planning.

If an article appeared promising in the initial review, I obtained the full text for an eligibility review in the next step. Siddaway, Wood, and Hedges (2018) suggest that although most articles collected through initial searches will not meet eligibility criteria, they recommended the researcher err on the side of "sensitivity," meaning the proportion of true positives correctly identified. Accordingly, I followed their advice to collect as many articles with potential as possible, before the formal exclusion process, to mitigate the risk of missing important information.

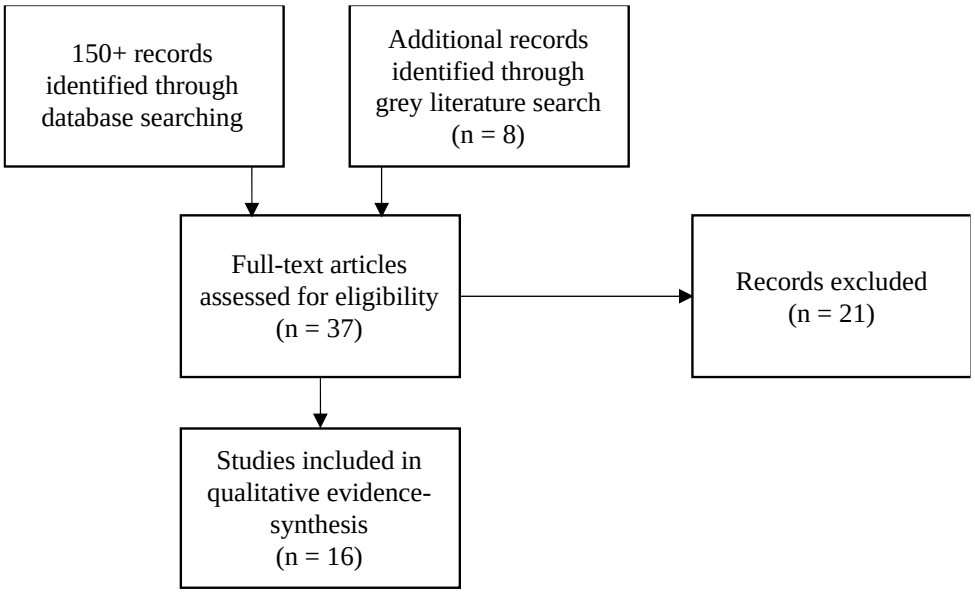
Another strategy suggested in the Cochrane Collaboration handbook's chapter on qualitative evidence is to "minimize reliance on topic-based searching and rely on citations-based approaches to identify linked reports, published or unpublished, of a particular study." To that end, I used manual article identification to supplement the automated search process, using reverse citation searches in reference sections of articles meeting eligibility criteria. The use of eligible studies' reference sections meant an iterative process where after the initial automated search I began reviewing studies for eligibility (step 3) and then returned to the search process (step 2). This iterative process was invaluable to gaining an understanding of when the search process approached

saturation, as I began to recognize most of the literature referenced in contemporary studies.

Lastly, correspondence with practitioners has proved an important source of information for grey literature (Higgins et al, 2018). Moreover, the Cochrane Collaboration's handbook's chapter on qualitative evidence suggest sources of qualitative evidence are more likely to include grey literature than quantitative evidence. Accordingly, this study reached out to four evaluation directors at nonprofits, two academic researchers on evaluation capacity, and two consultants that work with nonprofits, in an effort include relevant unpublished work and applied tools. I asked the evaluation experts and academics to share literature, workbooks, checklists, manuals and academic research that guide or influence their practice.

Figure 3.1

Qualitative Evidence Synthesis Flow Diagram



Eligibility Criteria and Selection

I used the structured and explicit research statement developed in the first step, which included population, intervention, outcome, and study design details, to inform the initial "sensitive" search process. The next step was to define all the eligibility criteria and begin the selection process using the eligibility criteria. Additionally, this is the step that most meaningfully modifies the systematic review process found in Khan, Kunz, Kleijnen, and Antes (2003). The authors' third step is to assess the studies' quality, a process most applicable when the main source of evidence is from randomized control trials. I collected articles with both quantitative and qualitative methodologies and therefore cannot assess quality as in a traditional systematic review. Accordingly, this step will finalize the study selection process by reviewing all academic and grey literature from the saturated search process.

Eligibility criteria on content included all studies proposing a theory, framework, or model of organizational evaluation capacity. I anticipated those theories, frameworks, or models to be in a diversity of forms from narrative, to conceptual, or in tool form (checklists, surveys). Models developed from qualitative or quantitative methods were eligible and did not need to include an assessment tool. Additionally, case studies on the application or testing of an existing EC model within a human service organization were eligible to understand if application modified a model. Models of individual practitioners' evaluation capacity or competencies were not eligible.

The study limited organizations of focus in the models to non-profit or governmental human-service organizations. Some of the organizations of focus in case

studies were not located in the United States, however, I only considered articles written in English for eligibility. Lastly, as the body of literature studying organizational evaluation capacity remains limited, but the practice of building evaluation capacity is common, applied tools and unpublished models were eligible, collected from the grey literature search. The time constraints of inclusion criteria were publication or creation in 2000-2021 to account for the most current beliefs and practices concerning evaluation capacity building.

With the eligibility criteria completed, I began the full-text reviews and selection process. Siddaway, Wood, and Hedges (2019) suggested this is the stage my focus needed to "shift from sensitivity to specificity" as I reviewed the full text of potentially eligible studies. Specificity in testing measures the proportion of true negatives correctly identified. In other words, becoming more specific meant to carefully inspect full texts to ensure I rightly removed all studies not meeting eligibility criteria.

I initially reviewed thirty-seven full texts from the academic literature, with sixteen meeting eligibility criteria. Seven articles used qualitative methods to develop a theory or model and six articles used quantitative methods to create or refine a structural model. The grey literature search returned some of the same documentation found from the academic literature review, as well as toolkits found on organizational websites like the World Bank or USAID. Upon eligibility review, three assessment tools that not published in academic journals met inclusion criteria (i.e., the Readiness for Organizational Learning and Evaluation (Preskill & Torres, 2000), the Institutionalizing

evaluation checklist (Stufflebeam, 2002) and the ECB Checklist (Volkov & King, 2007)). However, I found each of the three assessment tools cited in the academic literature.

Evidence Summary

Once I had undertaken full-text reviews of all the studies from the selection process I began the process of reviewing and summarizing the eligible studies. However, Siddaway, Wood, and Hedges (2019) point out summarizing the data is not enough and critical thought and reflection are required to advance the field's theoretical understanding through interpretation. The authors suggest the researcher must "zoom out" and link concepts, explore reasons for variations, and critique the overall evidence. Accordingly, I undertook a three-step process for summarizing and critically interpreting the evidence.

First, I wrote annotations for each of the eligible studies, in chronological order, considering and critiquing their contribution to the evidence base. The process of writing annotations and considering how the evidence builds over time helped me develop a conceptual understanding of the evidence base and begin considering overall narrative themes. Next, I identified and extracted key data points for each eligible article and tool (e.g., study design, sample, type of model, dimensions of the model, assessment items, validation tests, origin of instrument creation, and applied use of the model) to organize into a guiding table (Pajo, 2018) to facilitate the comparison of model elements across studies. I used an iterative process for creating the categories in the guiding table allowing for emerging categories to manifest or consolidate existing categories.

Next, with a strong grasp of the evidence base from the annotations and guiding table, I created an initial "lean" list of *a priori* codes (Creswell, 2007). The primary aims of the study informed the initial codes, with an emphasis on dimensional commonality and tests of validity or discussions of trustworthiness. For example, I expected models to have different terminology for similar concepts and used *a priori* codes to begin consolidating the model dimensions. I then reread the annotations and guiding table, marking the data with codes in Atlas.ti, to consolidate evidence within a code. The coding process was iterative and dynamic, allowing myself the ability to drop *a priori* codes for new emerging categorizations, especially for dimension comparison. The coding was complete when I had a final table summarizing the common model dimensions to support the interpretative phase.

Interpretation

Siddaway, Wood, and Hedges (2019) suggest meaningful output of the study is for the discussion to propose a new conceptualization or theory. Working towards that goal, in my last step, interpretation of the evidence began with thematic analysis using the evidence aggregated in codes. The four aims of the study became a starting list of *a priori* themes, and I used an inductive approach to find emerging themes across the research questions. When finding patterns in the evidence I would return to the full texts and used memoing as a means to begin narrating my conclusions.

In writing the interpretation and discussion content, I used the themes and codes to explicitly link my conclusions to the evidence. Found in Chapter 4 (Results of the Qualitative Evidence Synthesis) and Chapter 6 (Conclusion), I provide my interpretation

of the strengths and limitations of the literature, including a consideration for the direction of the evidence base (i.e., methodological trends, recommendations for practice, ideas for future research), and demonstrated commonalities across the diverse models to illustrate consensus in model theory and direct future testing. I also include an assessment of representativeness of the sample and critique my methods and process to search for more the possibility of sample bias and the subsequent possible implications for the analysis.

Concurrent Mixed Methods Single Instrumental Case Study

In sub-study 2, I describe the experience of a multinational NGO's application of an assessment of evaluation capacity. The study is interested in participants' negotiation of the implementation process and perspectives on the utility of the results. The method for the study is a concurrent mixed methods single instrumental case study to investigate and uncover the specific issues and considerations organizations face when implementing evaluation capacity assessments. Stake (1995) suggests a single instrumental case design can facilitate the creation of other theories through an in-depth study, reviewing the context and detailing ordinary activities. To that end, this study will fill a gap in literature, providing an account to support other organizations' implementation of evaluation capacity assessments and support development of implementation frameworks and guidance.

Philosophical Underpinnings

I use evolving constructivist philosophy as the philosophical underpinning of the study, consistent with a mixed methods design to generate and describe a case, as

suggested by Creswell and Plano Clark (2017). Although the I use a quantitative survey oriented towards post-positivistic claims, the study's research questions concern the organization's perspectives on the assessment process and use of the results, which orient towards qualitative data and the philosophy of constructivism. The evolving constructivist philosophy allowed me to take a post-positivist approach to the quantitative survey analysis but quickly transition to constructivism for the majority of the study. The value of the evolving constructivist philosophy, especially for this study, is to facilitate "different ways of knowing about and valuing" the case, which Creswell and Plano Clark (2017) suggest can contribute to new and different insights.

In general, the constructivist philosophy also allows participants to generate or inductively develop a pattern of meaning (Creswell, 2007). Creswell suggests subjective meanings are negotiated socially, through interaction with others, and constructivist research can illustrate the process of interaction among individuals. The notion of socially negotiating meaning is well-suited for this study as organizations, due to their structure, demand social negotiation and inherently contain different "realities" and perspectives. Additionally, the case study's emphasis was on the participants' perspective on the value and meaning of the assessment process and results.

Theory can influence a mixed methods case study through informing the case description. Theories often provide a guiding perspective for integrating the different data sources. In this case I apply a social science theory, as a theory of organizational evaluation capacity (Gagnon et al., 2018) guide the questions and techniques in the study. The dimensions of the theory informed the study's data collection and analysis, and

therefore provide direction on future decision making given the results are identifying capacity gaps for the sample organization to target for growth.

Case Study Rationale

Creswell (2007) suggests some clear benefits provided by case study. First, the method provides a meaningful way to understand a phenomenon within a context and demonstrate the real-life scenarios the phenomenon encounters. Another benefit is the creation of substantiative narratives from the perspective of those who participate in the real-life scenarios. Lastly, Creswell states case studies contribute qualitative empirical data to the research base that may otherwise lack rich detail. Each of these benefits describe the value this study provides the academic literature. The case in Chapter 5 provides a rich account of a real-world organization implementing and using an evaluation capacity assessment, detailing the necessary considerations and challenges they faced, which is a novel addition to the academic literature.

According to Stake (1995), an instrumental case study has *a priori* research questions informing the chosen bounded case, to develop understanding of a complex phenomenon in a given context. The research questions of this study aim to describe the experience of a multinational NGO's application of an assessment of evaluation capacity, providing some clear criteria for case selection. Additionally, to provide a meaningful account of the experience I aimed to provide thick description of the process, experiential understanding of the choices made, and detail diverse perspectives from participants.

Intent for Mixing Methods

A mixed methods case study design is consistent with the fundamental goal of a generic case study: developing a detailed understanding of a case through gathering diverse sources of data (Creswell & Plano Clark, 2017). The preceding section details why an instrumental case study is well-suited for the *a priori* research questions. The research questions also direct the choice of this complex mixed methods design, based on needing to use both quantitative and qualitative information to best describe the case. For this study, the implementation of the quantitative assessment is necessary to answer the experiential qualitative questions about the overall process and organization's use of results.

The study meets each of the four primary characteristics Creswell and Plano Clark (2017) suggest describes a mixed methods study. First, I collect and analyze both quantitative and qualitative data through the survey analysis, observational data, interview data, and document review. Second, I integrate the methods through the survey results informing the interview protocol, and the interview data helping to explain the sample organization's perspective on survey results. Third, I organize its procedures into specific research designs (concurrent mixed methods, instrumental case study) that provide the logic and procedures for conducting the study. And lastly, I have framed these procedures within theory (the model of evaluation capacity) and philosophy (evolving constructivism). The relative importance of the study is on the qualitative methods, as the organization's perspective on the process and use of the results answer the primary research questions.

Mixed methods representation. The procedural flow diagram presented in Figure 1 illustrates the process and integration points of the mixed methods. Using diagram notation for mixed methods from Creswell and Plano Clark (2017), I have followed their ten guidelines for drawing procedural diagrams. I used boxes for the quantitative and qualitative stages of data collection and data analysis. Within the boxes, uppercase or lowercase letters designate the relative priority of the quantitative and qualitative data collection and analysis (the diagram uses quan and QUAL). Single headed arrows illustrate the flow of the process which is important for interpretation as the I concurrently implemented the methods. Lastly, I list procedures and products for each stage with qualitative methods on the left and quantitative methods on the right.

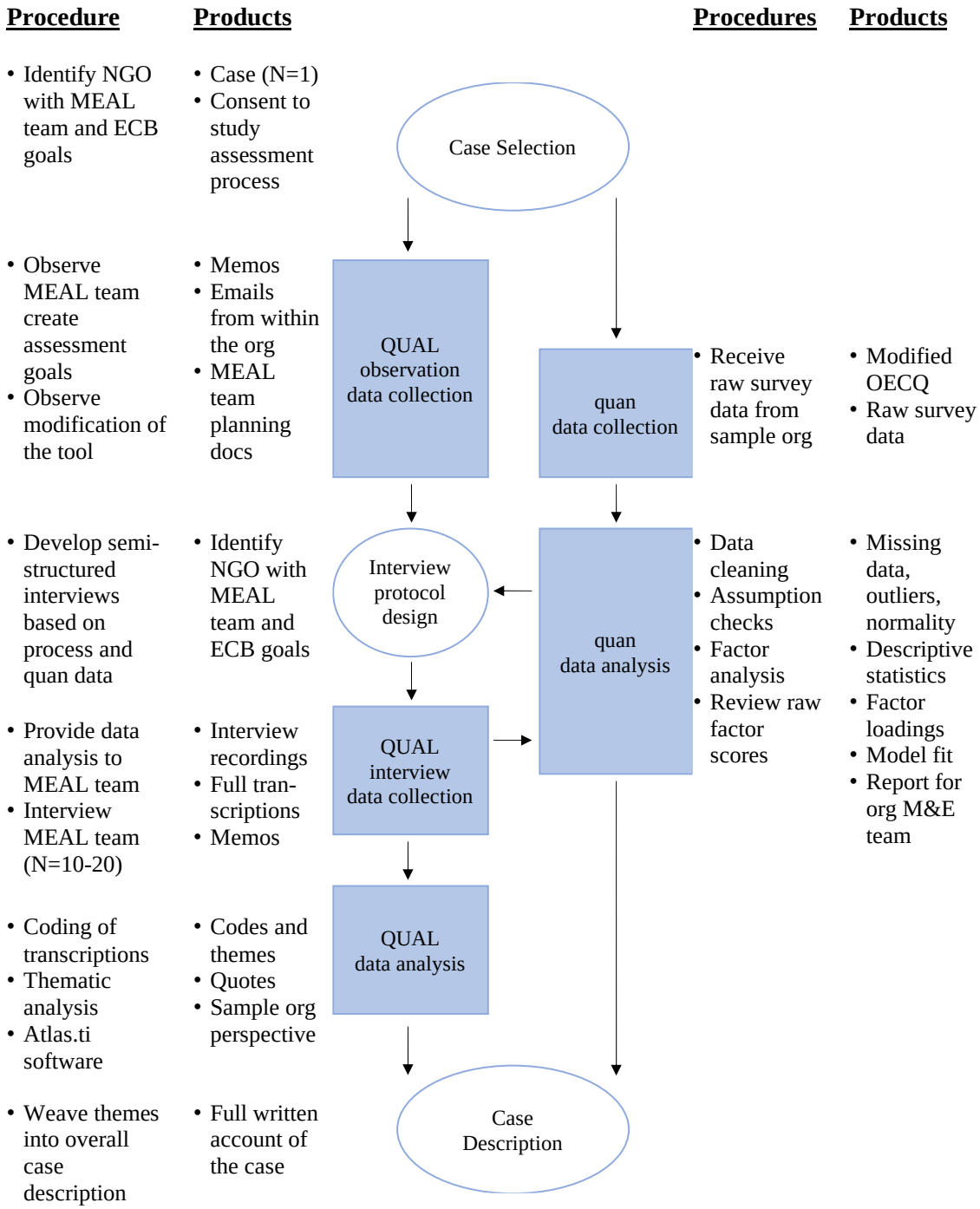
As seen in Figure 1, the process started with case identification and ended with the description and interpretation of the case using data from both methods. After agreeing to work with an organization on the study, during the case I observed meetings where the organization made choices for implementation of the survey, like instrument contextualization and survey sample composition, but attempted to minimize my impact on their actions to avoid biasing the organizations experience and choices. I was able to observe through silent attendance in Zoom meetings.

Once the organization completed the survey process and shared the raw data with me, I analyzed the survey results, including a factor analysis. Due to the statistical capacity of the organization, I provided the organization with a summary of the results to include data visualizations of each item and factor. As the organization reviewed the results, I used the initial quantitative results to inform my interview protocol. The

interview protocol also included questions to probe the decisions the organization made in the survey implementation process and their specific goals of the assessment. Data collected from the interviews then supported further quantitative data interpretation, creating a second integration of the methods, and subsequently both analyses shaped the final case description.

Figure 3.2

Concurrent Mixed Methods Case Study Procedural Diagram



Sample

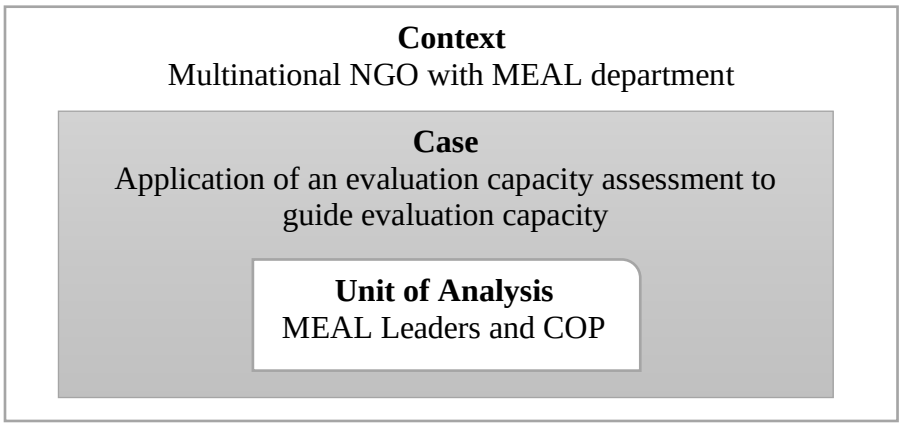
I utilized purposeful sampling in the concurrent mixed methods instrumental single case study, a nonprobability approach to selecting participants who can best address the research questions (Creswell, 2007). In this case, the purposeful criteria were a multinational NGO with a monitoring, evaluation, accountability, and learning (MEAL) department. MEAL leaders or a MEAL community of practice (COP) acted as the unit of analysis within the organization (see Figure 2). The sample size was one organization, but data collection included completed survey assessments from over 200 staff and 10 semi-structured interviews that included 25 staff.

It is critical to appropriately bound the case to inform data collection and analysis, including specifics identifying the participants and their location, the process, and the timeframe for investigating the case (Yin, 2014). Mills, Durepos, and Wiebe (2010) propose three ways of conceptualizing the boundaries of the case: commonsense, theoretical, and methodological. I state the commonsense boundaries in the preceding paragraph, bounding the case within a specific organization with a focus on MEAL staff as the unit of analysis. For the theoretical bounding, the study utilized a model of evaluation capacity that uses antecedent dimensions like organizational learning and staff support structures, that relate to a broader set of practices beyond executing evaluations. Lastly, with respect to methodological boundaries, the limited time frame for the study has implications on the case description. For example, although I will attempt to probe how the organization will use the results to inform future evaluation capacity building initiatives, I concluded the process before the organization could undertake those

initiatives. Accordingly, the bounds of the case study create limitations on the conclusions in my description.

Figure 3.3

Single Instrumental Case Study Sample



To recruit an organization, I leveraged my network from previous work experience in similar organizations that would qualify for the sample. I reached out to multiple organizations by email to explain my research interests, the potential value of participation to the organization, and detail the overall process. Two organizations expressed interest and I had multiple meetings with leaders from both organizations to discuss parameters of the study and answer any questions to help guide their decision. Ultimately, the goals and availability of one of the organizations proved best aligned with my research questions and I received their consent to participate. I keep the sample organization anonymous in the description, but I describe it in general terms in Chapter 5. I obtained Institutional Review Board approval from the University of Denver. The IRB granted me an expedited process as I did not plan to sample vulnerable groups and the interventions were interviews about the survey process.

Data Collection

Across the mixed methods I had four methods of data collection: observation, survey, interviews, and document collection. Yin (2014) proposes the use of multiple sources of evidence is necessary to detail a real-world context. Semi-structured interviews and observations provided the most critical data as the research questions focus on the perceptions and experience of the sample organization's staff. The sample organization contextualized the survey from research (Gagnon et al., 2018) before implementation, and the results helped shape the semi-structured interview questions. The document collection provided critical context to understand the organization, like its structure, policies, and goals.

Observation. Creswell (2007) contends observation is the act of noting a phenomenon in the field setting through the five senses. Unfortunately, due to the ongoing pandemic in 2021, I was unable to attend and observe meetings in person which restricted my ability to use all five senses. Using Zoom or similar video technology, I often observed as a nonparticipant in meetings, which Creswell (2007) defines as an "outsider of the group under study, watching and taking field notes from a distance." "Keeping a distance" was critical for the case description as I did not want to interfere and affect the implementation choices of the sample organization. I used memos as a means to keep notes and reflect on participants' responses and attitudes during observations. Additionally, the notes and memos I completed after each observing session informed the interview protocol, as I often needed the participants to describe the processes and decisions I observed.

Survey implementation. A journal article from Gagnon et al. (2018) describes the creation of the survey assessment tool and details the construct validity process through exploratory factor analysis (EFA). The assessment instrument, named the evaluation capacity in organizations questionnaire (ECOQ), includes 119 items organized by 11 factors in their model. Each item uses a Likert scale for response, from strongly disagree to strongly agree. I chose to offer the use of this survey to the sample organization due to its comprehensive and robust structure, as it responds to all the common dimensions found in the qualitative evidence synthesis of evaluation capacity assessment models (see table 3, Chapter 4).

Although the assessment tool does not reference a specific type of organization as a target, the sample organization has unique structure, interests, and limitations demanding contextualization. Research and the overall body of evaluation capacity building literature have repeatedly stressed the centrality of context in order to understand dynamic interactions, between hierarchies, systems, structures, and people, and how those interactions affect evaluation processes and use within organizations (King, 2007, 2017; Trochim, 2009; Suarez-Balcazar et al., 2010). Accordingly, I worked with the sample organization to make changes both to the length of the tool through reduction of contextually irrelevant items, as well as language changes to items to fit their context. Additionally, the organization added four demographic questions to help interpret and segment the results. However, the organization was comfortable with the factor structure and was able to modify items within the bounds of the original model. Once the organization's MEAL leaders had completed the contextualization of the tool, they

shared the survey with global program and MEAL staff and administered the survey online. The contextualization process is described in more detail in Chapter 5.

Interview protocol. I conducted 10 semi-structured interviews with 25 members of the organization with an emphasis on MEAL specialists, program support staff (program directors and technical advisors) and country offices (see Table 1). I conducted all interactions online through Zoom and recorded the conversations, with consent, for review and transcription purposes. The sample organization's MEAL leadership approved all interview participants and all participants' responses remain anonymous through the case description. Before each interview I sent the participants a document with a description of the research study, informing them of their rights, and sharing the protocols in place to keep their responses anonymous and protect all data.

Table 3.1

Interview Sample by Respondent Role

Role	Count
MEAL COP Lead	1
MEAL COP Members	11
Sr. Director of MEAL	1
MEAL Advisor	2
Learning Advisor	1
Program Directors	2
Technical Advisors	2
Country Office Staff	5
Total	25

A semi-structured interview protocol facilitated most interviews, with protocol questions finalized after the initial analysis of the survey results. I used a semi-structured process which suited my desired style to have structure to the conversation but remain

conversational. Furthermore, I desire the freedom to ask follow-up questions that deviate from the protocol based on the participants' unique perspectives and roles.

To create the interview protocol, I started with broad and general questions so that the participants could construct meaning of a situation (Creswell, 2007), and follow-ups allowed for more specific questions to ensure rich description. It is challenging to capture nonverbal expressions in online interviews; however, I completed short observational memos after each interview, reflecting on noteworthy moments of the participants' expression or response to support the transcription process (Kvale, 2008).

Document Collection. Yin (2014) proposes the use of documentation to corroborate and augment evidence from other sources is important for triangulation. The sample organization provided documents to support my understanding of the organization, how it is structured, its strategic objectives and goals, evaluation examples, as well as the scope of the MEAL department. Examples include the organization's strategic plan, staff organization charts, community of practice charters, evaluation policies, evaluation reports, and evaluation quality scores. As I engaged the sample organization, I regularly asked for documentation to support the overall case data.

Quantitative Data Analysis

I analyzed the survey data for the organization before collecting the qualitative interview data. As noted previously, I worked with the organization to contextualize the survey found in Gagnon et al. (2018) and the organization administered the survey. Once the survey closed, the organization shared the raw data with me in an excel file, with over 200 responses. My first step in the analysis was to clean and prepare the data. I reviewed

for entry errors or abnormally completed surveys, like cases with all the same entries for each item. After cleaning, 135 responses remained with 121 completing more than 95% of items (i.e., missing 4 items or less). Confident in the integrity of the remaining data, I explored the quantitative results through descriptive statistics (including mean item and factor scores), inter-item correlations, and an assessment of internal reliability using Cronbach's alpha. The sample organization was most interested in average item and factor scores, so I developed data visualizations and tables in excel for the organization to review items by each demographic (staff type, region, etc.).

Factor Analysis. Although the tool was modified for the sample organization's context, the organization attempted to keep items as similar to the originals as possible to prioritize fidelity to the model. They assumed the *a priori* theory of organizational evaluation capacity was appropriate for their context and believed testing the model fit could encourage more insight to be extracted from the original research (e.g., factor loadings between dimensions). I aimed to use confirmatory factor analysis (CFA), keeping the *a priori* factor structure proposed in the original model, to investigate if the results fit the underlying factor structure despite the reduction or modification of items by the contextualization of the tool. However, the size of the survey sample was a concern prior to attempting the confirmatory factor analysis. General recommendations have suggested a ratio of 4 or 5 respondents per variable or a minimum of 200 respondents (Floyd & Widaman, 1995), but this analysis has a ratio around 1:1 (113 completed surveys and 101 items) and about half the suggested minimum respondents. Moreover, I did not have the full existing SEM model that included any correlated error terms or residuals,

and Floyd and Widaman (1995) note lengthy questionaries can be challenging to fit due to many likely correlated errors.

Unfortunately, when I attempted the CFA, I was unable to achieve adequate fit to the *a priori* factor structure. I used multiple indices to determine appropriate model fit include chi-square, the comparative fit index (CFI), and the Root Mean Square Error of Approximation (RMSEA). Thresholds provided in Browne & Cudeck (1992) determined adequate fit. The chi-square test was significant ($x^2(4921) = 10866.4$, $p < .001$), below the threshold of .05, but is known to be highly sensitive to sample size. Therefore, I reviewed the fit under CFI and RMSEA using the following thresholds: CFI greater than 0.90 and RMSEA below 0.10. Both indices indicated poor fit, with a CFI statistic of .416 and a RMSEA statistic of .104.

The next step was to attempt an exploratory approach to describing the underlying factor structure and compare it the *a priori* model. The organization was also interested in scale reduction, for future and repeated administrations, and exploratory factor analysis (EFA) helps to remove items that do not load or meaningfully contribute to a dimension. My approach to the exploratory factor analysis used two steps, first a principal components analysis to determine the appropriate number of factors and then a common factor analysis to investigate the factor structure and relationships with individual items.

In the principal components analysis, the number of factors for the model was determined by exploring the component's eigenvalues, the scree plot, and a parallel analysis. Additionally, the existing *a priori* model also informed interpretation. The parallel analysis was critical in identifying a determination of 7 factors (data presented in

Table 4, in Chapter 5). During the common factor analysis, using the specified seven-factor model with Varimax rotation, I identified items with loadings of .30 or less, and those that loaded equally on multiple factors, for the interview protocol and presentation to the sample organization. I discuss the data and results of the exploratory factor analysis in detail in the case study results (Chapter 5).

Validity and reliability. Experts have tested the original instrument, the evaluation capacity in organizations questionnaire, for content and construct validity. Although the tool was adapted for the sample organization's context, the organization attempted to keep items as similar to the originals as possible to prioritize fidelity to the model. However, confirmatory factor analysis could not find adequate fit with the original model. Exploratory factor analysis illustrated a similar model (see Chapter 5) but the case to recreate construct validity would have been stronger had the model fit the *a priori* theory.

Threats to validity of the study include a small sample size, the composition of the sample, and contextualization of the tool. The sample consists of one organization, and a small portion of its staff. Accordingly, generalization claims about the validity of the model are not reasonable. Additionally, there is little guidance on who should respond to an evaluation capacity assessment survey, and different proportions of staff (i.e., MEAL, program, leadership, etc.) may have changed the results. Lastly, the contextual changes to the tool, while adopting the factor structure of the original research, made the comparison to the existing model more complex and difficult to interpret.

I used Cronbach's Alpha to examine the internal consistency of the items within each factor identified by the exploratory factor analysis, with coefficients between .7 and .8 indicating moderate internal consistency and coefficients at .8 and above indicating high internal consistency. Results of the analysis showed the factor with the smallest number of items, the capacity to do evaluation, had the lowest reliability coefficient of .762. The remaining factors all had coefficients above .83 indicating high internal consistency (see Table 6, in Chapter 5).

Qualitative Data Analysis

The qualitative data for analysis included all interview data, documents collected from the organization, and my memos. I undertook a six-step process as suggested by Creswell (2007) to analyze qualitative data that includes: (1) organizing and preparing the data; (2) reading and looking at the data; (3) coding; (4) utilizing the codes to describe the setting, people, or themes; (5) determining how the qualitative narrative will use the description and themes; and (6) interpreting the data set to determine findings.

To prepare the data, I reviewed each interview by first listening to the recording without transcription, and then again to transcribe verbatim. Next, I loaded the transcripts, along with the organizational documents and my memos, into Atlas.ti software for the coding process. However, before coding, I explored the data by reading each interview to memo initial reactions and create a few broad codes. After memoing, the coding started with the expectation of multiple readings of data and an iterative process of code and theme creation. Themes incorporated multiple codes, and, within each theme, I reduced the data to significant statements and quotes made by the interview

participants. Finally, I represented the findings in a discussion of themes with quotes and shared perspectives across interviewees in Chapter 5, with interpretations that respond to the research questions.

Analysis of the collected organizational documents consisted of first identifying information in the documents related to the interview data. For example, I identified evaluation policies that relate to important model factors discussed in interviews. The documents helped to triangulate critical observations or insights, supporting the codes and themes that arose from the interviews (Creswell 2007).

Trustworthiness. At the conclusion of the case, I review transferability, dependability, and confirmability to investigate the qualitative results. Threats to trustworthiness include the size of the survey sample, the selection of participants that represent the organization's diverse stakeholders, and the uniqueness of a single organization. Triangulation of data from multiple sources, as well as making participants' contradictions clear, support trustworthiness and an understanding of the limitations of the study. Beyond triangulation, to increase the trustworthiness of the results, I shared the major themes with the organization's MEAL leadership, in a final interview, as a means of checking the conclusions and making any final refinements. Furthermore, I attempted to achieve a rich and thick description of the case to provide enough detail for readers to consider the validity of the findings.

Integrating the Sub-Studies

In the conclusion of the dissertation, I integrate the findings from the qualitative evidence synthesis of organizational evaluation capacity literature in sub-study 1, with

the quantitative survey analysis and qualitative case data from applying an organizational evaluation capacity assessment in sub-study 2, to explicate a framework for the applied use of evaluation capacity assessments. To generate a framework for the applied use of organizational evaluation capacity assessments I utilized an inductive, exploratory research approach.

Inductive Research

Inductive research approaches attempt to generate theory from data (Eisenhardt, Graebner, & Sonenshein, 2016). Although not confined to qualitative research exclusively, the most well-known inductive methods are grounded theory, ethnography, and theory building from case studies. Inductive research is "bottom-up" and exploratory, but remains directed by empirical data (Woo, O'Boyle, & Spector, 2017). Inductive methods are appropriate when little is known about a phenomenon (Eisenhardt, 1989), are particularly problem-focused and attempt to better understand what happens and why (Woo, O'Boyle, & Spector, 2017), and excel at explicating processes and related "how" research questions (Eisenhardt, Graebner, & Sonenshein, 2016).

Eisenhardt, Graebner, and Sonenshein (2016) suggest inductive research approaches share three primary commonalities: (1) deep immersion in the focal phenomena with openness to many types of rich data, (2) the use of theoretical sampling to select cases to illuminate relationships among constructs or develop deeper understanding of processes, and (3) reliance on grounded theory-building processes, although not the exact steps of the traditional grounded theory method. The authors suggest the approach uses memoing during data gathering process to support theory

generation. Common analytic methods include developing thick descriptions, coding raw data for thematic analysis, the use of constant comparison between emergent theory and data, and engagement with the topic's literature to refine the theoretical product. The final product of building theory from inductive methods may be concepts, a conceptual framework, or proposition (Eisenhardt, 1989).

Accordingly, this study is well suited to conclude with an inductive approach to offer future practitioners guidance on how to implement an evaluation capacity assessment. Currently, a framework or approach to apply organizational evaluation assessments is absent in the literature, demanding new "problem-focused" theory on the "how" of assessment implementation. The integration of the methods and data from sub-studies 1 and 2 support the data collection and analysis commonalities proposed by Eisenhardt, Graebner, and Sonenshein (2016): the case study allows for deep immersion with the phenomena; the aggregate data is diverse to include my evidence-synthesis, quantitative survey data, observations, and interviews; and the case process allowed me to memo to support the theory development process.

Theory Building from Case Study

To guide my inductive research process, I borrowed from Eisenhardt's methodological work on "theory building from case study." Although I integrate findings from the qualitative evidence synthesis of organizational evaluation capacity literature in sub-study 1, with the quantitative survey analysis and qualitative case data from applying an organizational evaluation capacity assessment in sub-study 2, the case study process provided the most critical data to build the resulting theory. Eisenhardt

states that one of the reasons case study is a strong method for theory building is the validity of the theory is supported with “intimate” evidence and empirical observation. She states (1989) “this intimate interaction with actual evidence often produces theory which closely mirrors reality.” I found this to be true for my study and leaned heavily on “intimate” evidence from my case study to build a theory that, hopefully, “mirrors reality” for other organizations.

A major component of theory building from case study is the overlap of data analysis with data collection (Eisenhardt, 1989). As a primary example, she suggests taking field notes and memoing when impressions occur, without too much critical analysis, because you cannot know what will end up being useful. To that end, Eisenhardt suggests researchers ask, "What am I learning?" as they take field notes and memo. During the case study described in sub-study 2, I used memos to record my impressions about the efficacy and success of the choices the organization made in the process of implementation.

Another key feature of theory building from case study method is the comparison of the emergent ideas and theory with the relevant literature. Similar to using “intimate” data to support validity of the theory, Eisenhardt (1989) suggests tying the emergent theory to existing literature enhances the internal validity and generalizability of the end product. To mitigate individual bias, she states cross-case analysis helps the researcher look at the data in new ways. For example, conflicting evidence forces the researcher to reconcile the differences through deeper analysis or reconsider if a discovered pattern is spurious. For the conclusion, I did not have similar cases to compare, however, I was

able to use the qualitative evidence synthesis to return to the creation and/or testing of the models found in research, to improve the accuracy and refine my emerging theory. Accordingly, the theory building from case study method helped define the integration of the data from each sub-study, as formal integration of the data happened as I coded across the multiple data sets to find common concepts and perform the "cross-case" analysis (Creswell & Plano Clark, 2017)

Eisenhardt (1989) proposes a list of 8 steps in her research method. However, I was integrating and concluding two sub-studies, not undertaking a new case study for this research. The only new data not used in the previous sub-studies were memos recording my own impressions of the implementation process. Accordingly, I only adopted the data analysis and theory development portions of the theory building from case study method.

I used five nonlinear steps for the process to generate a framework for the applied use of organizational evaluation capacity assessments. As noted, the first step took place during the case study, as I wrote memos of my initial impressions from the implementation process. Upon the completion of sub-study 2 and the case description, for the second step, I consolidated, inventoried, and described the data from the multiple methods in study 1 and 2. In the third step, I identified any *a priori* constructs I brought to the formal data analysis process and begin coding and formulating a formal theory. At the latter stages of the coding process, I began to develop a "construct table" connecting key constructs in the generated theory with the data, as suggested by Creswell and Plan Clark (2017). The rough construct table slowly evolved into the final application

framework presented in the conclusion (Figure 13). In the fourth step, with a developing theory, I returned to the literature, relying on the data from qualitative evidence synthesis, to attempt a "cross-case" pattern analysis and looked for divergent evidence. I used the models of organizational evaluation capacity, their hypothesis testing, and the case studies supporting their creation, as checks on the developing theory. Lastly, at the point of saturation, when the theory was no longer progressing, I completed a narration of the theory in Chapter 6 (Conclusion).

Chapter Four: Results of a Qualitative Evidence Synthesis of Theories of Organizational Evaluation Capacity

In this qualitative evidence synthesis of organizational evaluation capacity models I will attempt to: (1) synthesize the extent of research theorizing organizational evaluation capacity models; (2) detail dimension commonality across EC models; (3) examine the extent the models have undergone tests of validity; and (4) identify possibilities for future research to expand the evidence base. In Chapter 3 (Methods), I detailed the method, eligibility criteria, and search strategy to identify all empirical research using explicit, systematic methods, to develop a comprehensive and representative sample and extract reliable findings.

Ranging from 2002-2021, my systematic review of the literature collected 16 articles and assessment tools representing the evolution and diversity of organizational evaluation capacity models. I describe the body of research by grouping it into three different periods, categorized by the types of models developed: conceptual models, structural models, and quantitatively developed models (see Table 2). Conceptual models include descriptions of evaluation capacity with at least two dimensions, structural models begin to relate the dimensions to one another, and the quantitively developed models use surveys and factor analysis to refine their structural model. Much of the research builds upon previous publications; therefore, I order the groups, and the

results within the groups, chronologically. Finally, I analyze the dimensions across the models, as well as the extent of validity examination for the quantitatively developed models.

Table 4.1

Articles Organized by Periods of Model Type

Conceptual Models	Structural Models	Quantitively Developed Models
Stufflebeam (2002)	Bourgeois & Cousins (2008)	Nielsen, Lemire, & Skov (2011)
McDonald, Rogers, & Kefford (2003)	Taylor-Powell & Boyd (2008)	Taylor-Ritzler, Suarez-Balcazar, Garcia-Iriarte, Henry, & Balcazar (2013)
Cousins, Goh, Clark, & Lee (2004)	Preskill & Boyle (2008)	Cousins, Goh, Elliot, & Burgeois (2014)
King & Volvok (2005)	Readiness for Organizational Learning and Evaluation Instrument (2011)	Bourgeois, I., Whynot, & Thériault (2015)
Naccarella, Pirkis, Kohn, Morley, Burgess, & Blashki (2007)		Cheng & King (2017)
Volkov & King (2007)		Gagnon, Aubry, Cousins, Goh, & Elliot (2018)

Assessment of Representativeness of the Sample

Although I have attempted to create an explicit, systematic process for gathering relevant models and tools, bias in the sample could still exist. I believe the primary risk to having adequate representativeness in the sample is missing data from practitioners. As I note in Chapter 3 (Methods), the inclusion of grey literature in this review is important given evaluation capacity building is a common practice yet academic literature on organizational evaluation capacity measurement is minimal. Of the twenty articles included in the review, 13 are from peer-reviewed journals and it is likely there are other operating models used by evaluation capacity building experts, consultants, and

internal evaluators. Additionally, due to time constraints, I did not exhaust program evaluation books that may include models of organizational evaluation capacity, although they would have likely shown up in the reverse citation search process. Finally, the samples used in the majority of the studies are problematic for generalization. The sampling method for most of the case studies is purposeful sampling, and convenience sampling within a set of shared entities (e.g., multiple departments residing in the same government office or nonprofits in an association) for the larger survey studies. Accordingly, the representativeness of the models may have a bias towards the type of organizations willing to participate in the research and therefore more experience in evaluation practice.

Conceptual Models

In an article on evaluation capacity building in public sector organizations, McDonald, Rogers, and Kefford (2003) established the notion of supply and demand influencing evaluation capacity. The authors state, "Many efforts at building evaluation capability have focused primarily or even exclusively on supply – on documenting and developing the skills, tools and resources that are available to produce evaluations." The authors encouragement to include dimensions of evaluation demand manifested in a few future models, and implicitly seen in almost every model included in the evidence-synthesis.

In 2004, Cousins, Goh, Clark, and Lee wrote an article describing and critiquing the current state of the knowledge base concerning integrating evaluation into organizational culture. The article sets the stage for Cousins' work with Bourgeois in

2008 and 2013, and Gagnon and Cousin's 2018 model, establishing sequentially more detailed models. The body of work creates a conceptual framework of "evaluative inquiry as an organizational learning system." The authors suggest several forces influence the integration of evaluation into organizational culture: facilitative leadership and modeling; ongoing training and technical support; the existence of prior knowledge, skill, and facility with evaluation logic; the availability of resources for evaluation; and exigencies for evidence about program and organizational performance and results. These influences surface in many of the future models of organizational evaluation capacity.

Included in this foundational period are two assessment tools commonly cited as elements of future organizational evaluation capacity models: Stufflebeam's (2002) institutionalizing evaluation checklist, and Volkov and King's (2007) checklist for building organizational evaluation capacity. I could not find supporting documents describing the creation of Stufflebeam's checklist, but in Stufflebeam and Shinkfield's (2007) book "Evaluation Theory, Models and Applications" a short description of a similar checklist called Institutionalizing and Mainstreaming Evaluation (p. 676-687) exists. He describes this 15-point checklist as designed to guide organizations through a process of planning and installing a sound organizational evaluation system. The authors do not break the 18-point checklist into dimensions but includes many similar components to other theories in the evidence-synthesis: external forces, staff structure, stakeholder participation, standards, policies, evaluation models, resources, systems, and communication channels.

Volkov and King's (2007) developed a checklist from their 2005 article examining the development and overall status of program evaluation in three Minnesota nonprofit organizations. Using the data the authors collected in their evaluation, they develop a conceptual framework for understanding and developing evaluation capacity building. The framework consists of three major categories: organizational context, evaluation capacity building structures, and resources. The evaluation capacity building structure category has the most relevant correlation to an evaluation capacity assessment as it includes the following dimensions: purposeful evaluation capacity building plan for the organization, infrastructure to support the evaluation process, purposeful socialization into the evaluation process, and peer learning structures. Volkov and King developed a checklist out of this structure, outlining components of each dimension, called "A Checklist for Building Organizational Evaluation Capacity."

Lastly, Naccarella et al. (2007) summarized the literature on evaluation capacity at that point and suggested the lack of definition for evaluation capacity was rooted in the varying conceptualizations of evaluation and evaluation capacity building. The authors state most definitions contain dimensions about equipping staff, organizational culture, resources, developmental stage, and evaluation use. The article then uses a case study of evaluation capacity building in Australia's federal mental healthcare programming to conclude that definitions of evaluation capacity building should address evaluation use, including process use, a major dimension of future models.

The four journal articles and two assessment tools cited during this period did not create clear frameworks or models of evaluation capacity. However, the items or

groupings in their checklists and case studies are foundational elements future studies utilize. For example, the notion of including the demand for evaluation, from McDonald et al. (2003), is found in the first structural model of evaluation capacity (Nielsen, Lemire, & Skov, 2011).

Structural Models

Building off the dimensions found in Cousins et al. (2004), as well as Bourgeois' dissertation, Bourgeois and Cousins (2008) create one of the first explicit models of organizational evaluation capacity. Assessing evaluation capacity in four government organizations in Canada, they create a framework with six dimensions across two factors: the capacity to do evaluation, and the capacity to use evaluation. The three "capacity to do evaluation" dimensions are human resources, organizational resources, and evaluation planning and activities. The three "capacity to use evaluation" dimensions are evaluation literacy, organizational decision-making, and learning benefits. Uniquely, the framework creates four levels of evaluation capacity, with each dimension having a description in each state: low, developing, intermediate, or exemplary. Similarly, they develop stages of evaluation capacity building: traditional evaluation, awareness and experimentation, implementing evaluative inquiry, and adoption of evaluations as a management function. The article examines the four stages of evaluation capacity building in relation to the four levels of evaluation capacity, and draw linkages between the two, a contribution not seen in other models. The analysis leads the authors to stress that organizations with high capacity still have low scoring dimensions, and organizations with less capacity have high scoring dimensions. In an article published five years later, titled "Understanding

dimensions of organizational evaluation capacity," they further build out the model with descriptive levels of competency for each level of evaluation capacity (Bourgeois & Cousins, 2013).

Expanding on the checklist from Volkov and King (2007), Taylor-Powell and Boyd (2008) develop a three-factor framework for evaluation capacity building in a complex organizational setting. The three factors are: professional development, resources and supports, and organizational environment. Each factor has 5-7 elements, and the authors map each onto other models of evaluation capacity. The article then uses a case study about logic model training to demonstrate the interconnections of the three dimensions. An important recommendation they make is to understand evaluation capacity building as organizational development, not just professional development, thinking about the individual, team, and organization simultaneously.

In 2008, Preskill and Boyle present a model of evaluation capacity building that contains elements of sustainable evaluation practice. They claim few comprehensive conceptual frameworks or models exist to guide practitioners' evaluation capacity building efforts and empirically test the effectiveness of evaluation capacity building processes, activities, and outcomes. The model draws on the fields of evaluation, organizational learning and change, and adult learning to create the model. The model has two circles: the left side of the model represents the initiation, planning, designing, and implementation of an evaluation capacity building intervention. The right circle is the most salient for assessing evaluation capacity. It represents the processes, practices, policies, and resources they believe sustainable evaluation practice requires. They

include evaluation policies and procedure, evaluation frameworks and processes, resources dedicated to evaluation, use of evaluation findings, share evaluation beliefs, integrated knowledge management evaluation system, strategic plan for evaluation, and continues learning about evaluation. The eight components map to the dimensions and sub-dimensions in Bourgeois and Cousins (2008) framework.

Mentioned in the Preskill and Boyle (2008) article is an assessment tool called the Readiness for Organizational Learning and Evaluation (ROLE) instrument. The instrument is more geared towards organizational learning than evaluation capacity, but future models commonly cite it as a major influence on their work. The ROLE consists of questions grouped into six factors: culture, leadership, systems and structures, communication, teams, and evaluation. The instrument states it can identify the existence of learning organization characteristics; diagnose interest in conducting evaluation that facilitates organizational learning; identify areas of strength to leverage evaluative inquiry processes; and identify areas in need of organizational change and development. Although the instrument is not explicitly about measuring evaluation capacity, its factors have significant overlap of future models of evaluation capacity.

The three journal articles and ROLE assessment tool cited during this period made major theoretical progress by illustrating comprehensive models of organizational evaluation capacity. Not only was the model presented by Bourgeois and Cousins in 2008 one of the first, but it remains one of the few to connect levels of evaluation capacity and stages of evaluation capacity building. Taylor-Powell and Boyd (2008) were one of the first to test their model and begin building interconnections, setting the

stage for future quantitative structural methods. And Preskill and Boyle's Multidisciplinary Model of Evaluation Capacity Building remains a seminal work in the overall field of evaluation capacity building.

Quantitatively Developed Models

Nielsen, Lemire, and Skov (2011) produced the first quantitative model of evaluation capacity using factor analysis to test the model's construct validity. The authors rightly suggested the empirical bases of the models to date differ, and noted the contributions are mostly case studies. Using organizational theory as its foundation, they built a model using two factors (demand and supply) and four dimensions (objectives, structure and process, technology, and human capital). Additionally, each dimension has 2-4 subcomponents. The authors created a quantitative survey with point totals for every subcomponent and had 287 Danish government evaluation practitioners complete the survey. The study used confirmatory factor analysis to demonstrate construct validity. The supply side had lower factor loadings and variance explained, but adequate overall model fit. The authors suggested the lack of dimensions on context; the role of culture; the role of leadership; the role of an evaluation champion; and the role of knowledge management to be the most notable omissions in the model.

Following the Nielsen et al. study, Taylor-Ritzler, Suarez-Balcazar, Garcia-Iriarte, Henry, & Balcazar (2013) also state the need for more empirical research on the factors that compromise evaluation capacity, as well as their relationships to one another. The article uses a theory created from a review of published conceptual models, evaluation capacity building principles, and factors believed to sustain the practice of evaluation in

nonprofits. The model has two factors, individual and organizational, used to predict evaluation capacity outcomes, broken into mainstreaming and use. No theories to this point have empirically assessed how individual and organizational factors relate or have used outcomes in a structural model. Dimensions of note included: process use, organizational support systems (e.g., incentives and rewards), internal pressures from program participants and staff, and external pressures from funders and accreditation requirements. Additionally, cultural factors related to the organization, the program, and the participants the organization serves. The instrument created 56 items and modified 12 more from other instruments, with 169 organizations completing the survey. Confirmatory factor analysis and structural equation modeling demonstrated adequate model fit.

In 2014, Cousins and Bourgeois, along with Goh and Elliot, substantially modified their model of organizational evaluation capacity used in 2004, 2008, and 2013, to quantitatively validate the factors (Cousins et al., 2014). In this version of the model, the factors organize into antecedent conditions affecting capacity, evaluation capacity process, and capacity consequences, creating outcomes similar to the work of Taylor-Ritzler et al. (2013). The first antecedent is sources of knowledge skills and abilities, and the second is organizational support structures. The four dimensions of evaluation capacity and processes are: capacity to do evaluation, evaluative inquiry, capacity to use evaluation, and mediating conditions. The evaluation consequences are organizational learning capacity and organizational consequences. The authors state a desire to inform future research as the model had not yet tested for validity, however the authors caution

disregarding qualitative inquiry. In Gagnon, Aubry, Cousins, Goh, & Elliott (2018), the authors test a similar model for construct validity. In the journal chapters proceeding this article, eights case studies of organizations inform the model's advancement. The organizations represent a range of sectors, (education and human resource development, community mental health and health, and societal and international development) and the authors believed their model generalized across the diversity. Their findings suggested three elements emerged as important considerations about organizational evaluation capacity: (1) administrative commitment and senior-level leadership, (2) organizational propensity to learn, and (3) the nature of evaluation expertise within the organization.

Bourgeois followed up her previous work alongside two other researchers (Bourgeois, I., Whynot, & Thériault, 2015) with a study focused on measuring evaluation capacity in different organizations, in order to identify transferable lessons to apply in diverse contexts. Bourgeois builds on the 2013 six-dimension model, from her work with Cousins, studying four different organizational contexts (non-profit, local government, and two federal agencies). The key takeaways stated by the authors include: (1) elements of organizational capacity vary considerably between different organizations, and do not follow a specific implementation pattern, but rather depend on each organization's unique characteristics; (2) capacity and institutionalization appear to go together; and (3) even though each organization differs from the others, some lessons learned transfer across the organizational types. As an example of the third takeaway, the existence of performance data, or lack thereof, can significantly affect evaluation capacity, or the leaders and

champions of evaluation within all three organizations played a critical role in ensuring adequate resourcing for evaluation activities and in promoting evaluation use.

Four years later, Gagnon, Aubry, Cousins, Goh, & Elliott (2018) published an article aiming to validate the model in Cousins et al. (2014). Using an instrument called the Evaluation Capacity in Organizations Questionnaire (ECOQ), 340 internal evaluators and organization members with oversight responsibility for contracted evaluation complete the questionnaire. Exploratory factor analysis and path analysis found a parsimonious model of 8 factors from 119 items. The model slightly evolved from the 2014 version, with the ongoing exploratory factor analysis creating a better fitting model. Model fit was adequate but not considered strong by the researchers. The model represents the most complex set of factors and relationships tested for construct validity to date, although the Taylor-Ritzler et al. (2013) model had stronger omnibus fit.

The final article, from Cheng and King (2017), studies evaluation capacity in a non-western context, in Taiwanese public schools. To examine Taiwanese schools the study uses the Delphi technique: consensus of opinions among a panel of experts through sequential questionnaires targeting a certain issue. The authors create the initial model through group interviews with eight experts and then a mail a survey to the larger expert sample, with two more follow-up surveys. The quantitative criteria accompanied by qualitative feedback dictated which items to retain, remove, modify, or add. The final version of the evaluation capacity model had three factors: evaluation culture, evaluation infrastructure, and human resources, each with 5-6 items. The items in their model significantly correspond with dimensions from many of the models in this evidence-

synthesis. The article provides evidence that models of evaluation capacity may have a moderate level transferability across cultures and sectors.

Dimensions

In total, considering similar models used in multiple studies, I identify seven distinct models proposed in the literature (see Table 3). The seven models are all from the structural and quantitatively developed model groups and share common dimensions, but differ in the organization of the dimensions. Many of the organizing factors in the models are binary. For example, inspired by the work of McDonald et al. (2003), Nielsen et al. (2011) uses evaluation demand and supply to organize their dimensions. The prolific work of Bourgeois and Cousins (2004; 2008; 2013), Bourgeois' own study (2015), and Gagnon et al. (2018) also use two factors: the capacity to do evaluation and the capacity to use evaluation. Taylor-Ritzler et al. (2013) uniquely created a model that uses individual and organizational factors to organize their dimensions. Preskill and Boyle do not provide a set of common factors to organize their 8 dimensions of sustainable evaluation practices. Taylor-Powell and Boyd (2008) organize their dimensions under professional development, resources and support, and organizational environment. Cheng and King (2017) organize their dimensions under three factors: evaluation culture, evaluation infrastructure, and human resources

Accordingly, differences in the number of high-level factors of organizational evaluation capacity exist, but the literature clearly demonstrates it is a multidimensional construct with meaningful consensus around the dimensions used. I organize the

elements of the seven distinct models of organizational evaluation capacity into nine common dimensions. The nine dimensions, with select indicators of their capacity, are:

1. Organizational culture and policies – the overall demand for evaluation in the organization, shared evaluation beliefs and commitment, and embedded policies for evaluation in programming.
2. Infrastructure and systems – knowledge management systems, data collection tools, software for data analysis.
3. Leadership – management processes to approve and encourage evaluation, decision support to evaluation teams, evaluation champions at high levels of the organization.
4. Organizational Resources – budget allotted to program evaluation, specialized departments to support program teams, number of evaluation staff members.
5. Human Resources – ongoing training available to evaluation team, career progression processes, leaning plans for staff members, capacity building in the field.
6. Capacity to do evaluation – overall quality of planning, implementing, and interpreting evaluations. Formal education and experience of staff. Ability to manage external consultants.
7. Capacity to use evaluation – evidence of program improvement based on evaluation findings, and evidence of organizational process improvement based on evaluation findings.

8. Communications - published and disseminated evaluation reports, communities of practice sharing evaluation findings, learning events (e.g., conferences or meetings).
9. Organizational Learning – continuous demand for data to improve, awareness of previous evaluations, organized resource depositories, leaning plans for programs or teams, non-program staff involvement in evaluation.

In Table 3, I map the seven models to identify which of the nine dimensions each model includes. A checkmark indicates a model includes the dimension verbatim; if the model includes a similar dimension but uses a slightly different term, the table provides the different term; terms italicized and in parentheses indicate a subdimension or item nested in another dimension of the model.

Table 4.2

Dimensions Across Models of Organizational EC

Models	Dimensions								
	Org. Culture and Policies	Infra-structure or Systems	Leadership	Org. Resources	Human Resources	Capacity to Do Evaluation	Capacity to Use Evaluation	Commun-ications	Org. Learning
Bourgeois & Cousins (2008, 2013); Bourgeois et al. (2015)		(*in Org. resources*)	Org. Decision Making	✓	✓	Planning and activities	Evaluation literacy	(*in HR*)	✓
Taylor-Powell & Boyd (2008)	✓	*(in org environment*)	(*in org environment*)	✓	Prof. development	(*in org. resources*)			
Preskill & Boyle (2008); ROLE (2011)	Shared beliefs and commitment	✓	✓	✓		Strategic Plan for Eval	✓	✓	✓
Nielsen et al. (2011)	Formalization	Technology		Structure and Processes		Human Capital	Utilization		
Taylor-Ritzler et al. (2013)	Awareness and motivation	(*in resources*)	✓	✓		Competence	✓	(*in eval use*)	✓
Cousins et al. (2014); Gagnon et al. (2018)	Org. Support: formalization	✓	*(in org learning)*	(*in capacity to do eval*)	Org. Support: training	✓	✓	(*in eval use*)	✓
Cheng & King (2017)	(*in eval culture*)	Evaluation Infra-structure	✓	(*in eval infra-structure*)	✓	Evaluation Culture and HR	(*in eval culture*)	(*in leadership*)	(*in eval culture*)

Table 3 demonstrates meaningful model commonality through the nine dimensions. Nielsen et al. (2011) fit the least based on dimension names, as it has only five of the nine common dimensions, but shares more commonality based on its survey items. Taylor-Powell and Boyd (2008) do not include three dimensions that have some common overlap around the use of evaluation (the capacity to use evaluation, communications, and organizational learning). The other five models have only one missing dimension. The only dimension missing in at least 3 models is human resources.

Validity

The early models and frameworks of organizational evaluation capacity involved conceptual models developed through expert experience. Case studies helped make the case for face and content validity, but statistical analysis and relational structure remained untested. Nielsen et al. (2011) was the first model to discuss construct validity. Nielsen et al. (2011) used factor analysis to test the demand and supply factor model proposed, to determine if their dimensions aligned with the correct factor. Through model fit indices (chi-square, the root means square residual, and the goodness-of-fit index) they concluded their model a plausible approach for measuring evaluation capacity.

Two other models use factor analysis to assess construct validity: Taylor-Ritzler et al. (2013) and Gagnon et al. (2018). Exploring a step further beyond Nielsen et al. (2011), both studies demonstrate relational structure using factor analysis and structural equation modeling. Moreover, both models have outcome measures in their structural model. Taylor-Ritzler et al. (2013) present fit for all four of their confirmatory factor analysis models, demonstrating adequate fit for three. However, their structural model

combining all four-factor demonstrated very strong fit with a RMSEA of .049 and CFI value of .990 (p. 197). The model proposed by Gagnon et al. (2018) demonstrated adequate overall model fit with a CFI value of 0.94 and a RMSEA value of .10, and SRMR of .05. The authors claimed these results suggested "somewhat good model fit."

Accordingly, there are only three models to date that have undertaken construct validity assessments, and none appear to have been replicated in the published literature. Additionally, the organizing factors of each are different: Nielsen et al. used a demand/supply model, Taylor-Ritzler (2013) uses an individual/organizational model, and Gagnon et al. (2018) expands on the capacity to do evaluation/capacity to use evaluation model. Each model could benefit from a replication with new samples. To respond to that research gap, Chapter 5 details a mixed methods case study implementing a modified version of the survey in Gagnon et al. (2018).

Further, in an article about surveying organizations, Fierro and Christie (2016) demonstrate that who provides the data to assess organizational evaluation capacity can lead to varying responses. In a paired study, the authors found evaluation practitioners gave less favorable ratings of evaluation capacity and practice than their managerial counterparts. Their findings suggest that methods used to assess organizational evaluation capacity and practice should triangulate responses from multiple individuals within an organization. They also suggest evaluation capacity assessments could benefit from mixed methods analysis to bridge survey findings with direct observations of program activities or document reviews, allowing for the corroboration or expansion of findings. None of the models in this qualitative evidence synthesis address the question

of variability in responses depending on the respondent from each organization and represents a threat to validity throughout the body of literature. In response to the gap, the subsequent mixed methods case study in Chapter 5 prioritizes a diverse sample of roles in the organization to facilitate comparative analysis of the survey data.

Chapter Five: Results of an Application of an Organizational Evaluation Capacity Assessment in a Multinational NGO

There are three primary research questions for this concurrent mixed-methods single instrumental case study: (1) what considerations are necessary to implement an evaluation capacity assessment? (2) How do the evaluation experts in the organization interpret the results? And (3), how does the organization use the results to make decisions about investing in evaluation capacity building initiatives?

The answer each of those research questions I need to start with a description of the organization and share some impactful characteristics, for example, an overview of their monitoring, evaluation, accountability, and learning (MEAL) practice and their funding profile. Next, to answer research question 1, on the necessary considerations to implement an evaluation capacity assessment, I will detail the organization's goals in undertaking the assessment and their process to adapt and administer the assessment tool. To answer research question 2, concerning how the organization interpreted the results, I will provide a summary of how they interpreted the quantitative data. Lastly, to answer research question 3, on the use of the results, I will provide themes of their overall perspectives on the assessment and summarize the next steps they plan to take after completing the assessment.

Description of the Organization

The organization participating in this case study is a large, multinational NGO focused on diverse humanitarian and development goals. To keep their identity anonymous, I will describe their structure, programs, staff roles, and other possible identifiers in general terms. Additionally, I will also describe their monitoring, evaluation, accountability, and learning practice, MEAL community of practice (COP), and funding characteristics, to provide additional important context and allow the reader to draw more insight from the case study.

The organization has a centralized unit and a decentralized, multifaceted network of country offices, affiliates, and country-based partner organizations. They work in over 100 countries and have country offices in a majority of those locations. The country offices have different governance models, with some acting as independent entities and others partnering with a western or "global-north" country who acts as a strategic and fundraising partner. The country offices fit into regional groupings that have governance and staff to provide program expertise, technical assistance, and strategic support. The country offices are responsible for program implementation and report metrics reaching over 100M people per year, with over 1,000 projects.

The centralized unit of the organization supports strategy and priorities of the overall network, responsible for branding, aggregate metrics reporting, and convening of the independent network. A Program Director described the role of the centralized part of the organization as "in service" to the network, saying their influence comes from "carrots, cajoling, supporting, engaging, facilitating, and very little of what we can do is

through a stick or punishment [to hold organizations] accountable." For programs, technical staff at this level are often related to an outcome area (e.g., gender empowerment) or a technical specialty (e.g., measurement, learning, etc.). The staff in the centralized or regional level support multiple countries within their specialization and therefore have a wider perspective into the organization's work.

The organization has built monitoring, evaluation, accountability, and learning activities into the processes and systems of the organization. As the program implementors, country offices have the primary responsibility for planning, executing, and supervising evaluations. However, not every country office has dedicated MEAL staff, with some having a program specialist only commit a part of their time to MEAL responsibilities. To support staff with MEAL responsibilities, the centralized unit of the organization has developed policies, tools, templates, and instruments for standardized use across the network, and organizes capacity building opportunities to increase the capacity of the network. They also collect internal metrics from the network to report on institutional progress, communicate impact to the public, and support new funding opportunities. I discuss the evaluation policies in a subsequent section and detail the expected practices, participation, outputs, and quality the policies mandate.

Within the centralized unit sits the MEAL COP. MEAL directors and advisors at the centralized level of the organization manage the COP, bringing together regional and country level staff from around the world to share lessons learned, promote best practices, and build internal capacity. The COP manages the organization's MEAL policies, tools, and overall standardization across the network, to ensure the organization

meets its goals and remains accountable to its stakeholders. Leaders of the MEAL COP were the main points-of-contact for this research and were the lead implementers of the survey process.

As noted above, many programs and county offices rely on partner countries, often in the global north, to support fundraising efforts. The common practice is a country office develops a proposal and sends it to the partner country in the global north, who then interacts with the donor. This is relevant to the study because evaluation capacity, as defined in the models in literature (see Chapter 4), includes using evaluations for activities like funding, public outreach, and communications. For a network governance model, this means the responsibilities that demonstrate an organization's evaluation capacity are diversified across many staff and levels of the organization. Additionally, not all funders provide funds for evaluation work, especially if the donor considers the work to be humanitarian aid. One MEAL COP member suggested institutional donors provide approximately 80% of funding for the organization, and the organization has minimal unrestricted funds to support evaluations when the donor does not fund them. In practice, that lead to projects funded by different donors, within the same country, with the same support staff, receiving different level of MEAL focuses.

Goals of the Study for the Organization

The MEAL COP was interested in an assessment of the organization's evaluation capacity for two primary reasons: to use the results as an input into their evaluation policies and to comprehensively search for and identify barriers to capacity building. In respect to updating their evaluation policies, after reviewing the models of organizational

evaluation capacity and accompanying survey tools, they believed identifying low scoring factors could provide direction on what the new policies needed to further highlight. The Sr. Director of MEAL stated:

> This study will give us the foundation to say we didn't just into revamp our evaluation guidelines based on our practices around the world; we actually had a study looking at factors that are proven to be the kind of factors we should [be investigating]. We need something solid like this to latch onto before we decide on the guidelines.

The MEAL COP Lead suggested the evaluation policies needed major revisions and left out important elements, calling the policies old and based on "very classic evaluation." The MEAL COP Lead bemoaned that the policies don't require theories of change, don't talk about failure, don't discuss contribution to change, and don't address the Sustainable Development Goals. Ultimately, the MEAL COP Lead wanted the new policies to address how important evaluation is to the organization and not only be considered "one more step in the program or monitoring and evaluation cycle."

The second reason was the organization had attempted previous MEAL capacity mapping in the past, which successfully lead to the creation of more standardization of formats and evaluation management tools. However, when discussing these processes, the MEAL COP realized they were very heavily focused on evaluation activities and skills, and not the broader view of organizational evaluation capacity illustrated in the models from research literature. Accordingly, they were interested in an assessment to collect data using a model and tool comprehensive of the organization and areas outside

of direct evaluation practice. For example, the adapted model from Gagnon et al. (2018) investigates dimensions like Organizational Support as antecedents or input, and Organizational Learning as an input and outcome. One MEAL COP member suggested, "we still need to focus on hard evaluation skills like data collection tools, statistical ability, and report writing, but those only really have meaningful impact if other conditions are in place, like a culture of organizational learning." Therefore, the organization hoped an assessment would help identify the systemic barriers to expanding evaluation capacity in the organization even beyond the hard skills to do evaluation.

The biggest participation concern for the organization was asking staff to fill out a lengthy survey, especially country staff who frequently receive requests from the regional and centralized levels of the organization. One Program Director said:

> Country offices are always being asked to provide program documentation, collect new information, fill out some kind of report, or engage in activities outside of their program portfolio which they already don't have enough time for. We need them to receive some benefit from engaging with thus survey, we can't just ask or something without giving back.

In a meeting to determine if they would proceed with assessment, the MEAL COP determined the counties would have three broad benefits. First, their perspective would inform the update to the evaluation policies which effect their responsibilities; second, the assessment could lead to more resources targeted to building capacity based on their needs; and third, the process could provide enhanced consensus around what constitutes evaluation capacity. And additionally, if a country had enough respondents, they could

use the results to improve their own practices and target specific areas for local capacity building.

Evaluation Policies

Literature supports the organization's goal of using the organizational evaluation capacity assessment to improve their evaluation policies. Research on evaluation policies details the impact they have on organizational evaluation capacity and effectiveness in multiple ways. First, an evaluation policy can be a communication tool within an organization and to its stakeholders, helping to clarify beliefs and expectations about evaluation (Preskill & Boyle, 2008; Trochim, 2009). Al Hudib and Cousins (2021) underscore evaluation polices potential to influence evaluation practice, to include when to evaluate outcomes, how to evaluate, who evaluates and their roles, and to dictate resources. Additionally, Trochim (2009) suggests written evaluation policies can make evaluation a more transparent and democratic endeavor, engendering participation and dialogue.

To protect the anonymity of the organization I am unable to explicitly state the organization's policies here but can generally relate their goals and substance. The organization's current evaluation policies state they aim to promote institutional accountability, continuous learning, and transparent sharing of evaluations both internally and externally. The objectives of the policies are to help the strategic and systematic collection, documentation and dissemination of lessons learned and impact; provide opportunities for stakeholders to participate in evaluation and provide honest perceptions

and assessments of program activities; strive to be transparent with findings; and be accountable to all stakeholders.

The actual list of policies is relevant to all programs in the organization's network and has overlap with many of the main dimensions in the model of evaluation capacity used in the assessment. For example, there are multiple policies on the capacity to do evaluation and evaluative inquiry: one policy mandates baselines and endline assessments; another policy mandates assessments of progress against organizational metrics, funder metrics, and strategic plans; another details what to include in evaluation documents (key questions, data collection instruments, etc.). There is one policy about stakeholder participation creating expectations for the inclusion of project beneficiaries in evaluation. Lastly there is a policy with a commitment to using the results for improvement, as well as some additional guidance sections that speak to evaluation use.

One important and challenging aspect of the implementation of the policies, and overall control of evaluation practice, is the use of external contactors. When donors fund endline assessments or impact evaluations, the donor sometimes hires third-party contractors to aid in the objectivity of the evaluation. In these cases, the organization has the ability to imbue the evaluation with the appropriate implementation of the policies through supporting the development of the questions, tools, and methods the evaluator will use. However, this forces the organization into a small window of impact on the evaluation's quality. A technical advisor said:

> [The hired evaluator] can only be as good as the time we put into what he's doing. You have a technical advisor who is working in like 12 different countries, and

> we have to essentially make the scope of an evaluation. When you have these smaller projects that you're trying to evaluate…sometimes I rush or miss a question and nobody catches it because the consultant's job is just to take your survey tools, collect the data and analyze it and write. So, there have been numerous times where I basically regretted not putting more time in upfront.

This is meaningful to the policies as they must account for the agency and opportunities for influence their staff have when using third-party evaluators. Additionally, as seen in subsequent sections in this chapter, the mix of internal and external evaluators used across the network had a major impact on the interpretation of the assessment's survey results.

Survey Adaption and Administration Summary

Organizational context plays a critical factor in evaluation capacity building strategy selection and overall success (Preskill & Boyle, 2008; Al Hudib and Cousins, 2021). Examples include, at minimum, organizational resources available, staff roles and characteristics, current evaluation practices, and desired learning objectives and expected outcomes. In the preceding section, I discussed some of the unique variables the organization needed to consider when adapting the survey tool, like the organization's decentralized network, different roles by location in the organization, and funding characteristics. In this section I will detail how the organization managed to collaboratively modify the instrument to their context.

Selection of Instrument

When I approached the organization and their MEAL COP about participation in this case study, I provided three examples of organizational evaluation assessments. I met with COP leadership to discuss my goals as the researcher, and learn more about their goals for participation, as outlined above. I then proceeded to provide a brief introduction to evaluation capacity assessments, using my common dimension table from Chapter 4 (Table 3). We discussed definitions, key concepts, and the value of viewing evaluation capacity beyond the skills and activities to do evaluation. After the review of the models, the MEAL COP leadership hypothesized what best fit their organization. We briefly discussed the evaluation policies, mapping what overlap existed with the evaluation capacity dimensions and what was missing (similar to the content in the evaluation policies section above).

The result of the discussion was choosing Gagnon et al. (2018) as the closet fit for the organization's context. Additionally, the group was optimistic that the number of items in the research's questionnaire, originally 119, would provide robust content to contextualize to their organization. Lastly, they believed staff beyond MEAL specialists could respond to the items, allowing a wider audience, like program staff and leadership, to participate. The MEAL COP leadership believed the opinions on the organization's learning culture, use of evaluation, and evaluation's impact on the organization were critical and would offer different perspectives from MEAL staff. Research from Fierro and Christie (2016) support a diverse sample, as the authors found evaluation practitioners gave less favorable ratings of evaluation capacity than their managerial

counterparts and therefore methods and samples used to assess organizational evaluation capacity should triangulate responses from multiple individuals within an organization.

Contextualization of Instrument

The questionnaire developed by Gagnon et al. (2018) used a federal government agency in Canada as the sample but does not target a specific type of organization as the intended audience. The next step for the MEAL COP was for two of their leaders to make direct edits to the survey and share the draft with me, and then with the wider COP. Their initial goal was to reduce the length of the tool through removal of irrelevant items and to rewrite items to better fit their practice, vernacular, and culture. However, an important component of their process was to prioritize fidelity with the original model. The MEAL COP Lead said, "The reason this study was accepted [by the broader COP] was because it has some theoretical framework behind. If we modify it significantly, it just becomes an evaluation capacity assessment of a different nature." Accordingly, the editors set out to keep the factor structure and modify items within the bounds of the original model.

During the iterative process of contextualization, there were a few meaningful considerations and discussions including: the demographics to include in the survey, the unit of analysis for each respondent given the decentralization of the organization, and the use of both internal and external evaluators across the network. First, given the complexity of the organization's network, it was important to have demographics to segment the results. Although the results would describe the evaluation capacity at the organization level, the COP wanted to be able to determine differences and nuances

between regions, countries, types of network entities, and types of staff. Accordingly, required demographics for every respondent included their role, the entity or type of office they support, and their years at the organization. I was able to create regional analyses by grouping country level data.

Second, as noted in the description of the organization, not every staff member, even MEAL staff, would have experience or responsibilities aligned with all items in the survey. Moreover, country staff would not be able to speak to the capacities of the broader organization based on their scope of work. Accordingly, wording of survey items need to be clear on the unit of focus. The solution was to ask each respondent to answer questions based on their "office." In practice, this meant that staff at the centralized level of the organization would respond based on their experience across the network, but country offices would respond based on their local capacity. The assumption was the diversity of country office responses would aggregate up to an accurate picture of the organization, while also allowing staff to extract country office specific analysis. Additionally, we could segment the results by centralized unit staff and country office staff to analyze meaningful differences.

Lastly, the mix of internal and external evaluators presented a challenge in writing items. A few MEAL COP members were concerned that some items may depend on the capacity of the external evaluators. However, this issue only concerned some of the capacity to do evaluation and evaluation inquiry items. Accordingly, the MEAL COP added items like "Our office has the knowledge and skills to oversee evaluations performed by external evaluators." Additionally, they wrote items with theoretical

phrases like “Our office possesses the technical competencies to conduct evaluations” to account for the possibility that external evaluators primarily demonstrated those competencies in their context, but the office still possessed the capacity.

The final draft of the modified survey contained 101 items, and 8 factors: Organizational Learning; Organizational Support; Capacity to Do Evaluation; Evaluation Inquiry; Stakeholder Participation; Use of Evaluation Findings; Process Use; and Mediating Conditions. The survey used the same underlying factor structure as the original model and only cut down the number of items by about 10%. Ninety-three items used a 5-point scale (either strongly disagree-strongly agree; or never-infrequently-sometimes-often-always), and the other 8 items used a 3-point scale (Low-Medium-High).

The organization kept changes from the original survey to a minimum to prioritize fidelity to the original model and survey. I have listed all items from the contextualized and original survey, grouped by factor, in Appendix C, indicating if each item had “minimal or no change”, was “modified,” was “new,” or was “removed” in the contextualized survey. The category “minimal or no change” indicates the survey used the item verbatim or the term “organization” was substituted for “office.” As noted previously, the organization made the unit of focus wording changes to facilitate the responses by country office staff who would not be able to speak to the capacities of the broader organization. The category “modified” means the organization made more substantial changes to the item, usually adding in prompts to programmatic work. Lastly, the “new” category means the item was created by the organization for their survey, and

"removed" means the item was in the original survey from Gagnon et al. (2018) but not used in the contextualized survey. In the final tally, the organization used 80 items with "minimal or no change" from the original survey, they modified 10 items, 3 items were new, and they removed 11 items (see subtotals in Appendix C for breakdown by factor). The "minimal or no change," modified, and new items add up to 93 because the survey used the 8 stakeholder participation items twice, once to indicate frequency or participation and the second asking about the level of participation.

Piloting and Distribution

Before the MEAL COP administered the survey, two country staff in the MEAL COP who were not involved in its modification piloted the survey. Each staff member completed the surveys for their country offices and provided feedback on item clarity, comprehensiveness, and user experience. Both staff members provided minor item edit suggestions and approved of the overall structure. However, both staff members shared concerns for the time it took to complete the survey, citing anywhere from 20-30 minutes. The MEAL COP leadership shared the concern and its potential impact on response rate. They had expected to cut the survey down further than 101 items, but ultimately decided the 101 items remained relevant and wanted to prioritize fidelity to the original model.

Once the organization's MEAL leaders had completed the contextualization of the tool, they translated the survey into Spanish and French to encourage response across the network. The team used Qualtrics to allow online completion of the survey and used an internal listserv to share the survey with global program staff at each level of the organization. Additionally, MEAL COP leadership promoted the survey in COP calls

with program staff across the network and asked regional leadership to promote it in country offices. The survey remained open for two weeks and the organization shared the raw data with me for analysis once it had closed.

Survey Results

The description of the survey results includes a breakdown of the sample, outcomes of a factor analysis, and the organization's perceptions on item results.

Sample

The raw data provided by the organization contained almost 200 responses, but I removed 40 surveys due to respondents only completing one factor or less. A smaller number of surveys had scoring patterns or other inconsistences justifying removal from the sample. After cleaning, 135 surveys remained for analysis, with 121 surveys 95% complete (missing 4 items or less), and 113 fully complete surveys. Staff from over 50 unique country offices completed the survey, with 15 offices providing 3-5 survey responses. I list the characteristics of the respondents, organized by the role, in Table 4. The majority of the respondents were MEAL specialists or technical program staff.

Table 5.1

Survey Sample by Respondent Role

Role	Count
MEAL Specialist	57
Other	11
Program support staff (e.g., admin)	7
Senior program staff	17
Technical program staff (e.g., project manager)	43
Total	135

The MEAL COP lead was pleased with the sample size and believed it represented optimism for growth in the network:

> I was surprised [by the number of responses]. When I look at other exercises that we are doing, or other areas where we are trying to explore capacity, it was a surprise that we had that amount of participation which shows a lot of interest in the topic, that colleagues wanted more clarity in terms of evaluation capacity.

However, not all staff agreed on the meaningfulness of the sample size. One Program Director was quick to note how relatively small the sample was, stating:

> 135 respondents are not [our organization], we have thousands of staff members. It's a minute percentage. And I don't know our percentage of program staff, but it is a teeny number of people. MEAL staff are 57 [of the 135 respondent sample]. So, I think that you're going to be skewed to having capacity on evaluation.

Not all staff interpreting the sample agree with the above program director's hypothesis on skewing; some felt the opposite, that MEAL staff would be more negative to evaluation capacity compared to other program or non-program staff. However, one consensus disappointment with the sample results was not promoting, and including another category, for high-level leadership in the centralized office. The MEAL COP Lead shared that if done again, they would have made a stronger effort to receive their opinions and facilitate their involvement, which would also build their investment in the results.

Factor Analysis

As previously noted, although the tool was adapted for the sample organization's context, the organization attempted to keep items as similar to the original items as possible, to prioritize fidelity to the model. They assumed the *a priori* theory of organizational evaluation capacity was appropriate for their context and believed testing the model fit could encourage more insight to be extracted from the original research (e.g., factor loadings between dimensions). I aimed to use confirmatory factor analysis (CFA), keeping the *a priori* factor structure proposed in the original model, to investigate if the results fit the underlying factor structure despite the reduction and modification of items. However, the size of the survey sample was a concern prior to attempting the confirmatory factor analysis. General recommendations have suggested a ratio of 4 or 5 respondents per variable or a minimum of 200 respondents (Floyd & Widaman, 1995), but this analysis has a ratio around 1:1 (113 completed surveys and 101 items) and about half the suggested minimum respondents. Moreover, I did not have the full existing SEM model that included any correlated error terms or residuals, and Floyd and Widaman (1995) note lengthy questionaries can be challenging to fit due to many likely correlated errors.

Unfortunately, but as expected, I was unable to achieve adequate fit to the *a priori* factor structure using CFA. I used multiple indices to determine appropriate model fit include chi-square, the comparative fit index (CFI), and the Root Mean Square Error of Approximation (RMSEA). Thresholds provided in Browne & Cudeck (1992) determined adequate fit. The chi-square test was significant ($x^2(4921) = 10866.4$, $p < .001$), below

the threshold of .05, but is known to be highly sensitive to sample size. Therefore, I reviewed the fit under CFI and RMSEA using the following thresholds: CFI greater than 0.90 and RMSEA below 0.08. Both indices indicated poor fit, with a CFI statistic of .416 and a RMSEA statistic of .104.

The next step was to attempt an exploratory approach to describing the underlying factor structure and compare it to the *a priori* model. The organization was interested in scale reduction, for future and repeated administrations, and exploratory factor analysis (EFA) can remove items that do not load or meaningfully contribute to a dimension. My approach to the exploratory factor analysis used two steps, first a principal components analysis to determine the appropriate number of factors and then a common factor analysis to investigate the factor structure and relationships with individual items.

Hahs-Vaughn (2017) suggests there are a number of indices that should be reviewed prior to conducting a factor analysis to assess factorability, including Kaiser-Meyer-Olkin measure of sampling adequacy and Bartlett's test of sphericity. The Kaiser-Meyer-Olkin measure of sampling adequacy is an index of shared variance in the variables, ranging from 0-1, with large values indicating adequate favorability. Bartlett's test of sphericity determines if the observed correlation matrix is statistically different from an identity matrix. Bartlett's test of sphericity was significant ($x^2(5050) = 10540.8$, $p < .001$), illustrating adequate factorability. However, the Kaiser-Meyer-Olkin was .471, slightly below the recommended value of .5, again illustrating the sample size challenge given the large number of items. Noting that risk, I proceeded with the analyzing the number of factors suggested for the model, determined by exploring the

components' eigenvalues, the scree plot, and a parallel analysis. The parallel analysis was critical in identifying a determination of 7 factors (Table 5) as the survey was lengthy enough to have many components with eigenvalues over 1.0 which was the cut-off value for use of Kaiser's rule. Additionally, the existing *a priori* model informed interpretation as seven factors was similar to the *a priori* structure. The seven factors explained 48% of the variance (Table 5).

Table 5.2

Parallel Analysis

Component	Parallel Analysis		Exploratory Factor Analysis	
	Mean Eigenvalue	95% Percentile	Eigenvalue	Percentile
1	3.662	4.029	23.347	23.116
2	3.436	3.673	6.245	6.183
3	3.271	3.475	5.213	5.161
4	3.134	3.305	4.394	4.35
5	3.014	3.144	3.531	3.496
6	2.914	3.036	3.265	3.233
7	2.812	2.949	3.045	3.015
8	2.725	2.857	2.718	2.691
9	2.635	2.778	2.429	2.405
10	2.555	2.667	2.236	2.214

To determine if the factors should use an orthogonal or oblique rotation, I ran the principal components analysis again using a specified seven-factor solution and oblique rotation, using the direct oblimin method. I reviewed the Component Correlation Matrix (Table 6) to review dimension correlation. The largest correlation was between component 1 and 2 (.244) but the coefficient was low, suggesting an orthogonal method was appropriate.

Table 5.3

Component Correlation Matrix

Component	1	2	3	4	5	6	7
1	1	0.244	0.232	-0.225	0.087	0.255	-0.163
2	0.244	1	0.14	-0.153	0.081	0.269	-0.158
3	0.232	0.14	1	-0.149	0.037	0.144	-0.182
4	-0.225	-0.153	-0.149	1	0.024	-0.158	0.109
5	0.087	0.081	0.037	0.024	1	0.099	-0.046
6	0.255	0.269	0.144	-0.158	0.099	1	-0.167
7	-0.163	-0.158	-0.182	0.109	-0.046	-0.167	1

Next, I ran the common factor analysis, using the specified seven-factor model with the orthogonal Varimax rotation. I identified items with loadings of .30 or less, and those that loaded equally on multiple factors, for the interview protocol and presentation to the sample organization. I did not proceed with removal of inadequate fitting items and more versions of the analysis, as the immediate goal was not scale refinement. I have provided factor loadings for each item in Appendix E, grouped by their *a priori* factor group. Although research papers normally present the rotated structure matrix in one long table, the organization was interested in the comparison to the *a priori* structure and therefore benefited from reviewing the items in the *a priori* factor groupings.

The results of the exploratory factor analysis were encouraging to the organization. The only meaningful difference at the factor level, in comparison to the *a priori* model, was the Evaluation Findings Use factor and Process Use factor loaded together, dropping the number of factors from 8 to 7. Each factor had a few items cross-loading, and all factors had at least one item that did not adequately load (except the "combined" Use factor). The organization viewed this result as confirmation of their ability to view their evaluation

capacity within these factors, to target specific areas for capacity building, and were excited at the possibility of future scale reduction to continue to use a smaller survey to follow-up on specific factors.

Reliability Analysis. I used Cronbach's Alpha to examine the internal consistency of each factor identified by the exploratory factor analysis. I used all items in the survey as the organization wanted to review all data in their analysis. Coefficients between .7 and .8 indicate minimal internal consistency and coefficients at .8 and above indicate adequate internal consistency. Results of the analysis showed the factor with the smallest number of items, the Capacity to Do Evaluation, had the lowest reliability coefficient of .762. The other six factors all had coefficients above .83 indicating adeqaute internal consistency (see Table 7).

Table 5.4

Reliability by Factor

Factor	Cronbach's Alpha	Items
Organizational Learning	.852	16
Organizational Support	.831	9
Capacity to Do Evaluation	.760	7
Evaluative Inquiry	.878	16
Stakeholder Participation	.846	16
Evaluation Use (Findings and Process)	.960	24
Mediating Conditions	.908	13

Item Review and Interpretations

I have shared item mean scores, standard deviations, and the percent of responses per response category in Appendix D, grouped by each factor. In this description of the items results, I present the data with a visualization favored by the organization. The vast

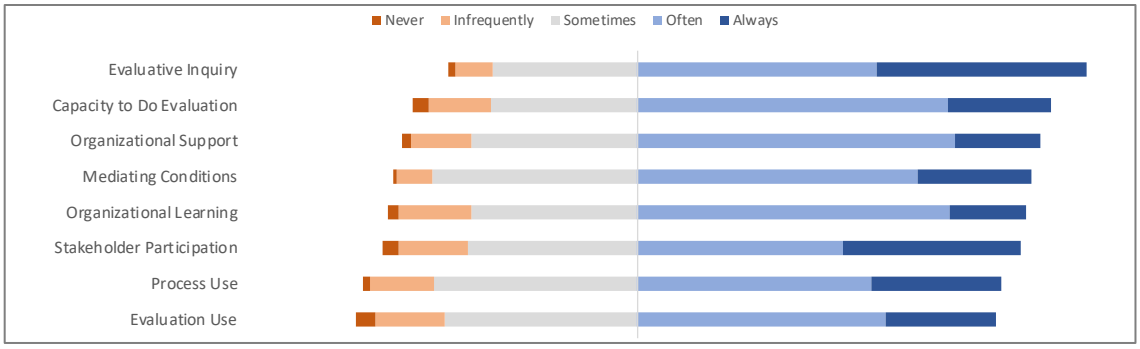
majority of the items used one of two five-point scales, and rather than assessing and comparing mean scores or standard deviations, the organization was better equipped to analyze the data using a stacked, colored-line visualization, viewed on the subsequent pages. The visualization uses two main colors, blue for the two positive categories (agree and strongly agree; or often and always), and orange for the negative scoring categories (disagree and strongly disagree; or infrequently and never). I present the middle of the scale (neutral/sometimes) in grey. I put a line down the middle of the visualization with the neutral/sometimes and negative categories to the left of the line, and the positive categories to the right of the line. I then ordered the items with the highest number of positive responses from the top. All items are positively worded, so no modifications had to be made to compare the results using this technique. The layout allowed the organization to quickly identify higher and lower scoring items from the factor and easily compare the items' range of responses compared to using means, percentages, and standard deviations.

The organization began with a review of composite scores by factor (Figure 3). I created these composite scores by averaging the items' scores within a factor, standardizing the sample size across the factors. The first takeaway from the organization was surprise at the positive skew of the results. For each factor, positive responses outweighed negative responses at a minimum ratio of 4:1. The organization expressed some disappointment at the lack of variability between the composite factor scores. Mean scores per factor ranged between 3.948 for Evaluative Inquiry as the highest, and 3.564 for Evaluation Findings Use as the lowest. However, the organization

found it compelling that, besides loading together in the factor analysis, Evaluation Findings Use and Process Use both scored as the factors with the lowest capacity scores. The MEAL COP said, “Going forward, because the results are so positive, it's hard to say we're going to try to highlight these factors in the policy because they're critical. I think we just have Evaluation Use for a high priority for the policy updates [at the factor level].”

Figure 5.1

Survey Results for the Model Factors



The organization moved to interpretation of items within factors, which demonstrated more variability in response ranges. As seen in Figure 4, the top scoring items “Staff can bring new ideas to improve programs” and “We have opportunities for self-assessment with respect to goal attainment” had negatives responses totaling less than 1% of the responses. However, the final two items “We have a system that allows us to learn successful practices from other organizations” and “Employees are given sufficient time to reflect on organizational successes and failures” had a 33% and 35% negative response rate, respectively. The Learning Advisor interviewed for the study affirmed the results, suggesting:

> I think our big weaknesses are time, so people are super busy…and therefore we learn more from ourselves and less from others. So those two most negative [items] don't surprise me. The positive ones about people able to bring ideas: I think that we're very decentralized, in most cases people have a lot of space for creativity. Some people have shy bosses and therefore can't, but in general management is largely supportive.

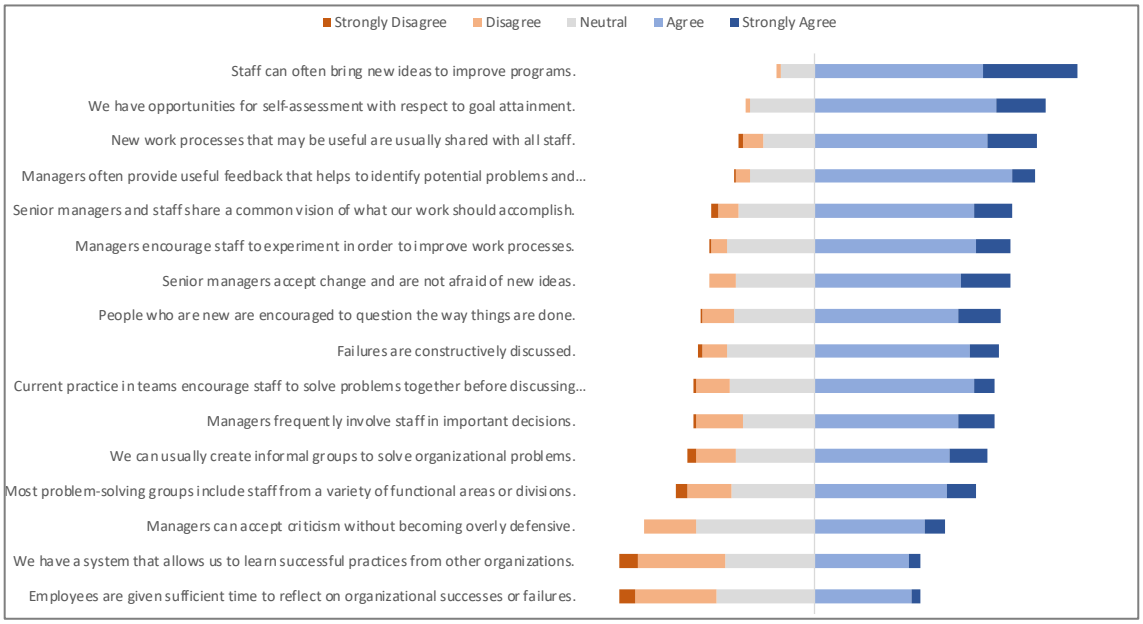
Additionally, there was a meaningful difference between the North America region and the rest of the network, with North America scoring lower on "Employees are given time to reflect on organizational success and failures." One Technical Advisor believed this to be based on roles and scope of practice, saying:

> I feel like country offices, depending on who the director is, are pretty good at quarterly stopping and saying 'what have we learned? What is our program doing? How can we make it better?' Whereas in Europe and the U.S., we don't have time…we have so many things happening all the time. Nobody stops and says, 'this is a really cool thing that country X is doing.

Two other lower scoring items were about decision-making "Mangers frequently involved staff in important decisions" and problem-solving "We can usually create informal groups to solve organizational problems." The organization also validated those scores with one MEAL COP member sharing: "We're pretty siloed. Engaging across divisional boundaries is not common and the problems are often systemic, so it makes sense people may feel left out of decision-making or don't see their own agency to make change."

Figure 5.2

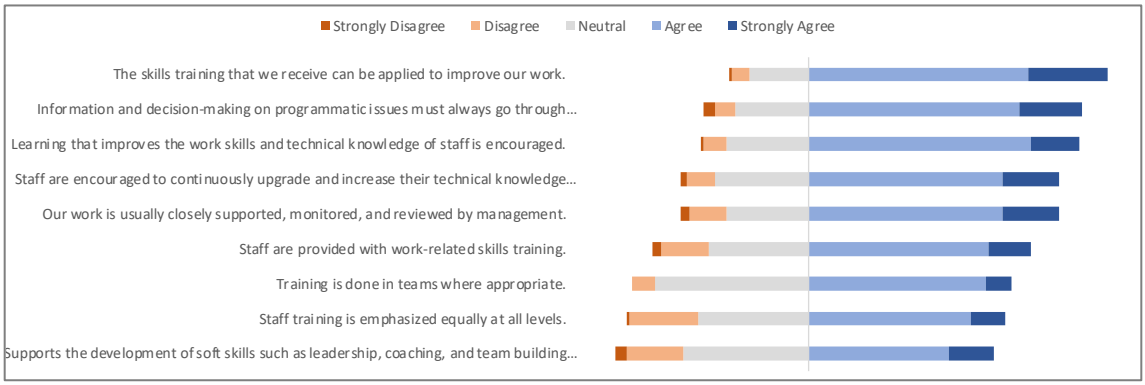
Survey Results for the Organizational Leaning (OL) Items



For the Organizational Support items (Figure 5), the organization expressed encouragement that items towards the top suggested the trainings implemented in the organization appeared impactful and well promoted. For example, two of the top scoring items were "The skills training that we received can be applied to improve our work" and "Learning that improves the work skills and technical knowledge of staff is encouraged." However, the two items that created the most discussion in the MEAL COP were "staff training is emphasized at all levels" and "Staff are provided with work-related skills training" with 19% and 15% negative responses, respectively. The interpretation reconciling the top and lower scoring items was "a supply issue," meaning that trainings provided were strong, but the frequency didn't allow for enough opportunities and forced the organization to focus on specific segments of staff.

Figure 5.3

Survey Results for the Organizational Support (OS) Items



Item results for the Capacity to Do Evaluation (Figure 6) contained the lowest scoring item on the survey: "We have long-term, dedicated financial support to ensure evaluation activities across all programming where evaluation is required." The results showed 36% of staff provided a negative response to the item and another 28% had a neutral response. The MEAL COP was not surprised by this outcome citing the funding challenges tied to humanitarian program funding, specifically. A Program Director explained the result:

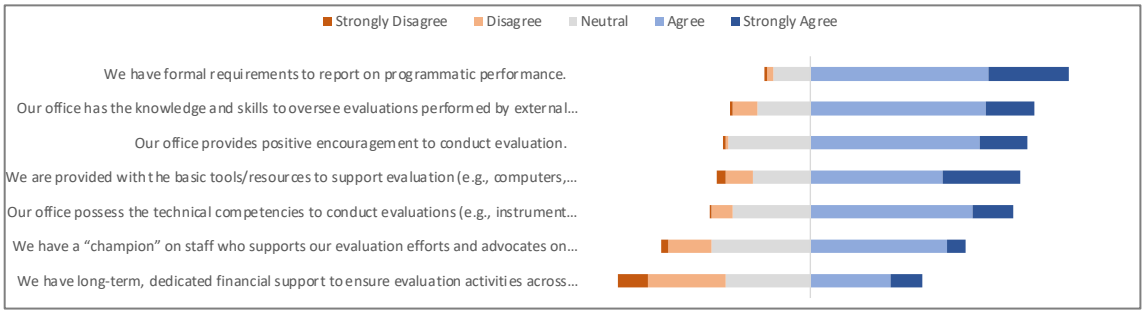
> Because of the world we live in, we are having a radical increase in our humanitarian program…humanitarian donors do not provide, generally, MEAL budgets. So overall I feel that this is very reflective of the truth, however, I do think that it could be interpreted as: management need to make sure that long term dedicated financial support. Well, we are dependent 80% on restricted funding, so donors actually need to be able to make that available.

The other meaningful item from this factor was "We have a 'champion' on staff who supports evaluation efforts and advocates on our behalf when required." The relatively lower response scoring to this item provided some new insight for the MEAL COP lead, who said:

> We don't have [developing champions for evaluation] in our evaluation framework or the way we do evaluation. So, thinking of policy and new ways of working, I was thinking we always talk about dedicated resources and that we don't have enough funding but maybe a design of champions could be something that can help in-between, to develop new pathways for funding or encourage use of evaluations where it hasn't previously existed.

Figure 5.4

Survey Results for the Capacity to Do Evaluation (CTD) Items



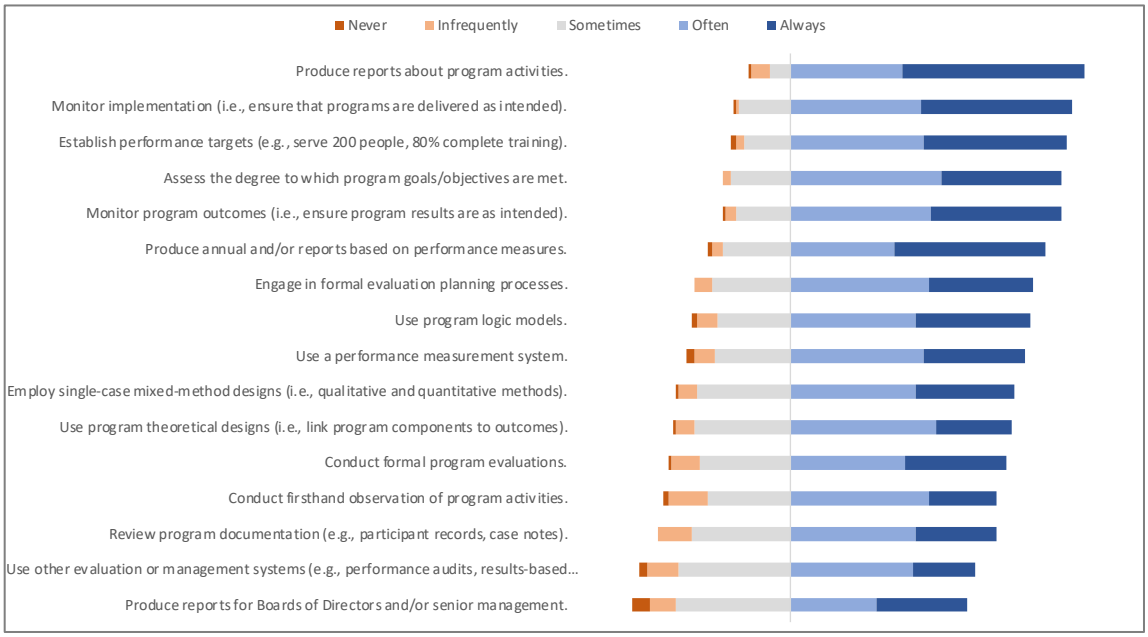
The Evaluative Inquiry factor was the relatively highest scoring factor in the survey (Figure 7). For each item, the factor asked the respondents to "indicate the extent to which your office has engaged in the following evaluation activities." The lowest scoring item in the set, "Produce reports for Boards of Directors or senior management" still had over 50% positive responses. For some of the lower scoring items, like the

production of reports for boards of directors, or "conduct firsthand observation of program activities" staff interpreted the results as reflective of the role program staff play in the organization. Although reports to Boards indeed do sometimes contain evaluation data, staff were skeptical if respondents would be aware, as it happens at the highest level of the organization. Additionally, staff pointed to the use of external evaluators for the lower scoring response to first-hand observations of program activities, believing the wording of the statement asking for extent the office has engaged in the following activities could have led respondents to interpret the items as meaning internal staff only.

The overarching interpretation of the Evaluative Inquiry items, as well as the Capacity to Do Evaluation Items, was that they demonstrated the "maturity-level" of the organization's capacity. In other words, respondents scored the most fundamental items in each factor highest, and the more secondary or advanced items scored lower. For example, 80% of respondents gave positive response to items "Produce reports about program activities" and "Monitor implementation" and "Establish performance targets." The high scores are not surprising given the systems, processes, and policies the organization has in place and practiced for decades. And with that perspective, some staff members suggested the results would be more effective by seeking to understand why those answers were not closer to 100% positive, and a doing a qualitative follow-up would be insightful to understand the issues. I will come back to the theme of organizational maturity and the assessment in a subsequent section, reviewing the overall interpretation themes.

Figure 5.5

Survey Results for the Evaluative Inquiry (EI) Items



There were two sets of items within the Stakeholder Participation factor. The first asked respondents to "indicate the extent to which the following stakeholder groups typically participate in evaluations in your office or for your office" seen in Figure 8, and the second set asked respondents to "indicate the level of participation in evaluation for the following stakeholder groups" seen in Figure 9. Both sets of questions used the same 8 stakeholder groups, but the frequency items used a five-point scale (never-infrequently-sometimes-often-always), and the participation level items used a three-point scale (low-medium-high).

Upon review, it is moderately suspicious that the results per stakeholder group are similar across the two questions. This could suggest that respondents did not take enough time to differentiate between the questions of frequency of participation and level of

participation. However, the organization did find some validity in the ordering of the stakeholders. For example, the lowest scoring group seemed appropriate to most staff, with one county office staff stating "project sponsors, donors and/or funders are only involved in the design process of the evaluation or making the term of reference. And after we make a decision about the findings to share with them." Accordingly, the organization could not decide if this result was a good outcome or a result suggesting they needed to be more involved.

A major focus of the organization is the empowerment of its beneficiaries, with a particular focus on vulnerable groups. Accordingly, the organization was concerned with approximately 50% of the respondents suggesting "Stakeholders representing different identities (e.g., race, disability, age, migration status, etc.)" were only sometimes, infrequently, or never involved in evaluations. The MEAL COP Lead suggested it needed to be more of a priority in the updated evaluation policies, saying "regardless of external or internal evaluations, or the type of donor funding it, we must make beneficiary participation a standard practice."

Figure 5.6

Survey Results for the Stakeholder Participation Frequency (SPF) Items

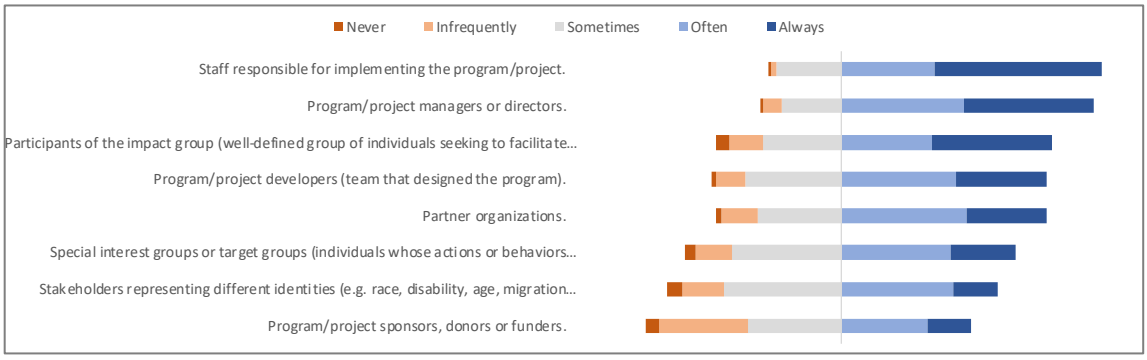
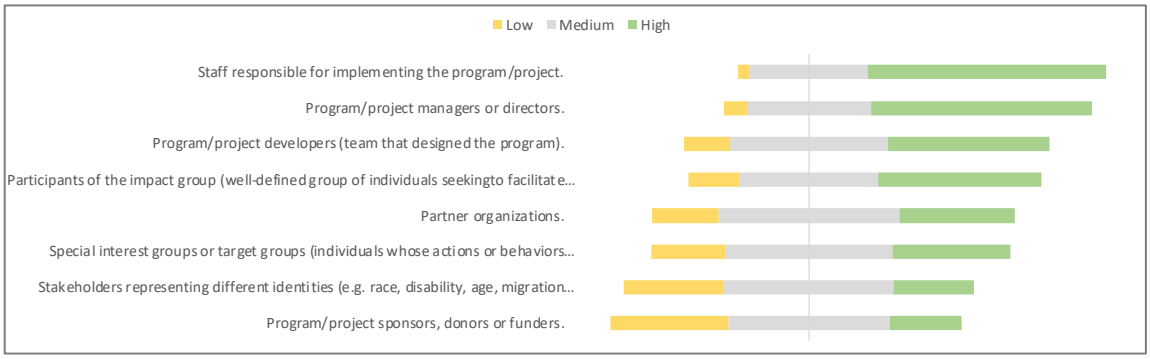


Figure 5.7

Survey Results for the Stakeholder Participation Level (SPL) Items



The two *a priori* factors Evaluation Findings Use (Figure 10) and Process Use (Figure 11) loaded together in the exploratory factor analysis. Additionally, they were the two lowest scoring composites of items. Due to the lower composite scores, as well as the organizations *a priori* beliefs about the organization's struggles with evaluation use, the organization put a lot of emphasis on the factors in their analysis. However, not all staff viewed the results through that relative lens, as the scores were still very positive looking only at its percentage breakdown. A Senior Director for MEAL said, "I see the

results as positive in that it tells me evaluations are truly being used to increase impact as opposed to a check-the-box compliance type accountability exercise. So, these are not necessary a bad result."

Similar to the Capacity to Do Evaluation and Evaluative Inquiry items, the results again promoted a discussion on the effect of organization's maturity on the results. The more fundamental evaluation uses like "Meet external accountability requirements" or "Justify program existence or continuation" or "Develop a better understanding of the program/policy/intervention being evaluated" had the highest scoring. It was the "secondary" uses of evaluation findings or process use, like "Improve management practices in the office" or "Question underlying assumptions about what we do" or "perform outreach and public relations" that scored lowest. Regardless, there seemed to be consensus Evaluation Use was a critical capacity issue for the organization to focus effort. One Technical Advisor said:

> It's our weakest link [in the organization]. We have these really cool programs, and we improve them, or we do a good job, and we have some good outcomes, and we actually measure it…but then we stop. We don't ever take that the step farther to actually share that information and take the impact to the next level.

A Sr. MEAL Advisor tied the issue to the funding structure and use of external evaluators:

> When [the organization] gets results…we use it within existing programs. But because of the way our projects are modeled, when a project's over, you don't get paid and you're done, and you move on…and that's it. We don't do anything with

the data. We don't have a dissemination event. We don't go back to the communities and talk about what we've learned. It's just kind of like we do the evaluation and then we give it to the donor and then we're done. It's really sad.

Lastly, a Program Director, put together a few of the items to summarize the "secondary" uses of evaluation the organization needed to improve on:

You can see outreach and public relations [scored low], there was something on networks and another on peers…so there's something around the using of our evaluation to both influence our management practice, but also to influence our donors and our networks and advocacy externally.

Figure 5.8

Survey Results for the Evaluation Findings Use (EU) Items

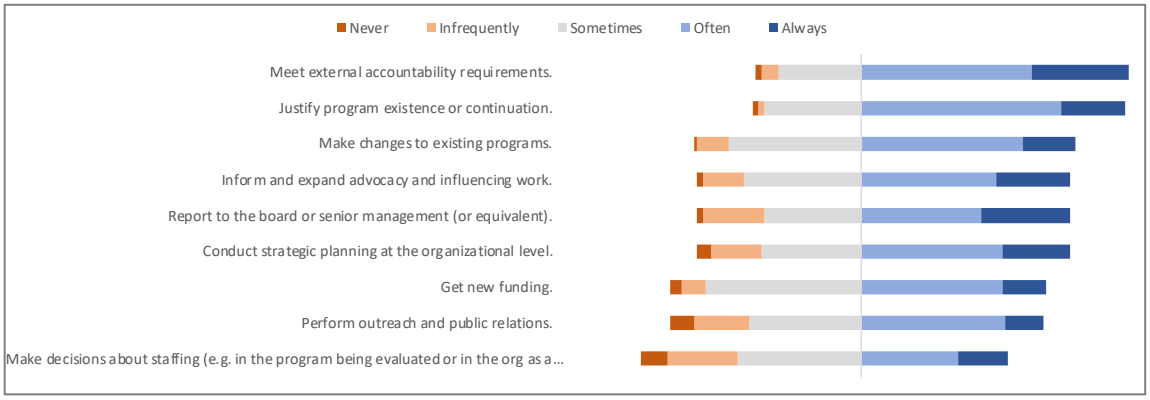
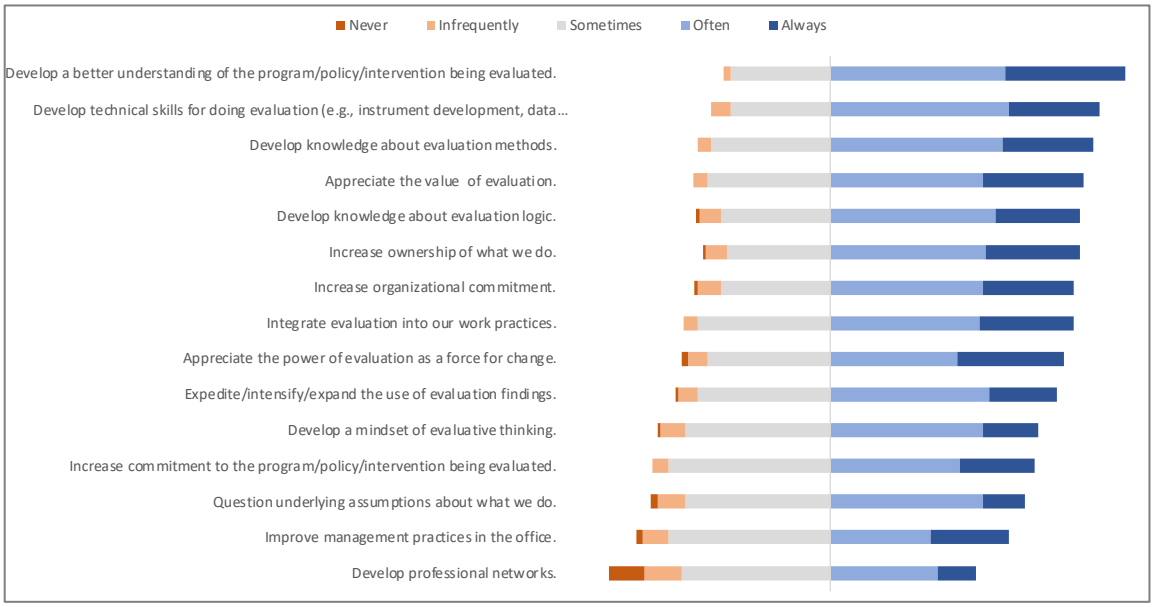


Figure 5.9

Survey Results for the Process Use (PU) Items



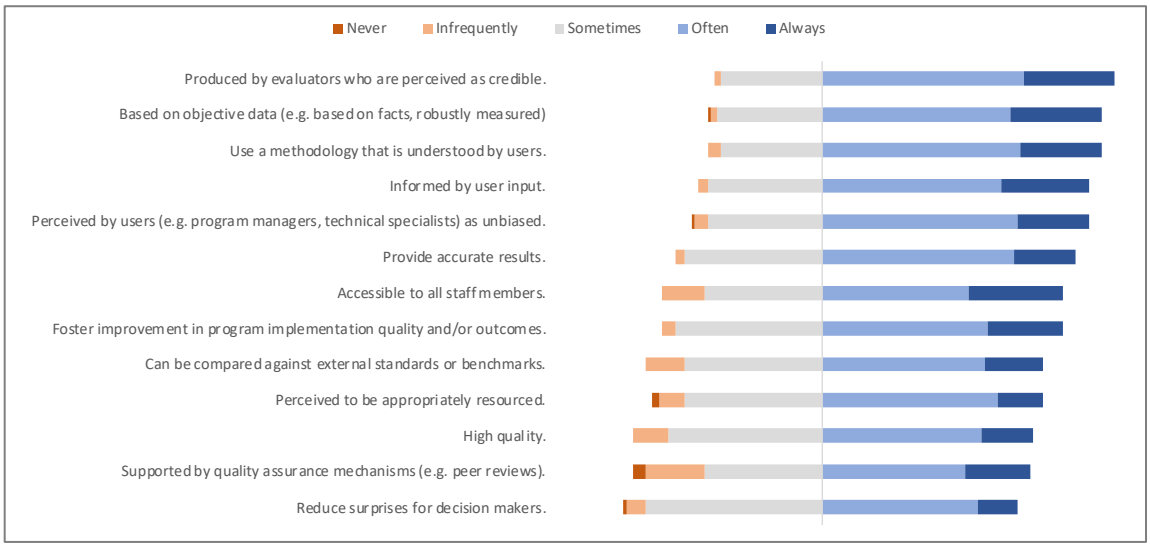
Lastly, the Mediating Conditions factor brings together conditions that influence the use of evaluation findings (Figure 12). The lowest scoring items asked respondents if evaluations in their office "reduce surprises for decision-makers," and "are supported by quality assure mechanism (e.g., peer reviews)." There was some commonality interpreted by the organization in the lower scoring items, suggesting the lower frequency of the use of quality assurance mechanism and the use of "external standards and benchmarks" illustrated a siloed approach to program evaluation.

One other key item for the organization was "evaluations in their office are perceived to be high quality." That item, along with the overall positively skewed results of the survey, led to a long discussion within the COP about a perceived lack of quality of evaluations in the network, and why staff were prone to suggest otherwise in the survey.

In the next section on overall interpretations of the assessment, one of the noteworthy themes is the connection between evaluation capacity and the quality of evaluations reviewed by the MEAL COP.

Figure 5.10

Survey Results for the Mediating Conditions (MC) Items



Themes from the Organization's Interpretation of the Assessment

In this section I discuss the primary themes from the organization's interpretation of the survey data and perceptions of the overall assessment process. I developed the themes from data collected in the 10 semi-structured interviews with 25 members of the organization, undertaken after the collection and review of the survey data. The themes cover the organization's perception of the data's validity; if the assessment met their goal to inform the next evaluation policies; the process use of the assessment; how the survey tool could have aligned or contextualized better; the assessment's connection to an evaluation quality scoring rubric; and who has the agency to improve capacity.

Positively Skewed but Valid Results

A consensus sentiment towards the results was suspicion towards the positively skewed responses. It appeared the respondents had a different perspective of the organization's evaluation capacity in comparison to the MEAL COP team who led the assessment, as well as the group I interviewed to discuss the results. The MEAL COP Lead said, "In general, I am suspicious about the results. I can't help it. They look too positive to me and it's hard to understand why. It's just a clear surprise from my side based on what I see on a regular basis." However, although this sentiment was common, when discussing the relative order of items scored within a factor most interviewees affirmed the results as valid and agreed the lower scoring items were what they perceived to be the biggest capacity gaps. Therefore, the MEAL COP still perceived the results to be valid, especially in relation to one another, but hesitated to celebrate the organization appearing to have as strong evaluation capacity as the scores suggest.

The main takeaway at the factor level was the need to improve the two *a priori* Evaluation Use factors (Findings Use and Process Use), as they had the lowest composite scores across the factors and aligned with the MEAL COP's *a priori* hypothesis that evaluation use was an area in need of improved capacity. However, there was consensus that analyzing specific item results was a more effective lens to improve the evaluation policies and target specific areas for improvement. A Sr. MEAL Director summarized:

> What I see as most valuable from this analysis are the elements that are at the bottom in each of the factors. Those either confirm some situations that we already know in terms of gaps that we have, but also provide some ideas of areas

that we may not have considered. And we should continue exploring or discussing, especially in light of the review of the policy. This gives us substance to push things we care about.

A Technical Advisor shared the sentiment saying:

> Even though everything was quote-unquote overwhelmingly positive…that's ridiculous. [We have] been stressing, for years, what are we doing wrong? What are we failing at? What could we do better? We have this whole entire series called failing forward. So, I think that pulling out the places where we are seeing [negative results], I think that's a good approach to improvement. I don't really know how else we reflect on our successes and failures, so finding those little glimpses of gaps is helpful.

Further supporting the MEAL COP's sense of validity in the survey results, multiple MEAL COP members spoke to patterns or themes they found in the data, within or across the factors. For example, mentioned earlier was a Program Director suggesting the results showed the organization's ability to "influence our management practice, but also to influence our donors and our networks and advocacy externally" needed to be a focus. A Learning Advisor independently agreed, saying he saw the results reflecting a need to more directly engage senior management in the use of evaluation and to create cultures of learning.

Meaningful Input for New Evaluation Policies

A primary reason the organization participated in this research was to inform an update to its evaluation policies, hoping low scoring factors and items could provide

direction on what the polices needed to further highlight and improve. Simply mapping the assessment factors and the evaluation guidelines was helpful in identifying gaps in the policies, and now the assessment results had highlighted areas for growth. The MEAL COP Lead was confident in using the results as intended, saying:

> [The assessment] gives us some substance to justify some points where we want to put attention. I was looking at the lowest scores, the use of the theoretical design for doing the evaluation, for instance, we are always pushing and saying, we're doing theories of change and those theories of change should inform evaluations. Now we have something to prove that we are not really doing that as much as we expect. Stakeholder participation is not what it should be, not aligned to our values. So, I'm going to use those kinds of findings to support things that we already suspected, and that should be highlighted.

Beyond identifying the areas for direct policy edits and additions, MEAL COP staff believed the assessment would help validate the policy changes in the perception of the network. Multiple staff said the organization's network can perceive new policies, promoted best practices, or culture change initiatives as lacking contextualization or relevant evidence, and this assessment would help mitigate that possible pushback. A Sr. MEAL Advisor summarized:

> This actually gives teeth to updating our evaluation policy. Otherwise, it would've been just a nice to have feel good, all the latest things in there…but now we can say we are doing X, Y, and Z, because through this evaluation we got to things we could be doing better, which is validated by people working [in the organization].

So, it's not just our thoughts. So, I think this is very useful in that regard, we can say it is backed up by data that tells us what our needs are.

The organization plans to revamp their evaluation policies in the coming months and expressed confidence the organizational evaluation capacity assessment had created a foundation for an effective outcome.

Process Use of the Assessment and the Concept of Evaluation Capacity

A few MEAL COP members suggested an indirect yet meaningful benefit from the study was the introduction of the evaluation capacity model. During the process to contextualize the tool, staff had robust discussions on the inclusions of items on Organizational Support, Organizational Learning, and items in the Mediating Conditions factor. The dialogue often focused on the MEAL COPs control over decision-making in those arenas, and therefore if the team should consider including the factors in the assessment. One Technical Advisor said the "the debate helped it click for me, that [the organization's] cultural inputs really mattered for our evaluation practice and also were affected by our evaluation practice." A Sr. MEAL Advisor reflected on the conceptualization of evaluation capacity conversations as learning moments, saying:

> I was really happy some of our MEAL staff in countries suggested we remove those pieces, because it created a learning moment. We were able to discuss the connections and I hope it changes how they interact with other parts of [their organization] and think about how management decisions affect evaluation, or resource development teams interact with evaluations.

Additionally, one MEAL staff member asked to return to the model in literature and review the directional loadings between factors. Although we hadn't been able to fit the sample model to the original structural equation model, the staff member wanted to understand what other inputs could affect Evaluation Use beyond focusing on the factor itself, asking:

> If you did make that assumption [that the organization fit's the research's structural model], you could say things like if we want to improve the use of findings, stakeholder participation is one of the links that does that. And if we have people more involved in evaluations, it makes sense that they would be more invested in using the evaluation findings.

Although it wasn't a direct goal of the organization, it appeared the very process of discussing and illustrating a model, or factor structure, of evaluation capacity was a capacity building experience for the staff involved in the process.

Aligning Results with Organizational Maturity

As noted in the item review, the organization's staff interpreted a few factors to demonstrate item scores in order of their maturity. For example, in Evaluative Findings Use, 80% of respondents gave positive response to items "Meet external accountability requirements" or "Justify program existence or continuation" which we would expect in an organization with established MEAL practice. More complex or advanced uses of evaluation like "Improve management practices in the office" or "Question underlying assumptions about what we do" naturally scored relatively lower. Apropos, as this should be a natural order, the organization struggled to interpret where to focus more:

should they be concerned the fundamental items weren't closer to 100% positively scored, or attempt to increase the more secondary uses of evaluation? Multiple COP members asked if comparative data from other organization's existed for comparison, hoping a benchmark could help provide more insight, but I could not find literature of other organizations applying the model beyond the original validation. The lack of benchmarking data underscores the significance of this research, and suggests future research ideas, but it left the organization struggling to reach consensus on where to focus.

In lieu of baselines from other organizations, multiple staff members suggested changes to future versions of the tool to account for the issue. A Technical Advisor suggested:

> [The organization] may have almost been too mature of an organization to use the tool. Some of the questions are about the systems in place and processes in place that we know we have. Like some of those things around: do you measure outputs? Do you have logic models? Do you have these types of systems and of course the answer is yes. So, I wonder if we can tailor it further to the kind of barriers or obstacles we are now identifying.

The MEAL COP Lead believed there was value in the inclusion of those fundamental items, suggesting 80% wasn't a high enough score and therefore tells the organization some capacity building issues remain. However, the MEAL COP Lead did express some regret for not tailoring the tool with a more focused eye towards pain points:

> Maybe we started too low on the capacity, on the minimum capacity that we would expect. And that prevented us from really going deeper into the pain points. What is stopping us from using our evaluations? How are we managing consultants or how are we managing to deal with five different evaluations at the same time when you have very limited resources? I think we can improve the next version to target these issues.

Improved Contextualization

Similar to staff suggesting modifications to the tool for future iterations based on organizational maturity, there were some lessons learned in reflection on the contextualization process of the tool. A few staff members brought up the desire to split results between humanitarian aid programs with minimal evaluation funds, and development programs that have funded evaluations. The staff tasked with adapting the survey discussed the issue during the contextualization process, but the MEAL COP found it difficult to create a demographic because staff worked across both program types within the same country, region, or technical specialty. Moreover, evaluation capacity of the organization should apply to both types of programs. However, when it came to factors like Stakeholder Participation, or individual items like "Our evaluations are perceived to be appropriately resourced," staff would point out the structural differences between the types of programming would affect the results. Accordingly, interpreting the results of those items was difficult as different expectations would be appropriate based on the type of program the item measured.

Another issue was the potential role respondent location in the organization played in the results. When reviewing results between country offices and regional offices, especially in North America and Europe, the western regional offices consistently scored items lower. The prevailing interpretation was the regional offices and centralized unit in the organization is more aware of gaps and issues across the network, and therefore have a more critical perspective. Another interpretation was the country offices had some implicit incentive to score themselves well to appease the regional and centralized units of the organization, but not all staff bought into that hypothesis. The MEAL COP Lead added:

> One thing I don't know how to address…is how this group of country offices are so different from the U.S. and European offices. I think this just has to do on what role you play during an evaluation process, so I guess the farther you are from the process, of course you are more critical on the process that you don't really influence that that much.

The prevailing conclusion was offering the survey across the organization served the purposes of providing feedback to improve the evaluation policies and target areas for capacity building, but future iterations of the survey should bound the sample with more clarity, to add confidence in the quantitative data's validity.

A similar issue concerned the role of fundraising. Some items asked the extent your office uses evaluation to "get new funding" or "perform outreach and public relations." A Program Director suggested:

> The country office doesn't have any ability to generate own resources, to fill evaluation gaps, for instance. They fully depend on their country partners in the north. So, what we could have done is segment questions for some roles and ask others different questions.

However, other staff pointed out country offices should still have some idea if their country partners are using evaluations for fundraising and should actively be promoting the use of their evaluations to be used, even if they can't control if it happens. Therefore, they argued, it was still appropriate to ask country offices the fundraising items, although the survey writers could have improved the items to better represent who should be using the evaluations for fundraising, rather than asking about the respondents' "office."

Lastly, the mix of internal and external evaluations in the network caused interpretation struggles. Similar to the program type issues, the MEAL COP team discussed the implications for the type of evaluator in the contextualization process, and it had an impact on items included in the final survey and how they wrote the items. However, some MEAL COP members remained concerned it clouded the results. One COP member said:

> When analyzing the results, we have to take into consideration that 66% of our evaluations, or at least the evaluations we have scored, are done by external parties. Therefore, some capacities of the study may be expected to score higher or lower depending on the role [the organization] plays in the evaluation process.

Maybe we should have only focused on internal evaluations for a clearer sense of our capacity.

Evaluation Capacity's Relationship to Evaluation Quality

The primary goal of the assessment for the organization was to inform the improvement of the organization's current evaluation policies. However, in a meeting I joined with the MEAL COP to discuss the survey results, COP members questioned how the assessment's results should align and compare to a rubric used in the COP to score evaluations. In a commitment to transparency, established in the current evaluation policies, the organization strived to place all evaluation results online for public review. In the recent past, the MEAL COP developed a scoring rubric with ten 1-point questions, to provide an easily interpretable and standardized score to the quality of each evaluation report.

The MEAL COP provided the scoring rubric to me, to compare to the evaluation capacity assessment. The scoring rubric asks multiple questions about what is in the evaluation report, like an executive summary, methodology, or impact metrics and provides a binary score for each question. It includes some focused concepts like discussing failure, unintended consequences, a theory of change, and disaggregated data by gender. The result is a set of questions that assess the final report more than the evaluation itself, and therefore is substantially different from the evaluation capacity assessment.

However, the MEAL COP as a whole was curious how capacity scores were relatively high when many evaluation's quality scores were poor. One MEAL COP

stated the issue as: “The average score is about five and a half out of 10. And so how can we make sure that those capacities, if they do exist, are translating into higher quality evaluation that we are using more consistently?” The MEAL COP Lead was under the opinion that the rubric needed to change, and that the capacity assessment offered the opportunity to imbue the rubric with new ideas:

> I was happy to hear from the team that we have a big inconsistency between what we did in the study, where we looked at these are the capacities we would like to have from evaluation, on our ability to do evaluations, versus the criteria we're using to actually assess the quality of the evaluation product at the end, when the evaluation has taken place. There's a big mismatch there, so we're looking at this from two very different perspectives. So, a very practical step that we need to make happen is that we need to adjust the criteria we're using to assess evaluation products, versus what we're trying to assess in terms of evaluation capacity. [The issue of adapting the quality criteria] had been circling a lot for mainly three years now and people don't seem to be, or didn't seem to be, open to change the criteria for assessing evaluation quality. So now again, I think this brings an entry point for that.

Furthermore, the evaluation capacity assessment had COP members finding new insights for embedding practices to encourage evaluation use. A Senior MEAL Advisor suggested if the evaluation quality scores were more visible by senior leadership and the country offices themselves, the culture of learning and use of evaluations outside of programs could increase:

> The quality of evaluation is not a variable that leadership are using to evaluate performance. And, if we have the policy being updated, then of course there should be some information that goes every year to the leadership, and they can see how [the network] is really meeting the commitments of the policy.

Moving forward, the MEAL COP discussed the importance of reconciling the evaluation polices and evaluation criteria, using the evaluation capacity assessment as an input to improve policies and a tool to assess progress:

> We will need to trace [the evaluation policies, quality criteria, and survey results] as a process. The evaluation policy says any good evaluation process should have these, let's say, 10 standards. From those 10 standards, what I would do in the evaluation capacity survey going forward would look at those elements over time. And then within the evaluation practice, the scoring criteria of evaluation reports needs to better reflect the policies' standards and be supported in the templates and tools we supply programs, no matter if you're doing internally or externally managed evaluations.

Who are the Change Agents?

As noted in the description on the structure of the organization, the MEAL COP, located within the centralized unit of the organization, are "in service" to the rest of the network. They have tremendous ability to influence practice and behaviors through the evaluation policies, development of standard tools, and conveying mechanisms like the MEAL COP, however other stakeholders may have different and equal agency to build

capacity. For example, one Sr. Director for MEAL discussed regional governances' ability to create change:

> I think, generally, in terms of the training and capacity strengthening efforts that we organize, this gives us a few areas that we might want to focus on going forward. A lot of that tends to happen on a regional level. So, I think we may want to do some analysis regionally if there are certain things that stand out because not all regions have the same needs.

Less straight forward, many MEAL COP members focused on the need for leadership to drive some of the change they wanted to implement. One Program Director used culture examples to drive home the challenge:

> So, you can say you have low capacity to develop a mindset of evaluative thinking, or you have a culture of defensiveness, and I probably would agree with that and I know why: people are under pressure to deliver. How you change that is not a very easy answer. It doesn't feel like capacity building. It doesn't feel like, oh, let's put in a new finance system. It feels like something far deeper. And I suppose where I question whether it's a capacity to build or an organizational will from the top down. And so, I could go to those leaders and ask them for their support in this, but it is not as simple as getting [that commitment]. They can say we want people to talk about failure but making that happen is quite complicated. So, we need to think hard about who to target and how capacity gets built in those areas.

Accordingly, the MEAL COP believed more leadership sponsorship on the evaluation capacity assessment process would have been impactful. The MEAL COP Lead hoped future activities would learn the lesson, saying, "this goes beyond MEAL, in terms of decisions and commitment, and all of it. So, hopefully in the next round, if we do this again, it must be sponsored by a higher-level platform."

Next Steps for the Organization

At the completion of the process the organization made the survey results accessible to staff throughout the network, on their internal server. I developed a pivot table with all demographics, factors, and items available to quickly segment the data. I also had the 5-point scale, stacked, colored-line visualizations embedded in the tool, so when the data was segmented the results would update to reflect the demographic choices. The MEAL COP believed staff would use the data, saying: "For colleagues to autonomously filter and look at their results, is excellent. Normally it's just a spreadsheet and everyone can do what they want. It is a really useful tool and easier for colleagues to use the data." A Sr. Director of MEAL agreed, saying:

> I tend to think positively when people see data [in our organization], they're usually first of all, curious about it, they want to know, and then second of all, they want to do something about it if possible. So, I'm modestly optimistic that this can be useful beyond what we plan to use it for, and I think it's useful no matter where people sit in the organization.

Some MEAL COP members believed the organization needed to further examine the low scoring items and spend time thinking about what changes would add the most

value. One MEAL COP member suggested "for each possible area to build capacity, we need to examine which would make a high difference to our overall evaluation quality. And which ones are actually feasible to shift." However, other staff pointed out not every area of weakness may need significant investment, and tools already exist to help build capacity. For example, discussing how to improve evaluation use and change common practices at the end of evaluations, one Sr MEAL Advisor suggested:

> We also have a set of online MEAL modules for staff to learn just generally about MEAL and then specifically how we do that [at our organization]. So, there may be some areas that we might want to focus on either an existing module, adding content, for example, or revising content or even new modules in areas where we have perceived weaknesses, like the use of evaluation findings.

Adding to that idea, another MEAL COP member pointed out they could begin targeting training at non-MEAL staff.

> I don't think our fundraisers have expertise on evaluation. So, can they read and understand evaluations so that they can communicate for funding? Or use research? I don't think we have a focus to train them right now. And I don't think we have made that connection yet…I think it could make sense to develop that capacity in the team.

Overall, there seemed to be two major next steps for the organization. First, the MEAL COP stated the changes to evaluation policies would begin within the next few months. The MEAL COP expressed confidence the results from the survey would not only identify areas for direct policy edits and additions but help validate and promote the

policy changes in the perception of the network. In general, the factors of stakeholder participation, evaluation use, as well as some specific items like use of evaluation champions, were the strongest themes they were taking from the assessment and into the policy discussions. Additionally, the MEAL COP desired to align the evaluation scoring rubric with the updated policies and capacity assessment.

The second major next step the MEAL COP was discussing was the reduction of the tool and continued administration to assess change over time or assess more bounded samples in the organization. The organization was disappointed other samples from organization's did not exist to aid in interpretation, but this first assessment created an internal baseline they could use to analyze change over time. The MEAL COP Lead suggested it could be an accountability tool, "to assess performance on how we are doing in making progress towards the policy or key elements of the policy." In addition, the organization discussed targeted qualitative follow-up to either understand why certain items score poorly, or to determine the right evaluation capacity building activities to increase capacity.

As the researcher, the process of supporting the organization through the evaluation capacity assessment process was insightful and encouraging. I feel honored to have worked with a group of MEAL professionals so committed to making a difference in their own organization, and most importantly, making a difference for vulnerable people around the world. I will provide ideas on how organizations can apply similar tools and processes in their own context in the conclusion of this dissertation, however,

this case reminded me dedicated individuals are always the most critical ingredient to any successful capacity building initiative.

Chapter Six: Conclusion

Summary and Discussion

The overarching goal of this dissertation was to support human service organizations in the implementation of organizational evaluation capacity assessments to inform their evaluation capacity building plans. Overall, the field of evaluation capacity building still needs to focus on ensuring evaluation capacity building efforts make a difference through reaching the right people, increase organizational learning, and are investigated for their influence (Preskill, 2014). Additionally, Nakaima and Sridharan (2017) suggest that the field could benefit from "better stories of the dynamics of organizational capacity building from specific case studies." My hope is the qualitative evidence synthesis of theories of organizational evaluation capacity and case study help move the academic outputs closer to regular application in the field. In this conclusion, I will summarize and discuss the two sub-studies of the dissertation and integrate their findings to theorize an application framework for an organizational evaluation capacity assessment.

Qualitative Evidence Synthesis

To support the ability to evaluate evaluation capacity building, researchers have hypothesized, illustrated, and attempted to measure evaluation capacity, the outcome of evaluation capacity building. The growing number of models and case studies on

theories of organizational evaluation capacity provided me the opportunity to systematize a review of their overlapping characteristics and find opportunities for further research. The four aims of the qualitative evidence synthesis of theories of organizational evaluation capacity were to: (1) synthesize the extent of research theorizing organizational evaluation capacity models; (2) detail dimension commonality across EC models; (3) examine the extent the models have undergone tests of validity; and (4) identify possibilities for future research to expand the evidence base. I discuss possibilities for future research to expand the evidence base in a subsequent section of this chapter.

To achieve the first aim of the sub-study, I used methods that closely resemble the methods of a traditional systematic review but with minor adaptations, like the absence of methodological quality assessments, to account for the heavily qualitative nature of the sample. Ranging from 2002-2021, I collected 16 articles and assessment tools representing the evolution and diversity of organizational evaluation capacity models. I describe the body of research by grouping it into three different periods, categorized by the types of models developed: conceptual models, structural models, and quantitatively developed models. Conceptual models include descriptions of evaluation capacity with at least two dimensions, structural models begin to relate the dimensions to one another, and the quantitively developed models use surveys and factor analysis to refine their structural model. Much of the research builds upon previous publications; therefore, in the narrative, I order the groups, and the results within the groups, chronologically.

I believe one of the most meaningful things about the narrative is identifying the legacies of early research and the continued evolution of their ideas. For example, one of the earliest conceptual models was from McDonald, Rogers, and Kefford (2003), establishing the notion of supply and demand as the primary evaluation capacity factors. Eight years later, Nielsen, Lemire, and Skov (2011) produced the first quantitatively developed model of organizational evaluation capacity using the supply and demand structure. Similarly, Cousins and Bourgeois make incremental steps from their conceptual model (2004) to a structural model (2008), and a quantitively developed model in (2014). Additionally, the tool used in the case study from Gagnon et al., (2018) included Cousins and builds on the 2014 model. The narrative underscores the meaningfulness of the body of literature building on itself and moving the field forward.

To achieve the second aim of the sub-study, I identified 7 distinct models to review dimension commonality across models. Comparing the models, I suggest 9 common dimensions (Table 3, Chapter 4): (1) organizational culture and policies; (2) infrastructure and systems; (3) leadership; (4) organizational resources; (5) human resources; (6) capacity to do evaluation; (7) capacity to use evaluation; (8) communications; and (9) a culture of learning. Moving forward, the 9 dimensions could support further expansion of current models, adapt and combine models, or create a new model taking survey items from different instruments.

The 7 distinct models fit the 9 dimensions reasonably well, with only two models lacking more than 1 dimension. Nielsen et al. (2011) fit the least based on dimension names; it has only five of the nine common dimensions, but shares more commonality

based on its survey items. Taylor-Powell and Boyd (2008) do not include three dimensions that have some common overlap around the use of evaluation (the capacity to use evaluation, communications, and organizational learning). The other five models have only one missing dimension. The only dimension missing in at least 3 models is human resources.

Landing on the 9 common dimensions was an iterative process. Methodologically, I leaned on my guiding table, where I extracted key data points for each eligible article and tool (e.g., study design, sample, type of model, dimensions of the model, assessment items, validation tests, origin of instrument creation, and applied use of the model) to consider different dimension options. The decision-making process was difficult because so many similarities existed between models, but they were nested in different ways. For example, human resources could fall under organizational resources. However, I decided to keep human resources separate as evaluation capacity building of individuals is such a priority in practice, it felt important to separate it from organizational resources like budgeting. Similarly, organizational learning could have fit under evaluation use or organizational culture, but when evaluating overarching organization learning environments the use of evaluations is not comprehensive, suggesting a need to review them separately. Similarly, organization learning is only a sub-culture of its larger influence. Ultimately, I believe my 9 common dimensions are a strong illustration of the consensus, but I think any of the models are appropriate if they best fit an organization's conceptualization of its evaluation capacity.

To achieve aim 3, I narrated the validity base of the models, which has evolved from content validity to construct validity. Three models have tested for construct validity using a single sample, using factor analysis and SEM (Nielsen et al., 2011, Taylor-Ritzler et al., 2013, Gagnon et al., 2018). Replication of models would further the evidence base for the models' applied use across diverse organizations. Moreover, the models have not established robust reliability, as inquiries of test-retest are absent for all of instruments, and the work of Fierro and Christie (2016) demonstrates the vulnerability of allowing a small sample to rate an organization.

Sub-study 1 is meaningful as it is the first piece of research to systematically compile all the existing models, chronologically analyze their development and influences, analyze the extent of dimension commonality, and review the levels of validity the models have investigated. I believe they offer a meaningful contribution to the academic literature by providing data and commentary on the evidence base, as well as provide practitioners clear comparisons to find the right models and tools for their own use.

Concurrent Mixed Methods Single Instrumental Case Study

A major gap in the research on organizational evaluation capacity assessments is the absence of guidance or frameworks for organizations to implement the models and their tools for applied practice, or even a descriptive case of an organization undertaking the process. Sub-study 2 details the process of a multinational human service organization applying an organizational evaluation capacity assessment tool to support capacity building goals, using a concurrent mixed methods single instrumental case

study. The three primary research questions for the concurrent mixed-methods single instrumental case study were: (1) what considerations are necessary to implement an evaluation capacity assessment? (2) How do the evaluation experts in the organization interpret the results? And (3), how does the organization use the results to make decisions about investing in evaluation capacity building initiatives?

To answer the first research question, I detailed the organization's goals in undertaking the assessment and their process to adapt and administer the assessment tool. The process involved multiple meetings with their MEAL COP to determine the assessments' goal and direction, many iterations and authors of tool changes, and a pilot of the tool with country office staff. However, as viewed in the section on the themes from organization's interpretations of the overall assessment, the process to consider the organizations unique factors and contextualization of the tool could have been better. Interestingly, the MEAL COP staff anticipated many of the variables that complicated the interpretation of the survey data, like the complex organizational structure, respondents' roles, types of programs, use of external evaluators, etc., but still faced challenges even after accounting for them in the tool contextualization.

A major obstacle was inherent to the goals of the study: measuring a set of dimensions across an organization that large is almost inevitably going to be complex, nuanced, and contain caveats. The tension between writing items specific enough to account for complexity, but general enough to be relevant across diverse governance, programmatic, and operational boundaries will remain a balance rather than a problem with a solution. Accordingly, the organization suggested future versions may be more

useful within specific boundaries of the organization, for example, within a region or program area. However, I believe the overarching organizational scope of the assessment in the case study fit the goals of the organization, which was to inform upcoming changes to their evaluation policies and identify areas for improvement that the centralized unit of the organization could respond with support.

To answer research questions 2 and 3, concerning how the organization interpreted and plans to use the results, I provided a summary of how the organization interpreted the quantitative data, themes of the organization's overall perspectives on the assessment, and summarized the next steps the organization expected to take after completing the assessment. The result was a narrative providing a window into both their "micro" perspectives on highlighted items and a "macro" discussion on the process.

As noted in the chapter, a consensus sentiment towards the results was they were positively skewed. However, interestingly, when discussing the relative order of items scored within a factor most interviewees affirmed the results as valid and agreed the lower scoring items were what they perceived to be the biggest capacity gaps. Therefore, the MEAL COP still perceived the results to be valid, especially in relation to one another, but hesitated to celebrate the organization appearing to have as strong evaluation capacity as the scores suggest.

The organization was disappointed there wasn't more variability between the factors. The similar results made for a more difficult interpretation of the results, which was useful for the case study description, forcing more detailed explanations from participants, but a hypothesis of the organization was the assessment would point to clear,

high-level areas for capacity building. Ultimately, that still happened, with the organization believing the results confirmed Evaluation Use was as an area in need of capacity building, but the result wasn't as self-evident as they anticipated. However, and more importantly, the organization felt confident in the results supporting the update to their evaluation policies. Beyond identifying the areas for direct policy edits and additions, MEAL COP staff believed the assessment would help validate the policy changes in the perception of the network.

Before undertaking the case study, the literature promoted multiple reasons why an assessment of organizational evaluation capacity prior to undertaking evaluation capacity building initiatives is meaningful. First, whether explicit or implicit, perspectives on what constitutes evaluation capacity inevitably shape evaluation capacity building initiatives (Naccarella et al., 2007), and therefore stakeholders should agree upon what encompasses organizational evaluation before planning activities. Further, measurement of organization evaluation capacity can assist in highlighting dimensions or resources in need of more concentrated focus. Lastly, the ability to find a baseline of evaluation capacity, and then repeat the measurement in the future to reveal change in evaluation capacity, would serve to evaluate the results of evaluation capacity building activities (Preskill, 2014). I believe the organization in the case study found all three of these reasons to be true. The conceptualization of evaluation capacity for their organization provided some process use benefit and will play a role in shaping their next set of evaluation policies. The organization identified Evaluation Use as a point of emphasis in future capacity building initiatives, and multiple areas for improvement from

item analysis. And affirming the third reason, they discussed the use of the assessment for future measurement in the organization, to measure progress against the new policies.

Sub-study 2 is one of the first cases to describe the experience of a human service organization's application of a previously developed assessment of evaluation capacity and detail the intended use of their findings. The research answers the call from Nakaima and Sridharan (2017) for "better stories of the dynamics of organizational capacity building from specific case studies." The description details the value and challenges of using an organizational evaluation capacity assessment, how the process can be imitated, and provides other organizations data to benchmark their own assessment's results.

A Framework for Applied Use

Lastly, to address the absence of guidance or frameworks for organizations to use organizational evaluation capacity assessment tools, I aim to conclude the dissertation by explicating an application framework for use of evaluation capacity assessments. During the case study, I wrote memos containing my own interpretations of what worked and what the organization could have improved in the assessment process. I used those memos, data from the qualitative evidence synthesis, and the interviews and documents from the case study to develop a "construct table" for the framework. The rough construct table slowly evolved into the final application framework presented Figure 13.

The Application Framework promotes 6 phases with steps for organizations to align an existing model of organizational evaluation capacity to their context, consider confounding variables, and plan for an impactful use of the assessment to meet desired goals.

Figure 6.1

Application Framework for an Organizational Evaluation Capacity Assessment

Phases	Steps		
Vision	Goal		
	Develop a detailed goal statement to inform what outcomes the assessment should achieve, the target audiences, the necessary stakeholders, and resources needed to undertake the process.		
Plan	Use	Context	Sample
	Based on the goal statement, determine how the assessment's results will influence the organization.	Identify the unique conditions of the organization, its programs, and MEAL practice.	Based on the goal statement and use plans, determine which stakeholders need to be in the sample.
Conceptualize	Tool Review	Factors	
	Review the literature on evaluation capacity assessments to inventory possible models for contextualization.	Based on the context considerations, hypothesize a set of factors relevant for the organization's evaluation practice and compare to models in literature.	
Modify	Tool Adaption	Pilot	
	Based on the context considerations and hypothesized factors, designate a model/tool to adapt, starting with the factors and then the items.	Based on the sample decisions and context consideration, identify multiple staff members to pilot the tool and provide feedback.	
Implement	Administer	Analyze	
	Determine how to deliver or administer the tool, and how the sample can be encouraged to respond.	Determine the necessary data points or comparisons needed to inform decisions to achieve the assessment's goals.	
Share	Disseminate		
	Based on the anticipated use of the results, develop a plan to present and promote the findings to influence change and achieve the outcomes determined at the beginning of the assessment process.		

Figure 13 maps out the 6 phases vertically (Vision, Plan, Conceptualize, Modify, Implement, and Share) and includes steps within each phase. Each phase should inform the subsequent phase and therefore an organization should complete all steps within a phase before starting the work of the next. Within the step boxes I provide some instruction and suggested outputs to develop. I prefer the term framework to something more prescriptive, like methodology or procedure, as an organization undertaking an evaluation capacity assessment is going to enter into the process with different motivations, incentives, and expectations on rigor. Accordingly, I believe the phases and steps presented provide meaningful direction and suggestions human service organizations can use to plan their assessment process.

Suggestions for Future Practice

Although I have established the uniqueness of each organization and their goals should drive the process and decisions for an organizational evaluation capacity assessment, I do have opinions about important considerations. In reflecting on the experience of the sample organization, I believe the most critical steps are in the vision and planning phases of the framework. Accordingly, I want to provide a few suggestions for future practice, to include: the prominent inclusion of organizational leadership in the process; the importance of a detailed goal statement for the assessment; the consideration of item "maturity" and expectations on the results; and the dissemination of the assessment's goals and process prior to implementation of the instrument.

Evaluation capacity building literature is very clear that organizational leadership is one of the most important factors in ensuring evaluation capacity building initiatives lead to sustained, improved practice (see Chapter 2). As noted in the case (Chapter 5), the MEAL COP Lead shared that if done again, they would have made a stronger effort to receive the opinions of leadership and facilitate their involvement in the process. The MEAL COP Lead's belief was increasing their involvement would have help build the organization's investment in the results and possibly lead to more opportunities and resources for capacity building. I believe one of the reasons the MEAL COP felt this so acutely at the end of the process was the assessment had highlighted items for improvement beyond the scope of the MEAL COP's influence, like overall funding, connection to public outreach, communications, and culture. One of the benefits of using a comprehensive organizational assessment, rather than only looking at the skills and abilities to perform evaluation, is to understand the dependencies and influences across the organization. The MEAL COP planned on presenting the findings to leadership but knew that if they have been involved in the process of planning the assessment, considering the model and dependencies prior to viewing results, they likely would be more amenable to bigger investments in capacity building, especially where the dependencies are a step away from the practice of evaluation.

The first phase of my theorized framework has one step: to determine the goal for the assessment. This may seem self-evident, but the chances of the assessment's success likely depend on the details of the goal statement. I suggest practitioners make clear what outcomes the assessment should achieve, the target audiences, the necessary

stakeholders, and resources needed to undertake the process at the outset. Ultimately, it will be difficult to target the right staff for the sample, ask the right questions or contextualize the instrument, and interpret the data without clarity from the goal. As noted in the case, the complexity of the organization led to difficulties modifying and writing items and interpreting the results. The complexity spanned a range of factors from the structure of the organization, the division of specialties/roles, types of programming, and diverse practice of evaluation. The primary goal for the case organization, in Chapter 5, was to improve evaluation policies at an aggregate level of the organization and therefore they needed a broad sample. Their assumption was the diversity of office and role responses would aggregate up to an accurate picture of the organization. Although the organization achieved its primary goal, some of the challenges in interpreting the data led them to believe future use of the tool would likely have a more bounded sample and targeted questions. A detailed goal statement, with investment from leadership, will ideally allow future practitioners to have greater success with proceeding steps in the application framework, like determining the sample, considering context for tool creation or refinement, and planning the use of the results.

With respect to instrument refinement or creation, the case (Chapter 5) demonstrated a missing step from the organization's process was to consider benchmarks for responses to specific items. This is most evident when item scores within a factor ordered based on their "maturity." In other words, items that are foundational components of MEAL practice scored higher (e.g., monitor programs, develop output indicators) and more advanced practices unsurprisingly scored lower (e.g., involve peer

reviews, produce reports for the board). Without forethought on what appropriate scores for those items should be, the MEAL COP struggled to understand if they should focus on foundational components that scored high but weren't perfect, or the advanced items that had more room for improvement but were "secondary" uses of evaluation or less elemental to MEAL practice. Accordingly, I don't believe practitioners need to set strict parameters for what adequate scores should be for each item but discussing expectations and priorities before implementation of the instrument would likely benefit interpretation of the data and decision-making on use of the results. Of course, benchmarking from other cases would greatly assist in this challenge, but for now the data does not exist in the literature.

Lastly, as noted in the case (Chapter 5), a few MEAL COP members suggested an indirect yet meaningful benefit from the study was the introduction of the evaluation capacity model. They believe the model and overall assessment process helped them better understand the inputs to MEAL practice from the rest of the organization, as well as the impacts staff outside of the MEAL team can have on evaluation use. Based on the results, they want to build greater connections with dependencies in the organization, like the resource development team. Although the survey was open to staff outside of MEAL to participate, there wasn't a dissemination of the process and explanation of the theory behind the assessment. The MEAL COP will now undertake that process, however, earlier engagement of groups like the resource development team may have led to clearer use of the results and greater collaboration in the capacity building efforts post-analysis. Accordingly, I believe a campaign to disseminate the process and theory of the

assessment to a wider audience within the organization may encourage larger participation in the survey and build greater awareness of how staff outside of programs and MEAL can impact the organization's evaluation practice and its effectiveness. In short, the process use benefits of an organizational evaluation capacity assessment can go beyond staff who directly work on MEAL.

Limitations

The most significant limitation of the dissertation is the representativeness of the samples in each study. In sub-study 1, I believe the primary risk to having adequate representativeness in the sample is missing data from practitioners. The inclusion of grey literature in the review was critical given evaluation capacity building is a common practice yet academic literature on organizational evaluation capacity measurement is minimal. Of the 16 articles included in the review, 13 are from peer-reviewed journals and it is likely there are other operating models used by evaluation capacity building experts, consultants, and internal evaluators. Additionally, the samples used in the majority of the studies are problematic for generalization. The sampling method for most of the case studies is purposeful sampling, and convenience sampling within a set of shared entities for the larger survey studies (e.g., multiple departments residing in the same government office or nonprofits in an association). Accordingly, the representativeness of the models may have a bias towards the type of organizations willing to participate in the research and therefore more experience in evaluation practice.

For sub-study 2, the major limitation is the extent conclusions and generalizations can be drawn from a single case. All organizations, especially large organizations, are

going to have unique factors from their structure, leaders, culture, funders, life-cycle, and overall mission. Accordingly, the lessons from the case study may have unique properties that another organization may not encounter. Furthermore, the case study did not have a survey sample large enough to make direct quantitative comparisons to the model existing in academic literature using structural equation modeling. The organization was satisfied with the sample size, based on their experience with similar internal data collection initiatives, but I likely needed a survey of 400-500 respondents to replicate the structural equation model with 101 items.

Another meaningful limitation of this study was my inability to engage the organization in-person. I first engaged the organization in the early months of 2020, right before Covid-19 quarantines ramped up in the United States. Accordingly, it was my only opportunity to meet at their office, discuss the assessment process in-person, and observe MEAL practitioners and their meetings. The organization, like all others, had to evolve during the pandemic, and did not pick up the project again until the summer 2021. I remain immensely grateful the organization proceeded in a virtual context, but also know my description of the process may have been richer in a non-pandemic time.

Suggestions for Future Research

Numerous opportunities to build upon the evidentiary base for organizational evaluation capacity assessments exist. Proposed here are four avenues of research: (1) confirming and replicating models found in the literature; (2) search for variation of models in different organizational demographics; (3) case studies on application of models; and (4) develop applied tools for organizations based on models.

As detailed in Chapter 4, the validity base for the models of organizational evaluation capacity is limited. Some models developed over time, like Cousins and Bourgeois' work from 2004 until present day, but all the quantitively developed models use a single sample and do not yet have a study replicating fit. Prime targets for replication are three models that have structural relationships and model fit indices (Nielsen et al, 2011; Taylor-Ritzler et al., 2013; Gagnon et al., 2018). Additionally, some of the instruments used may be challenging for organizations to complete due to length. For example, the Gagnon et al. (2018) modified by the organization in the case study includes 119 items. A major contribution to a model's applied appeal could be its structural replication despite the reduction in the number of items in the survey.

Secondly, the structural models with dimensional relationships are great candidates to assess invariance. In the process of new sample generation and replication, models could test for invariance across organizational factors. Human service organizations have multiple characteristics that could test for invariance: revenue size, staff size, sector, organizational life cycle, funding type, and program beneficiaries. The research could help advance an understanding of the extent of transferability of models, as well as specific dimensions that may differ given an organizational characteristic. However, this could be difficult given the need to contextualize the tools to fit organizational contexts.

Third, I have promoted other scholars suggesting the field needs case studies of organizations undertaking organizational capacity assessments. General case studies similar to what I have presented in Chapter 5 will remain valuable given the contextual

realities for organizations. However, another idea would be repeated measurements and longitudinal studies of an organization using the models to measure their evaluation capacity change over time. Researchers could undertake case studies to get a baseline of organizational capacity building, and then target specific areas for evaluation capacity building initiatives. After enough time had passed, the researcher could administer the instrument again to measure the change in evaluation capacity. The use of pre-intervention and post-intervention measurement would further the evidence that the models are useful in assessing targeted areas for capacity building and properly reflect the growth if changes have occurred.

Lastly, it would be beneficial to the field if the three models that have evidence of construct validity, along with some of the frameworks with content validity from case study analysis, became accessible as tools for evaluation capacity building practitioners. Using the examples of the checklist from Volkov and King (2007), Stufflebeam's (2002) checklist for institutionalizing evaluation, and the ROLE instrument from Preskill and Boyle (2008), researchers creating tools and instruments to bridge the gap from the academic literature to evaluation capacity practitioners would increase learning. The ability for organizations to use the models without needing to comprehend confirmatory factor analysis or structural equation modeling would add to the mainstreaming of the practice of evaluation capacity measurement.

My hope is this dissertation inspires future literature creation to guide and direct organizations on the process, challenges, and benefits of measuring organizational evaluation capacity.

References

Al Hudib, H. & Cousins, J. B. (2021). Understanding evaluation policy and organizational capacity for evaluation: An interview study. *The American Journal of Evaluation*, *43*(2), 234-254.

Al Hudib H., & Cousins, J. B. (2022). Evaluation policy and organizational evaluation capacity building: A study of international aid agency evaluation policies. *New Directions for Evaluation*, *2022*(173), 29–48.

Alkin, Marvin. (2013). *Evaluation roots: A wider perspective of theorists' views and influences.* Sage Publications, Inc.

Baizerman, M., Compton, D. W., & Stockdill, S. (2002). New directions for ECB. *New Directions for Evaluation*, *2002*(93), 109–120.

Bourgeois, I., & Cousins, J. B. (2008). Informing evaluation capacity building through profiling organizational capacity for evaluation: an empirical examination of four Canadian federal government organizations. *The Canadian Journal of Program Evaluation*, *23*(3), 127.

Bourgeois, I., & Cousins, J. B. (2013). Understanding dimensions of organizational evaluation capacity. *American Journal of Evaluation*, *34*(3), 299–319.

Bourgeois, I., Whynot, J., & Thériault, É. (2015). Application of an organizational evaluation capacity self-assessment instrument to different organizations: Similarities and lessons learned. *Evaluation and Program Planning*, *50*, 47–55.

Browne M.W., Cudeck R. (1993). Alternative ways of assessing model fit. In: Bollen K., Long J. (1993). *Testing Structural Equation Models*. Sage Publications.

Bryan, T.K., Robichau, R. W., & L'Esperance, G. E. (2021). Conducting and utilizing

evaluation for multiple accountabilities: A study of nonprofit evaluation capacities. *Nonprofit Management & Leadership, 31*(3), 547–569.

Carman, J. G. (2011). Understanding evaluation in nonprofit organizations. *Public Performance & Management Review*, *34*(3), 350–377.

Carman, J.G., Fredricks, K.A. (2008). Nonprofits and evaluation: Empirical evidence from the field. *New Directions for Evaluation, 119*, 51–71.

Cheng, S.H., & King, J. A. (2017). Exploring organizational evaluation capacity and evaluation capacity building: a Delphi study of Taiwanese elementary and junior high schools. *American Journal of Evaluation*, *38*(4), 521–539.

Cousins, J. B., & Bourgeois, I. (2014). Cross-case analysis and implications for research, theory, and practice. *New Directions for Evaluation*, *2014*(141), 101–119.

Cousins, J. B., & Bourgeois, I. (2014). Multiple case study methods and findings. *New Directions for Evaluation*, *2014*(141), 25–99.

Cousins, J. B., Goh, S. C., Clark, S., & Lee, L. E. (2004). Integrating evaluative inquiry into the organizational culture: A review and synthesis of the knowledge base. *The Canadian Journal of Program Evaluation*, *19*(2), 99–141.

Cousins, J. B., Goh, S. C., Elliott, C. J., & Bourgeois, I. (2014). Framing the capacity to do and use evaluation. *New Directions for Evaluation*, *2014*(141), 7–23.

Creswell, J. (2007). *Qualitative inquiry & research design: Choosing among five approaches* (2nd ed.). Thousand Oaks: Sage Publications.

Creswell, J., & Plano Clark, V. (2018). *Designing and conducting mixed methods research* (Third edition.). SAGE Publications.

Despard, M. R. (2016). Strengthening evaluation in nonprofit human service

organizations: Results of a capacity-building experiment. *Human Service Organizations: Management, Leadership & Governance*, *40*(4), 352–368.

El Hassar, B., Poth, C., Gokiert, R., & Bulut, O. (2021). Toward an evidence-based approach to building evaluation capacity. *Canadian Journal of Program Evaluation*, *36*(1), 82–94.

Fierro, L., & Christie, C. (2016). Evaluator and program manager perceptions of evaluation capacity and evaluation practice. *American Journal of Evaluation 38*(3), 376-392.

Floyd, F., & Widaman, K. (1995). Factor analysis in the development and refinement of clinical assessment instruments. *Psychological Assessment, 7*(3), 286–299.

Gagnon, F., Aubry, T., Cousins, J. B., Goh, S. C., & Elliott, C. (2018). Validation of the evaluation capacity in organizations questionnaire. *Evaluation and Program Planning*, 68, 166-175.

Hahs-Vaughn, D.L. (2017). *Applied Multivariate Statistical Concepts.* Taylor & Francis Group.

Higgins, J.P.T, Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M.J., Welch, V.A. (2020) *Cochrane handbook for systematic reviews of interventions* (version 6.1). Cochrane, 2020. Available from www.training.cochrane.org/handbook.

Holmes, A.G.D. (2020). Researcher positionality: A consideration of its influence and place in qualitative research: A new researcher guide. *Shanlax International Journal of Education*. 8(4). 1-10.

King, Jean A. (2020). Putting evaluation capacity building in context: Reflections on the Ontario Brain Institute's evaluation support program. *Evaluation and Program*

Planning, 80, 101452–101456.

King, J. A., & Volkov, B. (2005). A framework for building evaluation capacity based on the experiences of three organizations. *CURA Reporter*, 10–16.

Kvale, S. (2007). *Doing interviews*. SAGE Publications.

Labin, S. (2014). Developing common measures in evaluation capacity building: An iterative science and practice process. *American Journal of Evaluation, 35*(1), 107-115.

Labin, S., Duffy, J., Meyers, D., Wandersman, A., & Lesesne, C. (2012). A research synthesis of the evaluation capacity building literature. *American Journal of Evaluation, 33*(3), 307-338.

Lawrenz, Kollmann, E. K., King, J. A., Bequette, M., Pattison, S., Nelson, A. G., Cohn, S., Cardiel, C. L., Iacovelli, S., Eliou, G. O., Goss, J., Causey, L., Sinkey, A., Beyer, M., & Francisco, M. (2018). Promoting evaluation capacity building in a complex adaptive system. *Evaluation and Program Planning, 69*, 53–60.

Lynch-Cerullo, K., & Cooney, K. (2011). Moving from outputs to outcomes: A review of the evolution of performance measurement in the human service nonprofit sector. *Administration to Social Work, 35*, 364-388.

Nakaima, & Sridharan, S. (2020). Reflections on experiential learning in evaluation capacity building with a community organization, Dancing with Parkinson's. *Evaluation and Program Planning, 80*, 101441–101444.

Mark, M. M., Henry, G.T., & Julnes, G. (2000). *Evaluation: An integrated framework for understanding, guiding, and improving policies and programs*. Jossey-Bass Inc.

McDonald, B., Rogers, P., & Kefford, B. (2003). Teaching people to fish? Building the

evaluation capability of public sector organizations. *Evaluation*, *9*(1), 9–29.

Milstein, B., Chapel, T. J., Wetterhall, S. F., & Cotton, D. A. (2002). Building capacity for program evaluation at the Centers for Disease Control and Prevention. *New Directions for Evaluation*, *2002*(93), 27–46.

Mitchell, G. E., & Berlan, D. (2018). Evaluation in nonprofit organizations: An empirical analysis. *Public Performance & Management Review*, *41*(2), 415–437.

Naccarella, L., Pirkis, J., Kohn, F., Morley, B., Burgess, P., & Blashki, G. (2007). Building evaluation capacity: definitional and practical implications from an Australian case study. *Evaluation and Program Planning*, *30*(3), 231–236.

Nielsen, S. B., Lemire, S., & Skov, M. (2011). Measuring evaluation capacity results and implications of a Danish study. *American Journal of Evaluation*, *32*(3), 324–344.

Pajo, B. (2018). *Introduction to Research Methods: A Hands-On Approach*. Los Angeles: SAGE Publications.

Preskill, H. (2014). Now for the hard stuff: Next steps in ECB research and practice. *American Journal of Evaluation*, *35*(1), 116–119.

Preskill, H., & Boyle, S. (2008). A multidisciplinary model of evaluation capacity building. *American Journal of Evaluation*, *29*(4), 443–459.

Stake, R.E. (1995). *The Art of Case Study Research*. Thousand Oaks, CA: Sage Publications.

Readiness for Organizational Learning and Evaluation Instrument (ROLE). (2011). Retrieved from: https://www.fsg.org/tools-and-resources/readiness-organizational-learning-and-evaluation-instrument-role

Stufflebeam, D. (2002). Institutionalizing evaluation checklist. The Evaluation Center,

Western Michigan University. Retrieved from: http://www.wmich.edu/evalctr/checklists/institutionalizingeval.htm

Stufflebeam, D., & Shinkfield, A. (2007). *Evaluation theory, models, and applications*. San Francisco: Jossey-Bass.

Taylor-Powell, E., & Boyd, H. H. (2008). Evaluation capacity building in complex organizations. *New Directions for Evaluation*, *2008*(120), 55–69.

Taylor-Ritzler, T., Suarez-Balcazar, Y., Garcia-Iriarte, E., Henry, D. B., & Balcazar, F. E. (2013). Understanding and measuring evaluation capacity: a model and instrument validation study. *American Journal of Evaluation*, *34*(2), 190–206.

Trochim, William. (2009). Evaluation policy and evaluation practice. *New Directions for Evaluation*, (123), 13–32.

Volkov, B., & King, J. A. (2007). A checklist for building organizational evaluation capacity. Retrieved from https://wmich.edu/evaluation/checklists

Wade, & Kallemeyn, L. (2020). Evaluation capacity building (ECB) interventions and the development of sustainable evaluation practice: An exploratory study. *Evaluation and Program Planning, 79*, 101777–101779.

Weiss, Carol H., (1998). *Evaluation: Methods for studying programs and policies*. Prentice Hall.

Yin, R. (2014). *Case study research: Design and methods (Fifth edition).* SAGE: Thousand Oaks.

Appendices A-F: IRB Documents, Case Data, and Interview Protocols

Appendix A: Human Subjects Institutional Review Board Letter

UNIVERSITY of DENVER
OFFICE OF RESEARCH & SPONSORED PROGRAMS
Research Integrity & Education

DATE:	August 9, 2021
TO:	Ryan Smyth
FROM:	University of Denver (DU) IRB
PROJECT TITLE:	[1782202-1] Application of an Organizational Evaluation Capacity Assessment in a Multinational NGO: A Mixed Methods Case Study to Support Practice
SUBMISSION TYPE:	New Project
ACTION:	**EXEMPTION GRANTED**
DECISION DATE:	August 9, 2021
NEXT REPORT DUE:	*August 9, 2022*
RISK LEVEL:	Minimal Risk
REVIEW CATEGORY:	Exemption category # *2*
	Exemption 2: Educational Tests, Surveys, Interviews, or Observations

Thank you for your submission of the materials for this project. The University of Denver IRB determined this project qualtifies under an **Exempt category** according to federal regulations. This exemption was granted based on appropriate criteria for granting an exemption and a study design wherein the risks have been minimized.

Please note that maintaining exempt status requires that (a) risks of the study remain minimal; (b) that anonymity or confidentiality of participants, or protection of participants against any increased risk due to the internal knowledge or disclosure of identity by the researcher, is maintained as described in the application; (c) that no deception is introduced, such as reducing the accuracy or specificity of information about the research protocol that is given to prospective participants; (d) the research purpose, sponsor, and recruited study population remain as described; and (e) the principal investigator (PI) continues and is not replaced.

If changes occur in any of the features of the study as described, this may affect one or more of the conditions of exemption and may warrant a reclassification of the research protocol from exempt and require additional IRB review. For the duration of your research study, any changes in the proposed study must be reviewed by the University of Denver IRB before implementation of those changes.

Informed Consent Process

Informed consent is an important process when conducting human subject research beginning with providing potential subjects with a description of the project and assurance of a participants understanding. The DU IRB has granted this project exempt status with the use of a **Verbal Consent Script**. Informed consent must continue throughout the project via the use of the approved Verbal Consent. If requested, each participant is entitled to receive a copy of the Exempt Information Letter/ Consent script.

Unanticipated Problems Involving Risks to Subjects or Others (UPIRTSOs)

Any incident, experience or outcome which has been associated with an unexpected event(s), related or possibly related to participation in the research, and suggests that the research places subjects or others at a greater risk of harm than was previously known or suspected must be reported to the IRB. UPIRTSOs may or may not require suspension of the research. Each incident is evaluated on a case by case basis to make this determination. The IRB may require remedial action or education as deemed necessary for the investigator or any other key personnel. The investigator is responsible for reporting UPIRTSOs to the IRB within 5 working days after becoming aware of the unexpected event. Use the Reportable New Information (RNI) form within the IRBNet system to report any UPIRTSOs. All NON-COMPLIANCE issues or COMPLAINTS regarding this project must also be reported.

Continuation Review Requirements

Based on the current regulatory requirements, this exempt project does **not** require continuing review. However, this project has been assigned a **one-year review period** requiring communication to the IRB at the end of this review period to either close the study or request an extension for another year. The one-year review period will be posted in the Next Report Due section on the Submission Details page in IRBNet. During this one-year period, a staff member from the Office of Research Integrity and Education (ORIE) may also conduct a Post Approval Monitoring visit to evaluate the progress of this research project.

Study Completion and Final Report

A Final Report must be submitted to the IRB, via the IRBNet system, when this study has been completed. The DU HRPP/IRB will retain a copy of the project document within our records for three years after the closure of the study. The Principal Investigator is also responsible for retaining all study documents associated with this study for at least three years after the project is completed.

PLEASE NOTE: This project will be administratively closed at the end of the one-year period unless a request is received from the Principal Investigator to extend the project. Please contact the DU HRPP/IRB if the study is completed before the one-year time period or if you are no longer affiliated with the University of Denver through submitting a Final Report to the DU IRB via the IRBNet system. If you are no longer affiliated with DU and wish to transfer your project to another institution please contact the DU IRB for assistance.

If you have any questions, please contact the DU Institutional Review Board (IRB) at (303) 871-2121 or at IRBAdmin@du.edu. Please include your project title and IRBNet number in all correspondence with the IRB.

This letter has been electronically signed in accordance with all applicable regulations, and a copy is retained within University of Denver (DU) IRB's records.

Appendix B: IRB Approved Verbal Consent F

Version Date: 8/09/2021

VERBAL CONSENT SCRIPT

Introduction
My name is Ryan Smyth, and I am a student from the Morgridge College of Education at the University of Denver. I am emailing you to request participation in my research study. This study concerns the evaluation capacity of your organization and builds on the recent survey you completed for the MEAL COP. I obtained your contact information from your organization's MEAL COP lead.

Subjects Rights
Your participation in this research study is completely voluntary. You can withdraw at any time. Choosing not to be in this study or to stop being in this study will not result in any penalty to you. Your choice to not be in this study will not negatively affect any rights to which you are otherwise entitled, including your present or future employment at your organization.

Description of the study and study procedures
I am conducting the research study to understand the process and use of an organizational evaluation capacity assessment in a multinational NGO. The research builds on the recent work and survey provided by the MEAL COP. The name of the study is Application of an Organizational Evaluation Capacity Assessment in a Multinational NGO: A Mixed Methods Case Study to Support Practice. The IRB Project Number is 1782202-1.

If you agree to participate, we will discuss your interpretation of the survey results and how they can be used to plan evaluation capacity building initiatives. The interview will be about 30 minutes in length. With your permission, I will record the audio from our discussion to transcribe and use for analysis, through aggregation with other interviews.

Risks and Benefits
The risks from participating in this study may include indirect identification of your interview statements in a final narrative. However, all statements will not contain direct identifiers and all data will be stored privately and securely. Taking part in this study will help your organization understand how to use the data and make capacity building investments in the future and may help researchers to better understand how similar assessments should be implemented in the field to build evaluation capacity at other NGOs.

Confidentiality
Names and email addresses will not be captured in the formal data collection process, only used to connect online through Zoom links in emails. Email I have received listing participants will be deleted after the completion of the research. Interviewees will be identified in the data collection process by their job title if generic, or technical area (evaluation specialist 1, program specialist 1, etc.) if the job title is unique. Consent forms with identifiable information will be stored separately from the interview data collection. All folders containing interview data, including recordings, will be password protected on my DU OneDrive account, and destroyed upon the completion of the dissertation.

The results of the research study may be published, but your name will not be used.

Whom to contact with questions
If you have any questions or problems during your time on this study, you should contact my advisor, Robyn Thomas Pitts, Assistant Professor at the Morgridge College of Education, at Robyn.Pitts@du.edu.

If you have any questions regarding your rights as a research subject, please contact the the University of Denver's Institutional Review Board (IRB) Office at (303)871-2121.

Appendix C: Contextualization of the Instrument

Table C1

Comparison of Items between the Contextualized and Original Surveys

Organizational Learning Items	Comparison to Original Survey				Original Items
	Minimal or No Change	Modified	New	Removed	
Staff can often bring new ideas to improve programs.		x			We can often bring new ideas into the organization.
Failures are constructively discussed.	x				
Current practice in teams encourage staff to solve problems together before discussing them with a manager.		x			Current organizational practice encourages staff to solve problems together before discussing them with a manager.
People who are new are encouraged to question the way things are done.	x				
Senior managers accept change and are not afraid of new ideas.	x				
Managers encourage staff to experiment in order to improve work processes.	x				
New work processes that may be useful are usually shared with all staff.	x				
Senior managers and staff share a common vision of what our work should accomplish.	x				
Managers frequently involve staff in important decisions.	x				
We can usually create informal groups to solve organizational problems.	x				
Managers can accept criticism without becoming overly defensive.	x				
We have a system that allows us to learn successful practices from other organizations.	x				
Managers often provide useful feedback that helps to identify potential problems and opportunities.	x				
We have opportunities for self-assessment with respect to goal attainment.	x				
Most problem-solving groups include staff from a variety of functional areas or divisions.	x				
Employees are given sufficient time to reflect on organizational successes or failures.	x				
				x	Innovative ideas that are successful are rewarded by management.

				x	New ideas from staff are treated seriously by management.
				x	The organization's mission statement identifies values to which all staff must conform.
				x	This organization supports the development of skills such as leadership, coaching, and teambuilding among staff.
				x	We are rewarded for using performance information.

Organizational Support Items	**Comparison to Original Survey**				**Original Items**
	Minimal or No Change	Modified	New	Removed	
Staff are provided with work-related skills training.	x				
Information and decision-making on programmatic issues must always go through proper/formal channels.		x			Information and decision-making must always go through proper channels.
The skills training that we receive can be applied to improve our work.	x				
Staff are encouraged to continuously upgrade and increase their technical knowledge and education levels.	x				
Supports the development of soft skills such as leadership, coaching, and team building among staff.	x				
Learning that improves the work skills and technical knowledge of staff is encouraged.	x				
Staff training is emphasized equally at all levels.	x				
Training is done in teams where appropriate.	x				
Our work is usually closely supported, monitored, and reviewed by management.	x				
				x	Standard operating procedures have been established for almost every work situation.
				x	Staff training is relevant to our work.

Capacity to Do Evaluation Items	**Comparison to Original Survey**				**Original Items**
	Minimal or No Change	Modified	New	Removed	
We have long-term, dedicated financial support to ensure evaluation activities across all programming where evaluation is required.	x				
We are provided with the basic tools/resources to support evaluation (e.g., computers, software, copying, administrative support).	x				
We have a "champion" on staff who supports our evaluation efforts and advocates on our behalf when required.	x				
Our office possesses the technical competencies to conduct evaluations (e.g., instrument development, data collection and analysis).	x				

Our office has the knowledge and skills to oversee evaluations performed by external evaluators.	x				
Our office provides positive encouragement to conduct evaluation.	x				
We have formal requirements to report on programmatic performance.		x			We have formal requirements to report on performance.
				x	We are able to easily access information about "best practices" within our field.
				x	Performance measurement is integral to our organizational accountability framework.

Evaluative Inquiry Items	**Comparison to Original Survey**				**Original Items**
	Minimal or No Change	Modified	New	Removed	
Review program documentation (e.g., participant records, case notes).	x				
Conduct firsthand observation of program activities.	x				
Conduct formal program evaluations.	x				
Establish performance targets (e.g., serve 200 people, 80% complete training).	x				
Monitor implementation (i.e., ensure that programs are delivered as intended).	x				
Monitor program outcomes (i.e., ensure program results are as intended).	x				
Assess the degree to which program goals/objectives are met.	x				
Engage in formal evaluation planning processes.	x				
Employ single-case mixed-method designs (i.e., qualitative and quantitative methods).	x				
Use program theoretical designs (i.e., link program components to outcomes).	x				
Produce annual and/or reports based on performance measures.	x				
Produce reports about program activities.	x				
Produce reports for Boards of Directors and/or senior management.	x				
Use a performance measurement system.	x				
Use program logic models.	x				
Use other evaluation or management systems (e.g., performance audits, results-based management, quality assurance activities).	x				

Stakeholder Participation (Frequency and Participation) Items	**Comparison to Original Survey**				**Original Items**
	Minimal or No Change	Modified	New	Removed	

Program/project developers (team that designed the program).	x				
Program/project managers or directors.	x				
Program/project sponsors, donors or funders.	x				
Staff responsible for implementing the program/project.	x				
Participants of the impact group (well-defined group of individuals seeking to facilitate lasting change or impact) of the program/project.		x			Intended beneficiaries of the program.
Special interest groups or target groups (individuals whose actions or behaviors generate outcomes) of the program/project.		x			Special interest groups.
Partner organizations.			x		
Stakeholders representing different identities (e.g., race, disability, age, migration status, etc.)			x		

Evaluation Findings Use Items	**Comparison to Original Survey**				**Original Items**
	Minimal or No Change	Modified	New	Removed	
Make changes to existing programs.	x				
Conduct strategic planning at the organizational level.	x				
Get new funding.	x				
Justify program existence or continuation.	x				
Make decisions about staffing (e.g., in the program being evaluated or in the org as a whole).		x			Make decisions about staffing.
Report to the board or senior management (or equivalent).	x				
Perform outreach and public relations.	x				
Meet external accountability requirements.	x				
Inform and expand advocacy and influencing work.			x		

Process Use Items	**Comparison to Original Survey**				**Original Items**
	Minimal or No Change	Modified	New	Removed	
Develop knowledge about evaluation logic.	x				
Develop knowledge about evaluation methods.	x				
Develop technical skills for doing evaluation (e.g., instrument development, data collection and analysis).	x				
Develop a better understanding of the program/policy/intervention being evaluated.	x				

Integrate evaluation into our work practices.	x				
Improve management practices in the office.	x				
Expedite/intensify/expand the use of evaluation findings.		x			Expedite the use of evaluation findings.
Develop professional networks.	x				
Question underlying assumptions about what we do.	x				
Develop a mindset of evaluative thinking.	x				
Increase commitment to the program/policy/intervention being evaluated.	x				
Increase organizational commitment.	x				
Increase ownership of what we do.	x				
Appreciate the value of evaluation.	x				
Appreciate the power of evaluation as a force for change.	x				
				x	Foster a shared understanding of organizational functioning.
				x	Expand the use of evaluation findings.

Mediating Conditions Items	**Comparison to Original Survey**				**Original Items**
	Minimal or No Change	Modified	New	Removed	
Foster improvement in program implementation quality and/or outcomes.		x			Foster improvement.
High quality.	x				
Provide accurate results.	x				
Reduce surprises for decision makers.	x				
Perceived by users (e.g., program managers, technical specialists) as unbiased.	x				
Informed by user input.	x				
Supported by quality assurance mechanisms (e.g., peer reviews).	x				
Based on objective data (e.g., based on facts, robustly measured)	x				
Can be compared against external standards or benchmarks.	x				
Produced by evaluators who are perceived as credible.	x				
Perceived to be appropriately resourced.	x				
Use a methodology that is understood by users.	x				

Accessible to all staff members.			x		Accessible to senior management.

Subtotals for All Dimensions	**Comparison to Original Survey**			
	Minimal or No Change	Modified	New	Removed
Subtotal for Organizational Learning Items	14	2	0	5
Subtotal for Organizational Support Items	8	1	0	2
Subtotal for Capacity to Do Evaluation Items	6	1	0	2
Subtotal for Evaluation Inquiry Items	16	0	0	0
Subtotal for Stakeholder Participation (Frequency and Participation) Items	4	2	2	0
Subtotal for Evaluation Findings Use Items	7	1	1	0
Subtotal for Process Use Items	14	1	0	2
Subtotal for Mediating Conditions Items	11	2	0	0
Totals	**80**	**10**	**3**	**11**

Appendix D: Survey Descriptive Statistics and Response Breakdown

Table D1

Survey Results for the Organizational Learning (OL) Items

Organizational Learning Items	Descriptive Statistics			Response Percentages				
	n	*M*	*SD*	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Staff can often bring new ideas to improve programs.	134	4.172	0.678	0%	1%	11%	56%	31%
Failures are constructively discussed.	134	3.597	0.833	1%	8%	29%	51%	10%
Current practice in teams encourage staff to solve problems together before discussing them with a manager.	134	3.537	0.810	1%	11%	28%	53%	7%
People who are new are encouraged to question the way things are done.	134	3.642	0.879	1%	10%	27%	48%	14%
Senior managers accept change and are not afraid of new ideas.	134	3.724	0.844	0%	9%	26%	49%	16%
Managers encourage staff to experiment in order to improve work processes.	134	3.694	0.768	1%	5%	29%	54%	11%
New work processes that may be useful are usually shared with all staff.	133	3.812	0.845	2%	7%	17%	58%	17%
Senior managers and staff share a common vision of what our work should accomplish.	134	3.672	0.865	2%	7%	25%	53%	13%
Managers frequently involve staff in important decisions.	134	3.545	0.923	1%	16%	24%	48%	12%
We can usually create informal groups to solve organizational problems.	134	3.507	0.979	3%	13%	26%	45%	13%
Managers can accept criticism without becoming overly defensive.	134	3.328	0.839	0%	17%	40%	37%	7%
We have a system that allows us to learn successful practices from other organizations.	134	2.978	1.000	6%	29%	30%	31%	4%
Managers often provide useful feedback that helps to identify potential problems and opportunities.	134	3.746	0.690	1%	4%	22%	66%	7%
We have opportunities for self-assessment with respect to goal attainment.	134	3.918	0.661	0%	1%	22%	60%	16%
Most problem-solving groups include staff from a variety of functional areas or divisions.	134	3.410	0.983	4%	15%	28%	44%	10%
Employees are given sufficient time to reflect on organizational successes or failures.	134	3.007	0.962	5%	27%	33%	32%	3%
Total		**3.580**	**0.900**	**2%**	**11%**	**26%**	**49%**	**12%**

Note: For observed means, 1 = Strongly Disagree, 2 = Disagree, 3 = Neutral, 4 = Agree, 5 = Strongly Agree.

Table D2

Survey Results for the Organizational Support (OS) Items

Organizational Support Items	Descriptive Statistics			Response Percentages				
	n	*M*	*SD*	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Staff are provided with work-related skills training.	133	3.526	0.934	2%	13%	26%	47%	11%
Information and decision-making on programmatic issues must always go through proper/formal channels.	133	3.774	0.893	3%	5%	20%	56%	17%
The skills training that we receive can be applied to improve our work.	133	3.940	0.786	1%	5%	16%	58%	21%
Staff are encouraged to continuously upgrade and increase their technical knowledge and education levels.	133	3.707	0.868	2%	8%	25%	51%	15%
Supports the development of soft skills such as leadership, coaching, and team building among staff.	133	3.398	0.984	3%	15%	33%	37%	12%
Learning that improves the work skills and technical knowledge of staff is encouraged.	133	3.767	0.777	1%	6%	22%	59%	13%
Staff training is emphasized equally at all levels.	133	3.414	0.914	1%	18%	29%	43%	9%
Training is done in teams where appropriate.	133	3.541	0.713	0%	6%	41%	47%	7%
Our work is usually closely supported, monitored, and reviewed by management.	133	3.669	0.927	2%	10%	22%	51%	15%
Total		**3.637**	**0.884**	**2%**	**9%**	**26%**	**50%**	**13%**

Note: For observed means, 1 = Strongly Disagree, 2 = Disagree, 3 = Neutral, 4 = Agree, 5 = Strongly Agree.

Table D3

Survey Results for the Capacity to Do Evaluation (CTD) Items

Capacity to Do Evaluation Items	Descriptive Statistics			Response Percentages				
	n	*M*	*SD*	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
We have long-term, dedicated financial support to ensure evaluation activities across all programming where evaluation is required.	133	3.023	1.158	10%	26%	28%	26%	11%
We are provided with the basic tools/resources to support evaluation (e.g., computers, software, copying, administrative support).	133	3.797	1.021	3%	9%	19%	44%	26%
We have a "champion" on staff who supports our evaluation efforts and advocates on our behalf when required.	133	3.383	0.885	2%	14%	32%	45%	6%
Our office possesses the technical competencies to conduct evaluations (e.g., instrument development, data collection and analysis).	133	3.722	0.810	1%	7%	26%	53%	14%
Our office has the knowledge and skills to oversee evaluations performed by external evaluators.	133	3.797	0.833	1%	8%	17%	58%	16%
Our office provides positive encouragement to conduct evaluation.	133	3.850	0.712	1%	1%	27%	56%	16%
We have formal requirements to report on programmatic performance.	133	4.075	0.735	1%	2%	12%	59%	26%
Total		**3.664**	**0.946**	**3%**	**10%**	**23%**	**49%**	**16%**

Note: For observed means, 1 = Strongly Disagree, 2 = Disagree, 3 = Neutral, 4 = Agree, 5 = Strongly Agree.

Table D4

Survey Results for the Evaluative Inquiry (EI) Items

Evaluative Inquiry Items	Descriptive Statistics			Response Percentages				
	n	*M*	*SD*	Never	In-frequently	Some-times	Often	Always
Review program documentation (e.g., participant records, case notes).	130	3.746	0.934	0%	10%	29%	37%	24%
Conduct firsthand observation of program activities.	128	3.672	0.981	2%	12%	25%	41%	20%
Conduct formal program evaluations.	130	3.838	0.979	1%	8%	27%	34%	30%
Establish performance targets (e.g., serve 200 people, 80% complete training).	129	4.194	0.876	2%	2%	14%	40%	43%
Monitor implementation (i.e., ensure that programs are delivered as intended).	130	4.254	0.800	1%	1%	15%	38%	45%
Monitor program outcomes (i.e., ensure program results are as intended).	130	4.138	0.851	1%	3%	16%	42%	38%
Assess the degree to which program goals/objectives are met.	130	4.131	0.781	0%	2%	18%	45%	35%
Engage in formal evaluation planning processes.	130	3.969	0.871	0%	5%	23%	41%	31%
Employ single-case mixed-method designs (i.e., qualitative and quantitative methods).	130	3.885	0.920	1%	5%	28%	37%	29%
Use program theoretical designs (i.e., link program components to outcomes).	130	3.808	0.872	1%	5%	28%	43%	22%
Produce annual and/or reports based on performance measures.	130	4.138	0.946	2%	3%	20%	31%	45%
Produce reports about program activities.	129	4.349	0.881	1%	5%	6%	33%	54%
Produce reports for Boards of Directors and/or senior management.	129	3.612	1.127	5%	8%	34%	26%	27%
Use a performance measurement system.	130	3.885	0.985	2%	6%	22%	39%	30%
Use program logic models.	130	3.954	0.971	2%	6%	22%	37%	34%
Use other evaluation or management systems (e.g., performance audits, results-based management, quality assurance activities).	129	3.597	0.972	2%	9%	33%	36%	19%
Total		**3.948**	**0.949**	**1%**	**6%**	**23%**	**37%**	**33%**

Note: For observed means, 1 = Never, 2 = Infrequently, 3 = Sometimes, 4 =Often, 5 = Always.

Table D5

Survey Results for the Stakeholder Participation Frequency (SPF) Items

Stakeholder Participation Frequency Items	Descriptive Statistics			Response Percentages				
	n	*M*	*SD*	Never	In-frequently	Some-times	Often	Always
Program/project developers (team that designed the program).	129	3.767	0.996	2%	9%	29%	34%	27%
Program/project managers or directors.	128	4.078	0.927	1%	5%	18%	37%	39%
Program/project sponsors, donors or funders.	125	3.184	1.103	4%	27%	29%	26%	14%
Staff responsible for implementing the program/project.	128	4.250	0.878	1%	2%	20%	28%	50%
Participants of the impact group (well-defined group of individuals seeking to facilitate lasting change or impact) of the program/project.	129	3.806	1.146	4%	10%	23%	27%	36%
Special interest groups or target groups (individuals whose actions or behaviors generate outcomes) of the program/project.	127	3.551	1.029	3%	11%	33%	33%	20%
Partner organizations.	127	3.724	1.005	2%	11%	25%	38%	24%
Stakeholders representing different identities (e.g., race, disability, age, migration status, etc.)	127	3.386	1.024	5%	13%	35%	34%	13%
Total		**3.721**	**1.064**	**3%**	**11%**	**26%**	**32%**	**28%**

Note: For observed means, 1 = Never, 2 = Infrequently, 3 = Sometimes, 4 =Often, 5 = Always.

Table D6

Survey Results for the Stakeholder Participation Level (SPL) Items

Stakeholder Participation Level Items	Descriptive Statistics			Response Percentages		
	n	*M*	*SD*	Low	Medium	High
Program/project developers (team that designed the program).	127	2.315	0.687	13%	43%	44%
Program/project managers or directors.	128	2.539	0.614	6%	34%	60%
Program/project sponsors, donors or funders.	122	1.869	0.727	34%	46%	20%
Staff responsible for implementing the program/project.	128	2.617	0.549	3%	32%	65%
Participants of the impact group (well-defined group of individuals seeking to facilitate lasting change or impact) of the program/project.	123	2.317	0.717	15%	39%	46%
Special interest groups or target groups (individuals whose actions or behaviors generate outcomes) of the program/project.	125	2.120	0.725	21%	46%	33%
Partner organizations.	126	2.135	0.697	18%	50%	32%
Stakeholders representing different identities (e.g., race, disability, age, migration status, etc.)	122	1.943	0.719	29%	48%	23%
Total		**2.236**	**0.723**	**17%**	**42%**	**41%**

Note: For observed means, 1 = Low, 2 = Medium, 3 = High.

Table D7

Survey Results for the Evaluation Findings Use (EFU) Items

Evaluation Findings Use Items	Descriptive Statistics			Response Percentages				
	n	*M*	*SD*	Never	In-frequently	Some-times	Often	Always
Make changes to existing programs.	130	3.600	0.859	1%	8%	35%	42%	14%
Conduct strategic planning at the organizational level.	127	3.528	1.060	4%	13%	27%	38%	18%
Get new funding.	128	3.484	0.896	3%	6%	41%	38%	12%
Justify program existence or continuation.	127	3.835	0.784	2%	2%	26%	54%	17%
Make decisions about staffing (e.g., in the program being evaluated or in the org as a whole).	125	3.200	1.122	7%	19%	34%	26%	14%
Report to the board or senior management (or equivalent).	127	3.598	1.071	2%	17%	26%	32%	24%
Perform outreach and public relations.	127	3.315	1.052	6%	15%	30%	39%	10%
Meet external accountability requirements.	127	3.898	0.898	2%	5%	22%	46%	26%
Inform and expand advocacy and influencing work.	127	3.614	0.976	2%	11%	31%	36%	20%
Total		**3.564**	**0.993**	**3%**	**11%**	**30%**	**39%**	**17%**

Note: For observed means, 1 = Never, 2 = Infrequently, 3 = Sometimes, 4 =Often, 5 = Always.

Table D8

Survey Results for the Process Use (PU) Items

Process Use Items	Descriptive Statistics			Response Percentages				
	n	*M*	*SD*	Never	In-frequently	Some-times	Often	Always
Develop knowledge about evaluation logic.	125	3.704	0.942	1%	10%	27%	41%	21%
Develop knowledge about evaluation methods.	126	3.802	0.858	0%	6%	29%	42%	22%
Develop technical skills for doing evaluation (e.g., instrument development, data collection and analysis).	126	3.786	0.900	0%	10%	25%	44%	22%
Develop a better understanding of the program/policy/intervention being evaluated.	126	3.984	0.820	0%	3%	25%	43%	29%
Integrate evaluation into our work practices.	125	3.760	0.893	0%	7%	33%	37%	23%
Improve management practices in the office.	123	3.480	1.003	2%	13%	41%	25%	20%
Expedite/intensify/expand the use of evaluation findings.	124	3.621	0.907	1%	10%	33%	40%	17%
Develop professional networks.	125	3.096	1.088	9%	18%	37%	26%	10%
Question underlying assumptions about what we do.	124	3.419	0.912	2%	14%	36%	38%	10%
Develop a mindset of evaluative thinking.	125	3.512	0.904	1%	12%	36%	38%	14%
Increase commitment to the program/policy/intervention being evaluated.	123	3.618	0.883	0%	8%	41%	33%	19%
Increase organizational commitment.	124	3.702	0.971	1%	11%	27%	38%	23%
Increase ownership of what we do.	123	3.740	0.965	1%	11%	26%	39%	24%
Appreciate the value of evaluation.	125	3.800	0.898	0%	7%	30%	38%	25%
Appreciate the power of evaluation as a force for change.	124	3.718	1.017	2%	10%	31%	31%	27%
Total	**1868**	**3.650**	**0.951**	**1%**	**10%**	**32%**	**37%**	**20%**

Note: For observed means, 1 = Never, 2 = Infrequently, 3 = Sometimes, 4 =Often, 5 = Always.

Table D9

Survey Results for the Mediating Conditions (MC) Items

Mediating Conditions Items	Descriptive Statistics			Response Percentages				
	n	*M*	*SD*	Never	In-frequently	Some-times	Often	Always
Foster improvement in program implementation quality and/or outcomes.	123	3.756	0.793	0%	3%	37%	41%	19%
High quality.	123	3.569	0.831	0%	9%	38%	40%	13%
Provide accurate results.	123	3.764	0.736	0%	2%	34%	48%	15%
Reduce surprises for decision makers.	121	3.529	0.775	1%	5%	45%	40%	10%
Perceived by users (e.g., program managers, technical specialists) as unbiased.	122	3.803	0.799	1%	3%	29%	49%	18%
Informed by user input.	120	3.883	0.780	0%	3%	29%	46%	23%
Supported by quality assurance mechanisms (e.g., peer reviews).	122	3.475	1.038	3%	15%	30%	36%	16%
Based on objective data (e.g., based on facts, robustly measured)	121	3.909	0.796	1%	2%	26%	48%	23%
Can be compared against external standards or benchmarks.	122	3.607	0.858	0%	10%	34%	41%	15%
Produced by evaluators who are perceived as credible.	123	3.943	0.739	0%	2%	25%	50%	23%
Perceived to be appropriately resourced.	120	3.583	0.846	2%	7%	35%	45%	12%
Use a methodology that is understood by users.	121	3.884	0.766	0%	3%	26%	50%	21%
Accessible to all staff members.	123	3.732	0.942	0%	11%	29%	37%	24%
Total		**3.726**	**0.838**	**1%**	**6%**	**32%**	**44%**	**18%**

Note: For observed means, 1 = Never, 2 = Infrequently, 3 = Sometimes, 4 =Often, 5 = Always.

Appendix E: Rotated Structure Matrix of Survey Results by Dimension

Table E1

Rotated Structure Matrix for Organizational Learning (OL) Items

Organizational Learning Items	Components						
	1	2	3	4	5	6	7
Staff can often bring new ideas to improve programs.				0.431			
Failures are constructively discussed.				0.627			
Current practice in teams encourage staff to solve problems together before discussing them with a manager.				0.398			
People who are new are encouraged to question the way things are done.				0.545		0.337	
Senior managers accept change and are not afraid of new ideas.				0.708			
Managers encourage staff to experiment in order to improve work processes.				0.674			
New work processes that may be useful are usually shared with all staff.				0.529		0.386	
Senior managers and staff share a common vision of what our work should accomplish.				0.514			
Managers frequently involve staff in important decisions.				0.599	0.342		
We can usually create informal groups to solve organizational problems.							
Managers can accept criticism without becoming overly defensive.				0.471			
We have a system that allows us to learn successful practices from other organizations.	0.367					0.377	
Managers often provide useful feedback that helps to identify potential problems and opportunities.		0.381		0.377		0.303	
We have opportunities for self-assessment with respect to goal attainment.				0.399			
Most problem-solving groups include staff from a variety of functional areas or divisions.	0.399			0.526			
Employees are given sufficient time to reflect on organizational successes or failures.	0.367			0.502			

Table E2

Rotated Structure Matrix for Organizational Support (OS) Items

Organizational Support Items	Components						
	1	2	3	4	5	6	7
Staff are provided with work-related skills training.						0.614	
Information and decision-making on programmatic issues must always go through proper/formal channels.					0.301		
The skills training that we receive can be applied to improve our work.						0.595	
Staff are encouraged to continuously upgrade and increase their technical knowledge and education levels.						0.578	
Supports the development of soft skills such as leadership, coaching, and team building among staff.						0.615	
Learning that improves the work skills and technical knowledge of staff is encouraged.						0.648	
Staff training is emphasized equally at all levels.						0.483	
Training is done in teams where appropriate.						0.491	
Our work is usually closely supported, monitored, and reviewed by management.				0.492		0.398	

Table E3

Rotated Structure Matrix for Capacity to Do Evaluation (CTD) Items

Capacity to Do Evaluation Items	Components						
	1	2	3	4	5	6	7
We have long-term, dedicated financial support to ensure evaluation activities across all programming where evaluation is required.	0.323						
We are provided with the basic tools/resources to support evaluation (e.g., computers, software, copying, administrative support).				0.313		0.342	0.315
We have a "champion" on staff who supports our evaluation efforts and advocates on our behalf when required.							0.528
Our office possesses the technical competencies to conduct evaluations (e.g., instrument development, data collection and analysis).							0.644
Our office has the knowledge and skills to oversee evaluations performed by external evaluators.						0.304	0.573
Our office provides positive encouragement to conduct evaluation.							0.455
We have formal requirements to report on programmatic performance.				0.324			0.432

Table E4

Rotated Structure Matrix for Evaluative Inquiry (EI) Items

Evaluative Inquiry Items	Components						
	1	2	3	4	5	6	7
Review program documentation (e.g., participant records, case notes).			0.364				
Conduct firsthand observation of program activities.	0.405		0.492				
Conduct formal program evaluations.			0.583				
Establish performance targets (e.g., serve 200 people, 80% complete training).		0.322	0.411				
Monitor implementation (i.e., ensure that programs are delivered as intended).			0.734				
Monitor program outcomes (i.e., ensure program results are as intended).			0.652				0.3
Assess the degree to which program goals/objectives are met.			0.645				
Engage in formal evaluation planning processes.			0.642				
Employ single-case mixed-method designs (i.e., qualitative and quantitative methods).			0.754				
Use program theoretical designs (i.e., link program components to outcomes).		0.395	0.562				
Produce annual and/or reports based on performance measures.			0.439				
Produce reports about program activities.			0.554				
Produce reports for Boards of Directors and/or senior management.							0.471
Use a performance measurement system.							0.546
Use program logic models.		0.376	0.535				
Use other evaluation or management systems (e.g., performance audits, results-based management, quality assurance activities).	0.382						0.402

Table E5

Rotated Structure Matrix for Stakeholder Participation Frequency (SPF) Items

Stakeholder Participation Frequency Items	Components						
	1	2	3	4	5	6	7
Program/project developers (team that designed the program).			0.399				
Program/project managers or directors.				0.329			
Program/project sponsors, donors or funders.					0.466		
Staff responsible for implementing the program/project.					0.339		
Participants of the impact group (well-defined group of individuals seeking to facilitate lasting change or impact) of the program/project.					0.663		
Special interest groups or target groups (individuals whose actions or behaviors generate outcomes) of the program/project.					0.654		0.346
Partner organizations.					0.734		
Stakeholders representing different identities (e.g., race, disability, age, migration status, etc.)					0.675		

Table E6

Rotated Structure Matrix for Stakeholder Participation Level (SPL) Items

Stakeholder Participation Level Items	Components						
	1	2	3	4	5	6	7
Program/project developers (team that designed the program).		0.422					
Program/project managers or directors.		0.338					
Program/project sponsors, donors or funders.		0.317					
Staff responsible for implementing the program/project.					0.402		
Participants of the impact group (well-defined group of individuals seeking to facilitate lasting change or impact) of the program/project.					0.749		
Special interest groups or target groups (individuals whose actions or behaviors generate outcomes) of the program/project.					0.654		
Partner organizations.					0.645		
Stakeholders representing different identities (e.g., race, disability, age, migration status, etc.)					0.612		

Table E7

Rotated Structure Matrix for Evaluation Findings Use (EFU) Items

Evaluation Findings Use Items	Components						
	1	2	3	4	5	6	7
Make changes to existing programs.	0.594						
Conduct strategic planning at the organizational level.	0.724						
Get new funding.	0.542						
Justify program existence or continuation.	0.554						-0.314
Make decisions about staffing (e.g., in the program being evaluated or in the org as a whole).	0.691						
Report to the board or senior management (or equivalent).	0.681						
Perform outreach and public relations.	0.718						
Meet external accountability requirements.	0.395						0.337
Inform and expand advocacy and influencing work.	0.747						

Table E8

Rotated Structure Matrix for Process Use (PU) Items

Process Use Items	Components						
	1	2	3	4	5	6	7
Develop knowledge about evaluation logic.	0.644					0.3	
Develop knowledge about evaluation methods.	0.595					0.317	
Develop technical skills for doing evaluation (e.g., instrument development, data collection and analysis).	0.574					0.455	
Develop a better understanding of the program/policy/intervention being evaluated.	0.603						
Integrate evaluation into our work practices.	0.716						
Improve management practices in the office.	0.779						
Expedite/intensify/expand the use of evaluation findings.	0.761						
Develop professional networks.	0.649						
Question underlying assumptions about what we do.	0.721						
Develop a mindset of evaluative thinking.	0.726						
Increase commitment to the program/policy/intervention being evaluated.	0.755						
Increase organizational commitment.	0.784						
Increase ownership of what we do.	0.718						
Appreciate the value of evaluation.	0.712						
Appreciate the power of evaluation as a force for change.	0.781						

Table E9

Rotated Structure Matrix for Mediating Conditions (MC) Items

Mediating Conditions Items	Components						
	1	2	3	4	5	6	7
Foster improvement in program implementation quality and/or outcomes.	0.516	0.370					
High quality.	0.361	0.535					
Provide accurate results.		0.638					
Reduce surprises for decision makers.	0.48						
Perceived by users (e.g., program managers, technical specialists) as unbiased.		0.587					
Informed by user input.		0.414					0.326
Supported by quality assurance mechanisms (e.g., peer reviews).	0.402	0.447					
Based on objective data (e.g., based on facts, robustly measured)		0.665					
Can be compared against external standards or benchmarks.		0.629					
Produced by evaluators who are perceived as credible.		0.749					
Perceived to be appropriately resourced.		0.653				0.304	
Use a methodology that is understood by users.		0.701					
Accessible to all staff members.		0.504					0.415

Appendix F: Semi-Structured Interview Protocol

Ryan read the University of Denver's IRB approved verbal consent script and received participant's consent before hitting record and commencing interview.

1. What is your role, including in the organization's MEAL work and COP?
2. How would you describe organizational evaluation capacity? What are the elements?
3. What do you believe are some of the strengths of your organization's capacity?
4. Did you learn anything from the recent survey on evaluation capacity? For example, did you have takeaways from specific questions or the overall structure that made you think about evaluation capacity differently?

Ryan briefly reviewed results of the recent survey on evaluation capacity and tailored questions based on interviewees role and interests.

5. What surprises you about the survey results?
6. What experiences have you had that would make you think the results would be different?
7. Do you have concerns with the validity of the results?
8. How can you use the data to inform future investments in eval capacity building?
9. Do the weaknesses identified from the survey represent useful areas for investment?
10. What are other variables that should affect the choices of investment areas for evaluation capacity building?