An Integrated Mixed Methods Study to Construct a Usable Model for Impact Evaluation Capacity Development

Corey D. Smith
Western Michigan University, donovansmith@gmail.com

Follow this and additional works at: https://scholarworks.wmich.edu/dissertations

Recommended Citation

This Dissertation-Open Access is brought to you for free and open access by the Graduate College at ScholarWorks at WMU. It has been accepted for inclusion in Dissertations by an authorized administrator of ScholarWorks at WMU. For more information, please contact wmu-scholarworks@wmich.edu.
AN INTEGRATED MIXED METHODS STUDY TO CONSTRUCT A USABLE MODEL FOR IMPACT EVALUATION CAPACITY DEVELOPMENT

by

Corey D. Smith

A dissertation submitted to the Graduate College in partial fulfillment of the requirements for the degree of Doctor of Philosophy Interdisciplinary Ph.D. in Evaluation Western Michigan University June 2018

Doctoral Committee:

Daniela Schröter, Ph.D., Chair
Chris Coryn, Ph.D.
Mary Kane, M.L.S
To mitigate problems that development practitioners, evaluation managers, funders, and evaluators might face when considering how to plan and implement impact evaluations, there is a need for increased clarity around these issues. The primary purpose of this research was to build upon the results of a group concept mapping (GCM) study conducted at the International Labour Organization to examine two related, but separate issues: (1) what are the perceptions of evaluators and international development practitioners with regards to elements of organizational impact evaluation capacity? and (2) how do the responses of a larger group of evaluators and development practitioners either support or not support the conceptual structure of the original co-developed concept map? Descriptive and inferential investigations provided insight into the perspectives of international development practitioners with regard to impact evaluation capacity for organizations, and an examination of the validity of the framework presented by the original concept map, to develop a comprehensive impact evaluation capacity development framework.
This research employed a cross-sectional design using a survey developed around the ideas from the original concept mapping study. Using a sampling of items from the original study, the survey data was analyzed via exploratory factor-analysis (EFA), using the concept map structure as the theoretical model to identify the constructs underlying the survey instrument. Results were linked back to the original concept map to build a revised conceptual framework for impact evaluation capacity development. The framework was vetted via expert review to enhance its utility and relevance.

A six-factor model was identified from the EFA and used to develop a framework for understanding and developing impact evaluation within institutions who may be seeking to incorporate these types of evaluations into their organizational infrastructure. The model is presented as a sequence of steps that together form a comprehensive structure for effectively planning, designing, and implementing impact evaluations. It also serves as a sensemaking tool for practitioners to use when trying to link the existing literature and guidance around impact evaluation to their own work.
ACKNOWLEDGEMENTS

Completing this dissertation has been one of the most difficult things I have ever done. It has taken a lot of time, work, and conversation to get to this point. It represents the culmination of nearly 7 years of graduate work that was only possible because of the support of a number of individuals. There are many family, friends, colleagues, and peers who have helped me along the way, and to all those I don’t have space to name I say thanks.

First and foremost, I want to thank my committee chair, Dr. Daniela Schröter, who has been a friend and a mentor over the past five years, and who opened my eyes to all types of experiences during my graduate work. It was her encouragement that led me to Geneva, Switzerland which ultimately led me to this dissertation topic. I also want to thank Dr. Chris Coryn for his sound methodological advice, his encouragement, and his willingness to lend me any book that I might need, and that he might have. Mary Kane, I am so grateful to have met you through our work together in Geneva, and for your willingness to read, and re-read my work. Your clarity of thought was invaluable as I was trying to formulate my arguments and ideas for this dissertation.

Beyond academics I am grateful to a number of people for the professional opportunities they have given me to develop myself as an evaluator. Thanks to them I
was able to apply what I was learning in school to the real world on a daily basis. That has been invaluable to me as both a practitioner and a scholar. First, thank you to Dr. Meg Blinkiewicz who introduced me to the field of evaluation, and who took a chance on me when I was struggling to find my way in the world of work. It was your encouragement that led me to apply to the Interdisciplinary Ph.D. in Evaluation program, and I thank you for your continued friendship and support over the years. I also am thankful to Dr. Lori Wingate who offered me an opportunity to work with her at the Evaluation Center during my graduate work, and who always gave me sound and practical advice. Thank you too to Dr. Wendy L. Tackett who has also given me the chance to build my real-world skills, and who has been with me almost every step of the way on this journey. Thank you for your encouragement, guidance, and for putting your faith in me over all these years.

I also need to thank my colleagues from the International Labour Organization, Guy Thijs, Dr. Craig Russon, Peter Wichmand, Naomi Asukai, and Leilah Abdallah. Their support and embrace of me during my time in Geneva provided me with the foundation for my dissertation. Thank you for supporting the group concept mapping project that we undertook while I was there, and thank you for your advice and guidance along the way. I want to specifically thank Dr. Craig Russon for the words of wisdom that made me think intentionally about my experience in Geneva while trying to conceptualize a dissertation idea.
My time in the Interdisciplinary Ph.D. program would also not have been what it was without the friendship and support of my peers. Thank you to all my peers for long talks about evaluation, hilarious jokes, and all things related to our quest to become better scholars and practitioners. They have been a source of strength, and encouragement. Thanks to Kristin Hobson, Lyssa Wilson, Pedro Mateu, Satoshi Ozeki, Carl Westine, and Kelly Robertson who provided practical and philosophical advice as I worked through both my course work, and my dissertation.

I also owe great gratitude to my parents, Geoff Smith and Rhona Polonsky who have supported me in all ways throughout all my life as I tried to find a path forward. You have been as loving as any parents could be and I’m eternally grateful to you for what you have given me and my brother. Completing this dissertation is a testament to your development of me as a learner. Thank you.

Finally I want to thank partner, Natalie Schneider. You have stood by me all these years; you’ve been my cheerleader, my proofreader, and my sounding board. I know it hasn’t always been easy and I know that this endeavor has taken time away from us, but I couldn’t have done it without you. I love you and can’t wait to step side-by-side with you into our future.

Corey D. Smith
# Table of Contents

ACKNOWLEDGMENTS .................................................................................................................. ii

LIST OF TABLES ......................................................................................................................... viii

LIST OF FIGURES ......................................................................................................................... ix

CHAPTER

1. INTRODUCTION ......................................................................................................................... 1

   Problem Statement ..................................................................................................................... 5

   Study Purpose .......................................................................................................................... 10

   Significance/Contribution of this Dissertation ........................................................................ 12

   Organization of this Dissertation ........................................................................................... 18

2. LITERATURE REVIEW ............................................................................................................. 20

   Introduction ............................................................................................................................. 20

   The Paradigm Wars ................................................................................................................. 21

   The Attribution Issue ............................................................................................................. 26

   Organizational Policy Making on Impact Evaluation and Definitions .................................... 32

   The Conceptual Underpinnings ............................................................................................... 38

   Group Concept Mapping and Measurement .......................................................................... 47

   Chapter Summary ................................................................................................................... 49

3. METHODOLOGY ....................................................................................................................... 51
<table>
<thead>
<tr>
<th>Chapter</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><strong>Table of Contents—Continued</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>CHAPTER</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Study Design</td>
<td>52</td>
</tr>
<tr>
<td></td>
<td>Sample and Sampling Procedure</td>
<td>52</td>
</tr>
<tr>
<td></td>
<td>Participant Selection for Expert Review Panel</td>
<td>55</td>
</tr>
<tr>
<td></td>
<td>Communication with Survey Sample</td>
<td>56</td>
</tr>
<tr>
<td></td>
<td>Instrument Development</td>
<td>56</td>
</tr>
<tr>
<td></td>
<td>Item Reduction</td>
<td>57</td>
</tr>
<tr>
<td></td>
<td>Procedure</td>
<td>60</td>
</tr>
<tr>
<td></td>
<td>Survey Analysis</td>
<td>61</td>
</tr>
<tr>
<td></td>
<td>Survey Analysis for Examining the Perceptions and Opinions of Respondents</td>
<td>61</td>
</tr>
<tr>
<td></td>
<td>Factor Analysis of the Survey Data</td>
<td>63</td>
</tr>
<tr>
<td></td>
<td>Framework Development, Expert Review, and Face Validity</td>
<td>65</td>
</tr>
<tr>
<td></td>
<td><strong>4. RESULTS</strong></td>
<td>67</td>
</tr>
<tr>
<td></td>
<td>Respondents</td>
<td>68</td>
</tr>
<tr>
<td></td>
<td>Survey Results</td>
<td>72</td>
</tr>
<tr>
<td></td>
<td>Exploratory Factor Analysis</td>
<td>85</td>
</tr>
<tr>
<td></td>
<td>Test of Sampling Adequacy</td>
<td>85</td>
</tr>
<tr>
<td></td>
<td>Factor Extraction and Rotation</td>
<td>86</td>
</tr>
<tr>
<td></td>
<td>The Factor Structure</td>
<td>90</td>
</tr>
<tr>
<td></td>
<td>Comparative Analysis of the Theoretical And Observed Models</td>
<td>97</td>
</tr>
</tbody>
</table>
Table of Contents—Continued

CHAPTER

A Revised Framework for Impact Evaluation Capacity Development.......................... 102

5. SUMMARY, CONCLUSIONS, AND RECOMMENDATIONS ........................................ 110

Interpretation of the Results.......................................................................................... 111

Re-examining the Contribution of this Research ......................................................... 117

Implications .................................................................................................................... 118

Limitations ...................................................................................................................... 121

Future Directions .......................................................................................................... 123

REFERENCES .................................................................................................................. 126

APPENDICES

A. Factors, Item Codes, and Item Text ........................................................................ 140

B. Emails to Listservs for Survey Invitations and Follow-Up ....................................... 143

C. Survey ...................................................................................................................... 147

D. Invitation to Expert Panel ......................................................................................... 153

E. Expert Panel Facilitation Protocol ........................................................................... 155

F. HSIRB Approval ......................................................................................................... 157
LIST OF TABLES

1. Stated purposes for and definitions of impact evaluation among a sample of organizations................................................................. 36

2. Proportion of ideas in original map and reduced item set by conceptual cluster....................................................................................... 59

3. Respondent reported professional roles........................................................................................................................................ 69

4. Respondent reported country of primary office and representation by continent.................................................................................. 70

5. Mean importance rating at the cluster level..................................................................................................................................... 73

6. Importance ratings of items.............................................................................................................................................................. 76

7. Measures of factor score adequacy for the six-factor solution........................................................................................................ 89

8. Items and factor loadings for six-factor structure............................................................................................................................ 92

9. Titles and descriptions of the six factors....................................................................................................................................... 95
<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Timeline of key events related to the rise of impact evaluation in international development</td>
<td>25</td>
</tr>
<tr>
<td>2</td>
<td>Item reduction process</td>
<td>58</td>
</tr>
<tr>
<td>3</td>
<td>Illustration of the top five rated items as process markers for designing and implementing impact evaluations</td>
<td>74</td>
</tr>
<tr>
<td>4</td>
<td>Scree plot depicting factor extraction and principal component extraction</td>
<td>87</td>
</tr>
<tr>
<td>5</td>
<td>Scree plot from the parallel analysis</td>
<td>89</td>
</tr>
<tr>
<td>6</td>
<td>6-factor extracted model</td>
<td>91</td>
</tr>
<tr>
<td>7</td>
<td>Original concept map with the retained items from factor analysis representing the six factors</td>
<td>99</td>
</tr>
<tr>
<td>8</td>
<td>Revised institutional impact evaluation capacity framework</td>
<td>105</td>
</tr>
</tbody>
</table>
CHAPTER 1
INTRODUCTION

Impact evaluation has become an important topic for evaluators, funders, and practitioners in the field of international development (Hulme, 2000; Lensink, 2014; White, 2010). Cameron, Mishra, and Brown’s (2015) descriptive cataloging of impact evaluations completed between 1981 and 2012 demonstrates dramatic increases in the number of impact evaluations that have been conducted since 2000. Yet the definition of impact evaluation and how it is operationalized has been a topic of heated debate (Picciotto, 2014; White, 2010). As the debate continues, development practitioners, evaluators, and evaluation stakeholders cope with an array of guidance material, opinions, and technical papers dedicated to this type of evaluation.

Much of the recent growth in impact evaluation has been driven by demand for more precise knowledge of how investments in international development initiatives are changing people’s lives. Impact evaluation employing experimental or quasi-experimental designs became more widely practiced in development during the early 1990’s, when conditional cash transfer programs were closely studied to assess their value as tools for alleviating poverty (Davis, Gaarder, Handa, & Yablonski, 2012). An example of the prevalence of these types of studies comes from Mexico, where prior to 2013 half of the impact evaluations that were conducted in that country were of conditional cash transfer programs (Cameron, Mishra, & Brown, 2015).
At the World Bank’s 2003 conference on Evaluation and Development Effectiveness, two economists from the Massachusetts Institute of Technology (MIT) argued for the increased utilization of randomized control trials (RCTs) in impact evaluations (Duflo & Kremer, 2003). In 2004, a group of donors, including major U.S.-based foundations, provided financial support to create the Evaluation Gap Working Group, assembled by the Center for Global Development (CGD). In 2006, the Working Group published “When Will We Ever Learn: Improving Lives Through Impact Evaluation”, a report describing the lack of evidence for determining whether programs or policies in development do or do not work. The working group argued that there was a significant lack of empirical evidence in the field of development, stating that “after decades in which development agencies have disbursed billions of dollars for social programs, and developing country governments and nongovernmental organizations (NGOs) have spent hundreds of billions more, it is deeply disappointing to recognize that we know relatively little about the net impact of most of these social programs” (Savedoff, Levine, & Birdsall, 2006, p. 1). The group also described the challenges facing various actors in generating such evidence, argued for the importance and feasibility of doing these types of studies, and proposed steps to improve the evidence base for development assistance programming.

Publication of the CGD report escalated the debate on what constitutes impact evaluation in part because the report prescribed an approach for how impact evaluation should be done. There were three key tenets of impact evaluation that, per the CGD
report, differentiated impact evaluation from other forms of evaluation (Savedoff et al., 2006, p. 13):

1. The purpose of impact evaluations is to generate generalizable knowledge about what works across settings and time.

2. Impact evaluations require a counterfactual so that valid inferences can be made about program effects.

3. Impact evaluations should be done on new and untested programs.

The authors of the CGD report argue that impact evaluations should be done using RCTs, although they recognize that this design option is not always possible and, therefore, that other types of quasi-experiments might be used if this is the case (i.e., designs that employ propensity score matching, regression discontinuity, interrupted time-series, and difference-in-difference). The purpose of impact evaluations, according to the CGD working group, is to generate knowledge that can be used widely, and that ultimately becomes a public good (Savedoff et al., 2006). To this end, they write: “The value of impact evaluation is best understood as part of a broad scientific enterprise of learning, in which evidence is built over time and across different contexts, forming the basis for better policymaking and program design” (Savedoff et al., 2006, p. 13). This contribution of knowledge is one key distinction that is often made when contrasting impact evaluation with other forms of evaluation (e.g., performance evaluation, goal-based evaluations), that are intended to assess the implementation of programs and to what extent they achieved the outcomes outlined within their program design. Embedded within this purposive knowledge generation is an assumption that impact evaluation is to
be used for what Weiss (1998) called ‘enlightenment use.’ Impact evaluations, then, should generate knowledge for public use about what works, for whom, and under what conditions so that policy makers and program designers might make informed decisions.

Three years after the CGD published its report, the Network of Networks on Impact Evaluation (NONIE) released a guidance document outlining key principles, purposes, and approaches for conducting impact evaluations in development (Leeuw & Vaessen, 2009b). The document highlighted the relative value of experimental and quasi-experimental designs for assessing questions of causality, but it also emphasized the explanatory power of other mixed-methods approaches. The idea of a guidance document that was supported by a “network of networks”¹ was meant to help settle the growing debate around the most appropriate approach for doing impact evaluations. While the publication of the guidance document represented a major step towards achieving that goal, the debate did not go away completely and there is still preferential treatment for experimental and quasi-experimental designs.

Along with the growing interest in impact evaluation the debate between proponents of evaluation designs premised upon the use of counterfactuals (3ie, 2012b; J-PAL, 2017; Savedoff et al., 2006; White, 2010) and those who believe that rigorous impact evaluation can be done through more theory-driven approaches (Deaton, 2010; Dietz et al., 2013; Picciotto, 2012; Ton, 2012) has continued. These debates are important

¹ NONIE was comprised of the Organisation for Economic Co-operation and Development’s Development Assistance Committee (OECD/ DAC) Evaluation Network, the United Nations Evaluation Group (UNEG), the Evaluation Cooperation Group (ECG), and the International Organization for Cooperation in Evaluation (IOCE)
for generating new thinking, research, and approaches to evaluation, but for the non-evaluation community who is seeking to incorporate impact evaluation into their work, the debate can be confusing rather than thought provoking. To this end there is a need to offer a way for non-evaluation audiences to parse through the array of material being published on the topic of impact evaluation.

**Problem Statement**

Major international development organizations, including The World Bank, the International Initiative for Impact Evaluation (3ie), the United States Agency for International Development (USAID), and the United Nations International Children’s Emergency Fund (UNICEF), reference “impact evaluation” in a wide assortment of documents available to the public. These organizations and many others have allocated substantial resources towards producing and disseminating material relevant to or describing impact evaluation. They include guidance documents, evaluation policies, full length books, and/or technical notes. The ideas, prescriptions, and design guidelines within them reflect ideas that have been fiercely debated among evaluation scholars (Cook, Scriven, Coryn, & Evergreen, 2010), and some are more comprehensive than others.

The debate surrounding impact evaluation has not been particularly constructive (UNEG Impact Evaluation Task Force, 2013), but it has resulted in what might be described as a truce. This truce can be seen in recent claims from certain materials, such as “high-quality impact evaluations measure the net change in outcomes amongst a particular group, or groups, of people that can be attributed to a specific program using the best
methodology available, feasible and appropriate to the evaluation question(s) being investigated and to the specific context” (3ie, 2012b, p. 1) or, “a ‘mixed method’ approach utilizing quantitative, qualitative, participatory and blended (e.g., quantifying qualitative data) approaches is now widely accepted as advisable to address the types of interventions that are now predominant in international development” (UNEG Impact Evaluation Task Force, 2013, p. 4). Both excerpts exemplify language that is open to interpretation, emphasizing context-driven design and methodological plurality.

The arguments reveal epistemological differences in how to define rigor (Picciotto, 2014), how to define impact (White, 2010), and what evaluation designs are appropriate for particular evaluands within certain contexts (Leeuw & Vaessen, 2009a). Further, the arguments revolve around two key strands of argumentation; namely, that impact evaluations should utilize experimental designs where possible in contrast to designs that utilize other approaches to developing causal claims (e.g., theory-driven approaches) (UNEG Impact Evaluation Task Force, 2013). The divide appears to stem from different disciplinary backgrounds of evaluation professionals. On the one hand, economists have generally led the charge towards RCTs and other statistically-oriented designs, bringing with them the field of econometrics, while evaluators with other social science backgrounds have argued that impact evaluations can be done in a rigorous manner using theory-driven approaches (Picciotto, 2012).

While much of the debate has centered around how to do impact evaluations, the definition of impact has also been the subject of disagreement. In their paper, Belcher and Palenberg (2016) systematically reviewed definitions of ‘impact’ in the materials from
an array of international development organizations, research centers, and bilateral agencies. Their results describe the immense variation in how these organizations define and articulate the term. Their findings highlight an important issue; that is, the lack of clarity just in how to define the term ‘impact’ in development. Clarity on these issues is critical for clients of evaluation, evaluators themselves, and evaluations funders, since without them it can be difficult to move forward effectively with commissioning, conducting, or using these types of studies, let alone establishing any kind of standards for how they should be done.

But why does this debate pose a problem? First, evaluation clients (or potential clients) may begin to find increased internal or external demands for impact evaluation. This pressure might come from funders, boards or governing bodies, internal leadership, or stakeholder and constituent groups, who seek a better understanding of what is working, and to what end. Some of those making these demands might also be responding to the increased discussion on impact evaluation, and may also attempt to prescribe the approaches that they feel should be used. Those responsible for responding to these demands to conduct impact evaluation may be confused and potentially frustrated by the professions’ lack of cohesion and clarity. As they explore relevant methods and contexts, agencies may discover conflicting opinions around study design, quantitative and qualitative methods, or even how to define impact. Evaluation clients look to the evaluation community for guidance on selecting appropriate evaluation models and designs. Understanding the existing material around impact evaluation will help them establish a foundation for impact evaluation in their context, assess their
organizational capacity, seek appropriate external expertise where needed, allocate appropriate funding, and develop any internal systems they may need to conduct impact evaluations in ways that make sense for their organization.

Lack of clarity also poses a problem for evaluation practitioners. There is no doubt that debate within the field and among evaluators is beneficial and helps evaluation as a discipline to evolve, innovate, and move forward with new and different ideas, but, as with the positivist versus constructivist debates that took place in North America in the 1980s and 1990s, the impact evaluation debate becomes less useful when positions become intransigent. As illustrated earlier, there is some form of agreement now within the field and some acceptance that different approaches to impact evaluation are acceptable beyond those employing a counterfactual. But evaluation practitioners who look towards the professional literature to guide and support their practice also require clarity around impact evaluation and the implications of taking one approach over another. It is important that the evaluation community put forth a unified front to avoid jeopardizing the integrity and credibility of our profession.

Lastly, as philanthropic foundations, national development agencies, and multilateral organizations have come under more pressure from both internal and external sources, to show results of their investments, they have begun to turn to impact evaluation for answers to these questions. As funders, these types of organizations have a responsibility to articulate the parameters for evaluations. A simple declaration that a counterfactual-based approach is the way impact evaluations must be done may not suffice, particularly if a funder supports a diverse portfolio of programs or initiatives
operating at various institutional levels. There is relative agreement that evaluation questions should drive impact evaluation designs, approaches, and methods (Leeuw & Vaessen, 2009; 3ie, 2012b). It is important that funders understand this, or are at least able to critically assess the programs or projects they fund to develop impact evaluation policy and requirements that are appropriate and will lead to the generation of meaningful evidence.

Taking the point of view of these three different groups (i.e., clients of evaluation, evaluation practitioners, and funders of evaluation), it becomes clearer why this ongoing debate may pose problems. It is the imperative of the evaluation field to offer some way to make sense of that material that we, as a field, continue to write and publish. While a diversity of opinions and points of view on this topic enrich the conversation, it must not become overly difficult for users to identify key pieces of information that they can use to make decisions for themselves. A systematic process for interpreting the breadth of material, and the opinions that come with that material, offers individuals and organizations an opportunity to develop a framework for planning, implementing, and using impact evaluation in a way that is responsive to their needs and context.

Understanding the ways evaluation practitioners work in the realm of international development is important for knowing how the field views impact evaluation as it continues to gain traction among funders, evaluation commissioners, and evaluators themselves. A descriptive study allows for insight informed by many evaluation practitioners and can help guide evaluation training, and help the field better understand where priorities of practitioners may lie.
Study Purpose

There is a need for increased clarity around these issues to mitigate problems that development practitioners, evaluation managers, funders, and evaluators might face when considering how to plan and implement impact evaluations. The primary purpose of this research is to build upon the results of a group-concept mapping (GCM) study conducted at the International Labour Organization (Smith & Kane, manuscript in preparation) to examine two related but separate issues: (1) what are the perceptions of evaluators and international development practitioners of the ideas and concepts represented in the GCM data? and (2) how do the responses of a larger group of evaluators and development practitioners either support or not support the structure of the original concept map? This review will lead to both a descriptive and empirical investigation into the perspectives of impact evaluation capacity for organizations, and an examination of the validity of the framework presented by the original concept map, with the intent of developing a more refined impact evaluation capacity building framework.

In a GCM study, a series of idea clusters emerge from the analysis of the sorting of individual items or ideas with one another by participants. Those clusters of similar content are spatially bound through hierarchical cluster analysis (Kane & Trochim, 2007). In the study done at the ILO, eight clusters emerged that served as the eight key issues that required investigation as part of creating a comprehensive impact evaluation strategy. Each cluster consisted of a varying number of ideas that had been brainstormed...
by a network of evaluation professionals in the organization who had participated in the GCM study.

To serve the two purposes described previously, the data from the original GCM study was used as the basis for the design of a survey instrument. The survey explored the perceptions and attitudes of a broader group of evaluators and practitioners as they pertain to the key ideas and concepts represented in the GCM study. This descriptive data offers insight into how practitioners view impact evaluation, and what they believe is most important when considering an organizational strategy for using it effectively to examine the work of international development organizations and entities. The results can be used to develop a tool or process that may be adopted by organizations seeking to develop a cohesive and comprehensive impact evaluation strategy. The framework aims to be broadly acceptable, balanced by specificity that ensures its adaptability and utility.

The key questions for this research relate to the two purposes described previously. The first two questions draw upon the same data set (i.e., the survey), but the data was analyzed in different ways. The third question is less quantitative and relates to the tool that emerges based on the initial analyses.

Question 1: What components of institutional impact evaluation capacity development do international development practitioners and evaluators believe are most important?

- Sub question 1.1: What, if any, characteristics of respondents explain differences or variations in response patterns (e.g., role, training background, experience)?
• Sub question 1.2: How, if at all, do the attitudes and perceptions differ from or align with the results from the original concept mapping study?

Question 2: In what ways do the constructs measured by the survey instrument relate to concepts that were present in the original concept map?

• Sub question 2.1: How do the results validate or invalidate the framework from the original concept map?
• Sub question 2.2: What are the implications for utilizing this approach for validating other concept maps?

Question 3: What kind of framework do the results of this research lend themselves to?

• Sub question 3.1: In what ways does the emerging framework reflect other evaluation capacity building frameworks?
• Sub question 3.2: Is the framework useful for organizations engaged in international development work?

**Significance/Contribution of this Dissertation**

Given the increasing calls for research on evaluation (ROE) (Coryn, Wilson, Westine, Ozeki, Fiekowsky, Greenman II, & Schröter, 2017) it is important to place this study in context to describe its contributions to the field. In outlining a framework for classifying and cataloguing RoE, Mark (2008) describes four overarching subjects to guide and stimulate future RoE, and why such a framework is of use. The framework suggests all the major reasons that evaluation, as a field, needs an improved empirical foundation to guide its practice, and inform its theory (Smith, 1993). The research introduced in this
chapter does not fall cleanly into one of these four subjects of inquiry, but as Mark (2008) writes, “these are not mutually exclusive categories. To the contrary, important studies of evaluation often will examine relationships across these categories. Most research on evaluation will in fact cut across these incorporating elements of some or all” (p. 117). Yet, to maintain some sense of order, I argue that this research falls most cleanly into the first subject of inquiry, evaluation context.

The subject of evaluation context is described by Mark (2008) as, “the circumstances within which evaluation occurs” (p. 117), and that inquiries into context might occur at the societal, organizational, or even evaluation-specific level. This research collects observations from evaluators who represent many types of organizations, and many different professional roles. However, its aim is to explore evaluation practice as it occurs, or might occur, at the organizational level. This investigation is an attempt to develop an organizational impact evaluation capacity building framework, and it means examining issues that cut across Mark’s subjects of inquiry.

Mark (2008) notes, “the categories that emerged resemble, to some degree, other frameworks familiar to evaluators” (p. 117), highlighting the frameworks similarities to logic modelling and Stufflebeam’s CIPP (context, input, process, product) model for evaluation (Stufflebeam & Coryn, 2014). The framework also bears some resemblance to the framework presented in the impact evaluation concept map that serves as the basis for this research. In both frameworks, issues of context serve as the starting point. An examination of Mark’s (2008) framework shows some overlap with the concept map with...
regard to key issues it addresses and describes (e.g., use, training, approaches, practices, professional issues). The cross-referential nature of the concept map to the Mark framework supports the argument for the logical relevance of the framework presented in the concept map. Furthermore, if these issues need to be examined in the field of evaluation more broadly, it is no doubt important to explore them within a particular approach to evaluation. By drawing upon Mark’s justification for developing a framework to classify the subjects of inquiry in RoE, one begins to understand the potential contribution of this effort to develop an empirically based framework for institutional impact evaluation capacity (Mark, 2008).

There has also been a call for comparative research on evaluation (Henry & Mark, 2003). The first research question presented in this dissertation asks how evaluators perceive key issues related to impact evaluation capacity at the organizational level, and how their perceptions vary based upon key factors (e.g., position, experience, role). This is the type of comparative research that Henry and Mark called for to continue building a more empirical basis for discussing how to move the field forward in ways that will positively affect practice.

Having placed this dissertation into the broader context of the RoE agenda (see also Szanyi, Azzam, & Galen, 2012), this dissertation should make three key contributions, the first two of which are tied directly to the evaluation field. The survey results from this research offer a descriptive overview of how international development practitioners, whether evaluation consultants, evaluation managers, or internal evaluators, perceive
key issues related to institutional capacity for conducting impact evaluation. Such research is still important for the evaluation field given its evolving nature (Henry & Mark, 2003; Mark, 2008). Examining the perceptions and opinions of a diverse group of evaluation professionals on the topic of institutional support and capacity development for conducting impact evaluation offers insight into how practitioners view the issues facing the use of impact evaluation methods and approaches in development work moving forward.

Like the détente that has settled over the quantitative-qualitative debates described by Rossi (1994) as a ‘crisis in evaluation,’ a truce exists between the experimentalist econometricians and the evaluators who Picciotto (2012) describes vaguely as, “seasoned development evaluators schooled in qualitative methods” (p. 215). But while that truce has quietly affected the official policy language of development organizations, there is still strong preference for the experimentalist versions of impact evaluation. This research will contribute, not by settling the debate, but by offering a way to understand it.

The second contribution that this research makes to the evaluation field is the development of a framework for institutional impact evaluation capacity. The framework offers a broad-based approach to take stock systematically of an organizational context with the aim of creating a comprehensive approach for conducting impact evaluation that is responsive to that organization’s needs. Previous evaluation capacity building models speak to evaluation more broadly (Bourgeois & Cousins, 2013; Cheng & King, 2016; King & Volkov, 2005; Preskill & Boyle, 2008; Tarsilla, 2012). These have merit and utility for
practitioners looking for some way to guide their formulation of a strategy to respond to the new demands, whether internal or external, for impact evaluation. Outlining a general model for evaluation capacity building, Preskill and Boyle (2008) write that these types of frameworks can be used for two primary purposes, to…” (a) guide practitioners’ ECB efforts and/or (b) empirically test the effectiveness of ECB processes, activities, and outcomes” (p. 444). While implied in the latter reason, it should be made clear that outlining an institutional impact evaluation capacity building framework can guide future research into impact evaluation by identifying key areas for focus, similar to what Mark (2008) and Henry and Mark (2003) did for RoE, more generally.

By producing a framework for conceptualizing impact evaluation, this study will help evaluation funders, clients, and practitioners make better decisions when planning, designing, or implementing impact evaluations. A framework is useful for defining parameters which might need to be considered by users without prescribing what should specifically be done within those parameters. Therefore, this dissertation offers a way to explore impact evaluation conceptually in ways that are accessible to a wide range of audiences.

The methodological sequence being explored in this study that pairs GCM with a factor-analytic approach to examine the validity of an original concept map has not been widely used. While there are numerous examples of researchers using GCM to inform instrument development (Rosas & Ridings, 2016), few studies then use the results of the instrument development to comment on the original map. This research also offers an opportunity to contribute to the methodological literature by exploring this link more
closely. The third contribution of this research is methodological in nature, and although it was originally characterized as unrelated to the evaluation field, it likely has some relevance to future RoE efforts. Although Rosas and Ridings (2016) comprehensively reviewed a number of studies that pair GCM and measurement scale development, one finding was the lack of comparison of the final measurement model via various forms of psychometric testing to the original concept map in order to comment on the relationships between what has been observed, and what had been conceptualized in the map. Most studies did not outline any steps that had been taken to make this final connection, and the few that did, did so on a limited basis. For example, some researchers compared the proportion of items included in the final psychometric model from each of the original concepts on the map to describe how the original conceptual model aligned with the empirical model that had been constructed.

This analysis presents an approach for empirically validating the framework presented by the original conceptual model. The aim is to translate that conceptual model into a tool that will guide evaluators as they think strategically about impact evaluation as it relates to their unique organizational context. But, simply offering it as a framework that had been co-constructed by a small group of professionals in the ILO would not give it the credibility or subsequently, the generalizability necessary for the type of adoption that has any sort of meaning. The methodological literature will benefit from an additional study tying together GCM with psychometric modeling, and back again. Knowing what has been done, and the relevant limitations, allows this research to extend the boundaries in some sense to consider tandem approaches in the future.
Organization of this Dissertation

This dissertation is organized into five chapters. To this point the first chapter has introduced the topic and offered an overview of the issue on hand. The chapter also describes the general purpose of the dissertation by briefly touching on what this research hopes to explore, and some of the questions it hopes to answer. The second chapter provides a comprehensive literature review. The literature review starts with a description of the ongoing debate that has developed around impact evaluation focusing primarily on the period following the CGD (2006) publication of its report that discussed the dearth of strong evidence for the effectiveness of development interventions.

The literature review then examines existing frameworks for evaluation capacity and the methods used in their development. This discussion focuses in part on their relatively new inclusion of issues pertaining to organizational capacity for doing and using evaluation. The next review section describes the key concepts from the GCM study to provide background on the topics, and offers a frame of reference for interpreting the data presented in the results. The last critical section of the review examines studies done that have utilized GCM for the purposes of creating a measurement instrument of some sort. This includes a discussion of methodology and validation to offer empirical support for the approach described in Chapter 3.

Chapter Three begins with a delineation of the research questions followed by the methodology being used to address them, including a discussion of the sampling frame and access to the group. It also includes the survey instrument being proposed for use in
this study along with the methodology used in its development. The data analysis plan and procedures are also presented in this chapter.

The results of the analysis of survey data are discussed in Chapter 4. The first section presents a descriptive analysis of the survey data to examine the perceptions and opinions of respondents to the ideas and concepts asked about on the survey itself. This offers insight into how practitioners understand some of the issues surrounding organizational capacity related to impact evaluation. The second section in the results presents findings from the factor analysis done on the data in order to examine the degree alignment with the original concept map. It will include an assessment of how well the survey findings support the original structure of the concept map.

The final chapter presents an interpretation of the results from Chapter 4. The intent is to present the results in the form of a framework as described in the study’s purpose. The framework is used as the vehicle used to interpret and discuss data from the results chapter.
CHAPTER 2
LITERATURE REVIEW

Introduction

The purpose of this dissertation is ultimately to produce a framework for impact evaluation capacity, focused primarily on the institution level. This chapter presents an overview of key literature related to this dissertation research. In particular, the review of the literature is couched in Rogers and Peersman’s (2014) call for more research on the enabling environment for impact evaluation, that is, “Individual impact evaluations operate within a larger context of local guidance, policy, capacity development and formal and informal incentives” (Rogers & Peersman, 2014, p. 88). It is in this space that this dissertation fits into a broader agenda.

The chapter begins with a brief overview of the quantitative vs. qualitative methods debate in the North American evaluation community, to frame the second part of the literature review that describes the debate beginning in the early 2000’s related to the increased use of impact evaluation to study development assistance. I connect these reviews here because of the many similarities between the two debates. In fact, Picciotto (2012) chided the development assistance community for delving into methodological debates around impact evaluation saying that “evidently the lessons of past evaluation debates had not been internalized by the economics profession...”
(p.215) in an article commenting on the proposition that experiments and quasi-experiments were the gold standard for evidence generation.

I then present a review of the study that lies at the foundation of this dissertation research, and then describe existing literature related to evaluation capacity. A brief discussion about the use of group concept mapping (GCM) for scale development will conclude the chapter, recognizing the role the methodology played in establishing the foundation for this work.

**The Paradigm Wars**

As evaluation evolved, debates around its theory and practice came and went, helping to shape the methods, approaches, and perspectives that we have today. Perhaps the most hotly contested issue was the quantitative vs. qualitative debate that began in the 1980’s (Picciotto, 2012). While the debate has subsided over time, it remains an issue that periodically emerges, whether in response to new publications, approaches, or external forces. For example, the debate re-emerged in 2003 when the Institute for Education Science released new guidelines on study design stating studies using experimental or certain quasi-experimental designs would be prioritized for funding. Although focused in the United States, this statement ignited passions in the evaluation community so much that it prompted dueling statements from within the American Evaluation Association; one opposing the guideline, and another supporting it (Donaldson, Christie, & Mark, 2015).
The development assistance community has been having the same type of debate over the past 10 years. Highlighting some similarities helps to illuminate the backdrop behind the impact evaluation debates in international development.

As the field of evaluation in the United States grew out of the social agenda charted by the Lyndon B. Johnson’s War on Poverty, it fell to social researchers to respond to the demand for analysis and accountability of the investments made to alleviate poverty and address other pressing social issues facing the United States at that time. Shadish, Cook, and Leviton (1991) described the evolution of evaluation as a response to the changes in social policy that were occurring in and after the 1960’s. This included the mandated evaluation of public, federally funded programs, most notably the Title I legislation.

To meet this growing demand, evaluators emerged mainly out of academia, bringing with them the social science training they had received through masters and doctoral programs (Shadish, Cook, & Leviton, 1991). But their practice was diverse, and the approach to evaluation taken by anthropologists was different from the approach of economists. Engagement of academics across disciplines was one reason the evaluation field developed a diverse set of theories and approaches that became the basis for the professionalization of evaluation. This diversity also brought shifts in thinking about evaluation, from a focus on studying outcomes to a more inclusive concern about program quality and implementation; and from a reliance on quantitative methods to the incorporation of qualitative methodologies (Guba & Lincoln, 1981; Shadish, Cook, & Leviton, 1991).
This brief historical perspective is meant to describe how those debates between quantitatively oriented and qualitatively oriented evaluators in North America are not so different from the ongoing discussion in international aid around impact evaluation. Since the Center for Global Development (CGD) released their 2006 publication “When Will We Ever Learn?”, demand for and use of impact evaluation for examining development interventions has grown (Cameron et al., 2015). At the same time quantitatively oriented experimentalists, often hailing from economics, argued that the use of RCTs is the “gold standard” for generating evidence of what works; while evaluators hailing from other social science disciplines argued other models, mainly theory of change oriented approaches, can also be effective ways to explore causality and attribution. The debate has been particularly divisive at times (UNEG Impact Evaluation Task Force, 2013). Many organizations and individuals have attempted to make definitional contributions by articulating what impact means, and how impact evaluation should be done (e.g., 3ie, Abdul Latif Jameel Poverty Action Lab, the World Bank, DFID, AusAID). Because the results of the present research are meant to offer a way to interpret the impact evaluation debate, it is important to discuss that debate and its evolution. The timeline in Figure 1 presents key points in the evolution of the debate and can serve as a reference for some of what is discussed in this chapter.

The CGD (2006) report highlighted the lack of evidence of effectiveness from billions of dollars in investment in development, and recommended the issue be addressed by doing more impact evaluations. The authors of the report suggested that
these impact evaluations ought to be done using counterfactual designs, emphasizing experiments and certain quasi-experiments as the best way to prove what worked, and what did not. The CGD approach to impact evaluation was largely driven by economists, not evaluators (Picciotto, 2012), and the broader evaluation community met the report with some criticism (Deaton, 2010; Picciotto, 2012; Ravallion, 2009).

Arguing that the application of experimental methods was the best way to generate evidence of what works in social programs and policy making (Savedoff et al., 2006), development economists seemed to place less value on the work of evaluators in the field of international development. Their arguments seemed to suggest that the evidence that evaluators had produced was insufficient to clarify what works. The experimentalist approach to impact evaluation rose to prominence quickly (3ie, 2012b; Gertler, Martinez, Premand, Rawlings, & Vermeersch, 2011; White, 2009). Picciotto (2012) writes about the struggle between experimentalists and other aid evaluators. Having not been a part of the mainstream evaluation profession during the earlier paradigm wars, other evaluators were unprepared for the “onslaught” from the

---

2 Discussed later in this chapter. They are often referred to as experimental or quasi-experimental designs at different points in the chapter.
experimentalists, which forced them to give ground (Picciotto, 2012). That description, while perhaps somewhat hyperbolic, captures the way the experimentalist perspective has quickly risen to prominence in international aid.

Figure 1. Timeline of key events related to the rise of impact evaluation in international development.
The Attribution Issue

At the center of the impact evaluation debate is the issue of attribution; that is, observed outcomes that can be said to be a result of a particular program or intervention. It is commonly agreed that impact evaluations should address the issue of attribution and both the peer reviewed literature, in evaluation policies of organizations, and relevant guidance materials support this (3ie, 2012a; Independent Evaluation Group, 2012; Lensink, 2014; OECD, 2002; UNEG Impact Evaluation Task Force, 2013; White & Raitzer, 2017). The debates arise when considering how attribution should or can be assessed, and what types of designs and methods produce credible evidence to support attributional claims (White, 2010). Two main perspectives have emerged: that impact evaluation can be done using only counterfactuals; and that other, factual approaches can assess attribution effectively through theory-driven designs (Leeuw & Vaessen, 2009b; UNEG Impact Evaluation Task Force, 2013).

It is also important to clarify the differences between attribution, and contribution. The terms are often used interchangeably when discussing linkages between programs and their measured effect (Belcher & Palenberg, 2016; UNEG Impact Evaluation Task Force, 2013), but they represent different types of claims. Attribution is most often associated with causal claims that directly link an intervention with impacts (DFID, 2012) and the issue of attribution is most often discussed in studies that utilize counterfactual designs. Contribution is more closely associated with the recognition of programs as parts of larger systems, and with claims that describe their role as one element of those systems in generating impacts (DFID, 2012). Contributory claims are
most often associated with non-counterfactual designs and support a more systems-driven perspective.

Shadish, Cook, and Campbell (2002) describe a counterfactual in this way: “In an experiment, we observe what did happen when people receive a treatment. The counterfactual is knowledge of what would have happened to those same people if they simultaneously had not received treatment” (p. 5). Observing the counterfactual is impossible because we cannot simultaneously engage individuals and not engage them in treatment. Those who argue that counterfactuals are necessary for impact evaluation argue that without them, evaluators face a problem: they do not know what might have happened if a “treatment” had never existed (Lensink, 2014). Adherents to the experimentalist approach argue that random assignment is the most effective way for estimating that counterfactual scenario, an approach that has been termed the “gold standard” by these researchers.

The use of counterfactual designs allows evaluators to answer causal questions about interventions and their observed effects. The control group, when of a certain size, serves as the basis for a best estimation of what may have happened if an intervention were not implemented (Gertler et al., 2011). Random assignment is the most effective way for creating two groups that reduce bias for estimating their equivalence, effectively called selection bias (Shadish, Cook, & Campbell, 2002). Sometimes randomization is not possible or feasible. In this case an evaluator may use a quasi-experimental design, the most sophisticated of which utilize statistical methods to control for that selection bias that randomization mitigates in an RCT.
Similar to the early years of evaluation in North America, evaluators have reacted to the rise of the counterfactual perspective in international development. The issue isn’t necessarily the designs themselves; rather, the unbridled claims by those who advocate for these approaches and claim methodological superiority. Scriven calls this “methodological imperialism” (2015, p. 118), and argues that many scientific disciplines utilize methodologies that do not use experimental designs to make causal claims and do so effectively.

Evaluators who have taken issue with these claims of methodological superiority generally use the OECD/DAC (2002) definition for impact (Stame, 2010; UNEG Impact Evaluation Task Force, 2013; White, 2010). It describes impact as “positive and negative, primary and secondary long-term effects produced by a development intervention, directly or indirectly, intended or unintended” (p. 24). Evaluators who adopt this definition are generally willing to employ evaluation designs that accept observations of the factual as legitimate approaches for establishing cause and effect, in contrast to a counterfactual (UNEG Impact Evaluation Task Force, 2013). Arguments against counterfactual designs include narrowness of scope, and inability to address complexity. Such studies provide answers to a few specific questions about effects (e.g., was there an effect on the intervention group or not?) (Armytage, 2011; Deaton, 2010; Picciotto, 2012; Ravallion, 2009; Stame, 2010). According to the former director of the World Bank’s research department, experiments are insufficient for addressing the knowledge needs of decision makers because they focus only on two parameters: the average effect of the intervention on those who received it, and those who did not. But policy
makers need answers to questions about the way interventions work as well. Cartwright (2011) suggests that RCTs provide insufficient information on generalizability and that policy makers “need evidence for ‘it-will-work-for-us’ claims” (p. 1401), as opposed to its utility elsewhere.

The scope of experimental and quasi-experimental designs is also criticized because it doesn’t adequately address issues of complexity. Many development interventions operate in complex environments where the intervention may be one part of a causal package (DFID, 2012). Counterfactual designs are not well suited to such contexts; Picciotto (2012) writes, “Experimental ‘black boxes’ are poorly suited to the evaluation of complicated or complex programmes in unstable environments. Yet, this is where knowledge gaps are the deepest” (p. 223). This very issue prompted DFID to publish a guide detailing alternative approaches for impact evaluation. Citing their own experience with experimental and quasi-experimental designs, DFID described them as inapplicable to the majority of their funding portfolio (DFID, 2012). At the 2016 European Evaluation Society Conference, Belcher and Palenberg (2016) suggested that counterfactual designs adopt an intervention perspective and are insufficient for addressing issues of complexity, which demand a systems perspective that examines the contribution of a program or policy as part of a larger causal package.

The OECD/DAC criteria set is the most commonly used in international aid. Impact is one of the five OECD/DAC criteria\(^3\) used to develop evaluative claims about

---

\(^3\) The other four include relevance, effectiveness, efficiency, and sustainability
programs and policies. In his analysis of the OECD/DAC criteria as they relate to Scriven’s
definition, Picciotto (2014) places the criterion relevance in relation to the idea of merit,
the criterion effectiveness in relation to the idea of worth, and the criterion efficiency in
relation to value. Counterfactual designs for impact evaluation, he argues, are just one
part of a toolkit. While effective for determining the causal connection between an
intervention and observed results, they do little to address the other criteria for
determining that intervention’s overall value. The question then is whether a pure
assessment of impact is sufficient for making evaluative claims.

Considering whether impact evaluations using counterfactual designs are in fact
evaluation relates closely to the critiques described earlier in this chapter: that such
studies do not provide sufficient or useful information for policymaking. Discussing
challenges for impact evaluation White (2014) points out that many such studies are not
evaluations, and “do not immediately yield information on a programme design that can
be adopted more widely” (p. 21), and writes that “achieving better development
outcomes depends on a whole host of factors that need to be captured in a
comprehensive evaluation of a particular programme. Therefore, the challenge is to
conduct impact evaluations that do indeed evaluate development programmes” (p. 22).

As with the quantitative versus qualitative debates in the United States in the
latter part of the 20th century, the right way to conduct impact evaluations has often
been presented as an either/or scenario. Relative to impact evaluation, attempts to
resolve this debate have included eliminating the either/or proposition by adopting
language that recognizes the strengths of different approaches, and the value that can
be found in combining them. Perhaps the most ambitious effort to do this was by the Network of Networks on Impact Evaluation (NONIE), a collaborative group made up of key actors in international aid. The publication of the Network of Networks for Impact Evaluation (NONIE) manual in 2009 marked what Picciotto (2012) called “an uneasy truce” (p. 216).

The NONIE manual adopted the OECD/DAC (OECD, 2002) definition of impact, thus positioning itself around two underlying premises: the issue of attribution, and the issue of counterfactuals. Asserting that the most important part of designing evaluation is to prioritize the evaluation questions, the NONIE manual indicated that studies ought to be designed in service to the questions, not a particular methodological approach (Leeuw & Vaessen, 2009a). Trying to resolve the debate, Leeuw and Vaessen write in the NONIE guidance document, “The two approaches towards impact evaluation are not mutually exclusive, but overlap at a certain point” (p. 8). The NONIE guide lays out approaches that reflect both camps. Since development interventions range in their size, scope, and complexity, these characteristics should drive decisions about methodology and design. The publication of the NONIE document was important for the field as it served both to recognize that experiments and quasi-experiments are superior for establishing evidence of causation in some circumstances, and to recognize that they are incongruent with some situations where other approaches might be more effective for addressing the attribution question.

A truce does seem to have been struck: the language has shifted to describe the need to make methodological decisions based on evaluation questions, and
programmatic contexts. Evaluators and evaluatees acknowledge the need for increased use of mixed-methods to generate knowledge of how programs work, not just if they do. Opposing perspectives still remain, and how this affects policymaking for impact evaluation and the practice of impact evaluation is perhaps yet to be seen.

**Organizational Policy Making on Impact Evaluation and Definitions**

The impact evaluation debate has centered around methods and designs. That debate and the inconsistency of it seems to have affected the ways the field defines impact and impact evaluation. The two broad definitions for impact that were described above come from the OECD/DAC and 3ie, but variation exists in the definitions of impact (Belcher & Palenberg, 2016). The only commonly agreed upon point is that mixed methods should be employed, and this has been one way to mitigate the paradigmatic debate. Ongoing arguments among scholars have contributed to a lack of cohesion at the organizational level (Picciotto, 2012). In this section, I review some key organizational policies related to impact evaluation to highlight this variation.

The work of the CGD was instrumental in establishing the International Initiative for Impact Evaluation (3ie), an organization that has become one of the most influential actors in the impact evaluation arena for international development. In their guide for grantees, 3ie states they do not advocate a particular approach or methodology for impact evaluation. However, the document suggests that 3ie supported the CGD report perspective that impact should be measured using counterfactual designs as described by their description for eliminating selection bias of a control group, which they indicate
can be done using experimental or quasi-experimental approaches (3ie, 2008). This indicates an assumption that such approaches are preferred. Despite the publication of the NONIE document, there continue to be differences in the ways organizations make decisions about how to define impact, and how to conduct impact evaluation.

Several development organizations have described their approach to impact evaluation in response to the rising demand for its use. The World Bank has been one of the most active institutions supporting the practice of impact evaluation. Establishing the Development Impact Evaluation (DIME) research group is an example of their involvement. DIME is responsible for conducting impact evaluations of World Bank-funded work, as well as the work of other multilateral development banks. The World Bank also funds impact evaluation through their Strategic Impact Evaluation Fund (SIEF). In addition to such structural commitments to impact evaluation, they have published a wide array of material outlining their approach to these types of studies. These materials generally support the use of counterfactual designs, and prioritize quantitative methods (Gertler et al., 2011; Independent Evaluation Group, 2012). The DIME’s short introduction to impact evaluation states, “To be able to make causal statements from the effect of external factors, IEs are based on counterfactual analysis” (p. 1), articulating the general orientation of the World Bank towards impact evaluation (Gertler et al., 2011).

Another major actor in the realm of impact evaluation of development work is the Abdul Jameer Latif Poverty Action Lab (JPAL). As one of the early voices advocating for impact evaluations, JPAL has been a strong supporter of RCTs for evaluation. Esther
Duflo, director of JPAL stated, “Creating a culture in which rigorous randomized evaluations are promoted, encouraged, and financed has the potential to revolutionize social policy during the 21st century, just as randomized trials revolutionized medicine during the 20th” (Atabaki, 2004, p.1). JPAL’s Guide to Evaluation describes evaluation as an approach that uses RCTs to examine the social programs and policies of international development actors (JPAL, 2017).

In contrast to the above, other organizations take a broader view of impact evaluation designs that might be appropriate for their work. As an example, DFIDs’ (2012) guidance document states:

Up to now most investment in impact evaluation has gone into a narrow range of mainly experimental and statistical methods and designs that according to the study’s Terms of Reference, DFID has found are only applicable to a small proportion of their current programme portfolio. This study is intended to broaden that range and open up complex and difficult to evaluate programmes to the possibility of impact evaluation (p. i).

This declaration by a leading bilateral funder of international aid is notable. It at once recognizes that the institution has adhered to the counterfactual design approach to impact evaluation, but have come to realize its limited utility in most of the work DFID supports. The document focuses on theory-driven approaches for impact evaluation that are more responsive to complex settings, highlighting the importance of institutions developing a strategic approach to addressing impact evaluation that makes sense within their context.
UNICEF has also made a large investment in developing guidance materials around impact evaluation. Led by Patricia Rogers, UNICEF published a series of briefs written by various authors addressing different aspects of impact evaluation (Rogers, 2014b). The effort is a thorough attempt to provide practitioners with a clear guide for decision making on key decisions one must make when designing, planning, and implementing impact evaluations. While recognizing the ongoing debate, the documents offer guidance for decision making regarding types of impact evaluation that are appropriate in different settings, in contrast to prioritizing one approach as superior. Discussing the issue of attribution in brief no. 6, Rogers (2014a) writes, “Several different strategies may be used to undertake causal attribution, each of which has its own particular strengths, limitations and suitability according to the specific programme and evaluation context” (p. 4). Three broad approaches are outlined, including counterfactual approaches, consistency of evidence with causal relationship, and ruling out of alternatives.

Other organizations have published various guides and manuals outlining their own position on impact evaluation. A review of the purposes and definitions for impact evaluation in these different materials revealed agreement on some fundamental aspects while simultaneously illustrating the divergence in conceptual clarity that was discussed earlier in this chapter. A sample of some of these definitions and purposes can be found in Table 1.
Table 1. Stated purposes for and definitions of impact evaluation among a sample of organizations

<table>
<thead>
<tr>
<th>Organization</th>
<th>Referenced Publications</th>
<th>Stated Purpose of Impact Evaluation</th>
<th>Definitions of Impact Evaluation or Impact</th>
</tr>
</thead>
<tbody>
<tr>
<td>AusAID</td>
<td>Impact Evaluation: A Discussion Paper for AusAID Practitioners (2012)</td>
<td>Establish value of innovative and effective programs including pilots • Test causal logic of existing programs • Generalizability • Prove worth</td>
<td>Could extrapolate that Impact is a treatment effect (e.g., income, productivity, poverty).</td>
</tr>
<tr>
<td>DFID</td>
<td>Broadening the Range of Designs and Methods for Impact Evaluations</td>
<td>Accountability • Generalizability • Learning</td>
<td>An impact evaluation is a systematic and empirical investigation of the impacts produced by an intervention—specifically, it seeks to establish whether an intervention has made a difference in the lives of people. (p. 2)</td>
</tr>
<tr>
<td>JPAL</td>
<td>Introduction to Evaluations</td>
<td>Learning</td>
<td>“to demonstrate that development programmes lead to development results, that the intervention has cause and effect.” (p. 1)</td>
</tr>
<tr>
<td>UNEG</td>
<td>Impact Evaluation in UN Agency Evaluation Systems: Guidance on Selection, Planning, and Management</td>
<td>Learning (effectiveness) • Accountability</td>
<td>Impact evaluations measure program effectiveness typically by comparing outcomes of those who received the program against those who did not.</td>
</tr>
<tr>
<td>USAID</td>
<td>Technical Note: Impact Evaluations (2013)</td>
<td>Effectiveness • External validity • Testing untested programs • Upscaling</td>
<td>References the DAC definition (p. 6) and the 3ie definition (p. 7).</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>State that most UN agencies have adopted DAC definition (p. 6).</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Measure the net change in outcomes amongst a particular group, or groups, of people that can be attributed to a specific program using the best methodology available, feasible and appropriate to the evaluation question(s) being investigated and to the specific context. (p. 1).</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Those that measure the change in a development outcome that is attributable to a defined intervention.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Based on models of cause and effect and require a credible and rigorously defined counterfactual... (p. 1).</td>
</tr>
<tr>
<td>Organization</td>
<td>Referenced Publications</td>
<td>• Stated Purpose of Impact Evaluation</td>
<td>Definitions of Impact Evaluation or Impact</td>
</tr>
<tr>
<td>------------------------------------------</td>
<td>-----------------------------------------------------------------------------------------</td>
<td>---------------------------------------------------------------------------</td>
<td>-------------------------------------------------------------------</td>
</tr>
<tr>
<td>Bill and Melinda Gates Foundation</td>
<td>Evaluation website <a href="https://www.gatesfoundation.org/How-We-Work/General-Information/Evaluation-Policy">https://www.gatesfoundation.org/How-We-Work/General-Information/Evaluation-Policy</a></td>
<td>• Upscaling • Replication • Pilot testing</td>
<td>• Provide information about the impacts produced by an intervention. • Use OECD/DAC definition of impact</td>
</tr>
<tr>
<td>UNICEF</td>
<td>Overview of Impact Evaluation (2014) &amp; Overview of Strategies for Causal Attribution</td>
<td>• Formative purposes • Summative purposes • Inform decision making</td>
<td>• Adopt the OECD/DAC definition of impact • Impact evaluations should quantitatively measure the attribution of IFAD programmes to the observed outcome. • IFAD highlights quasi-experiments and experiments as the approaches that should be used in impact evaluation.</td>
</tr>
<tr>
<td>IFAD</td>
<td>Evaluation Manual: Second Edition</td>
<td>• Accountability • Learning</td>
<td></td>
</tr>
</tbody>
</table>

Not all these guidance documents name authors, nor are they all official policy documents. Those that do name authors seem to be products of contractual work undertaken with the intent of producing a guide for stakeholders within a particular organization (e.g., DFID, the Asian Development Bank, UNICEF). Who is chosen to develop these documents is important to consider because the choice often affects the approach to impact evaluation the document describes. Authors of these documents can be traced to other publications discussing impact evaluation in both the peer-reviewed literature, and other guidance-type materials. Examining these connections begins to shed light on the types of orientations contracting organizations may already...
have, or may be looking to take towards impact evaluation. For example, DFID (2012) states that the organization has realized that experimental and quasi-experimental approaches have extremely limited applicability for the array of projects in their funding portfolio. This prompted them to seek out knowledge of other approaches for evaluating impact. As in any field, certain key actors are often asked to produce material around impact evaluation. Being aware of the author(s) of a guidance document can sometimes be enough information for understanding what approach to impact evaluation that document will take.

There seems to be a consensus that impact evaluations should be used for accountability and learning purposes. But descriptions of how impact evaluation should be done vary across organizations. Much of that variation can be traced back to the debates among scholars about how to generate evidence of attribution for impact evaluation. Because some of those scholars have been engaged in producing some of the materials reviewed in this section it is perhaps unsurprising that those views are reflected. But even policy documents produced by evaluation offices indicate specific positions on the issue. The ongoing debate, while settled to some extent, makes it difficult for practitioners, funders, and evaluators to determine how best to proceed.

**The Conceptual Underpinnings**

A project conducted at the International Labour Organization (ILO) to establish a conceptual framework for impact evaluation capacity serves as the basis for this research (Smith & Kane, manuscript in preparation), in response to a call by the governing body for the evaluation office to address the growing demand for impact
evaluation. Until that point impact evaluation was included in the evaluation infrastructure of the organization, but it was decentralized, so that it was the responsibility of individual departments to plan, fund, implement, and manage. In their approval of the 2013-14 annual evaluation report the governing body stated:

EVAL [the ILO evaluation office] needs to work with the technical departments to ensure that they use the Office’s guidance, established definitions and tools, in addition to conducting evaluability assessments to ensure quality and to justify investments. EVAL needs to take on a more proactive role by producing an additional protocol that outlines impact methodologies and approaches that are based on international and United Nations Evaluation Group best practices and standards. Moreover, EVAL could strengthen its role by improving knowledge sharing, advocacy, technical support and quality assurance for impact evaluations (p.20).

In response to this mandate, the evaluation office developed a three-part strategy that would guide the office’s work, and outlined how the office would support departments in designing, conducting, and using impact evaluations. The first part of this strategy was to develop a guiding conceptual framework built upon the ideas and perceptions of key stakeholder groups in the ILO. It was important for EVAL to recognize the work and expertise that already existed in the organization. To do this, Group Concept Mapping (GCM) was selected as the methodological approach for completing that part of the strategy.
GCM is a mixed-methods approach that solicits the ideas and opinions of a group and uses multivariate statistical analysis to synthesizes the way they perceive those ideas in order to produce a two-dimensional conceptual map. The resulting map serves as a visual reference for the group to use in examining their collective knowledge and perception around a particular construct (Kane & Rosas, 2018; Kane & Trochim, 2006; Trochim, 1989). GCM has been applied extensively in health, social work, and human services research as well as in evaluation contexts. Introducing GCM as a tool for planning and evaluation Trochim (1989) notes the difficulties associated with conceptualizing an issue, problem or topic, compounded when an entire group is engaged in the process. GCM is a systematic and empirically based participatory process for drawing upon the knowledge of individuals to develop a conceptual framework.

The results from this study yielded eight conceptual clusters that formed two distinct regions on the map: Context and Strategy, and Knowledge and Capacity, each of which consisted of four conceptual clusters. The Context and Strategy region included Foundation and Support, Rationale, Definitions and Communication, and Context and Utilization, representing institutional and structural issues related to developing impact evaluation capacity. The Knowledge and Capacity region included Integration and Harmonization, Knowledge and Asset Mapping, Capacity Building, and Resources, containing operational issues for developing impact evaluation capacity within the institution. These conceptual domains are well represented in other evaluation capacity frameworks, many of which include both structural components and resource components (Cheng & King, 2016; Cousins, Goh, Elliott, Aubry, & Gilbert, 2014; GAO,
2003; King & Volkov, 2005; Labin, Duffy, Meyers, Wandersman, & Lesesne, 2012; Taylor-Ritzler, Suarez-Balcazar, Edurne Garcia-Iriarte, Henry, & Balcazar, 2013). The similarities between the conceptual framework from the study of impact evaluation at the ILO and other evaluation capacity frameworks suggests that this work can be described as an impact evaluation capacity framework.

Before further discussing evaluation capacity frameworks and models, it is important to make a distinction between evaluation capacity, and evaluation capacity building (Cheng & King, 2016). Evaluation capacity is described by Cheng and King (2016) as the ability both to conduct evaluation and use it. It has been discussed and described at various levels—the individual, the organization, and the societal level (Nielsen, Lemire, & Skov, 2011). It is also important to recognize that there are distinct capacity-related issues for evaluators, evaluation managers, and clients.

Evaluation capacity building (ECB) is the process that leads to evaluation capacity. It is most often conceptualized as an activity that is separate from actually conducting evaluations (Labin et al., 2012). Although some evaluators may incorporate elements of ECB into their evaluation work, existing models of ECB treat it as a separate exercise. In a 2002 New Directions in Evaluation issue Stockdill, Baizerman, and Compton proposed the following conceptual definition for ECB calling it, “a context-dependent, intentional action system of guided processes and practices for bringing about and sustaining a state of affairs in which quality program evaluation and its appropriate uses are ordinary and ongoing practices...” (p. 8). The definition is notable because of its institutional focus; it describes a system, as compared to the
competencies of an individual. In another study of the ECB literature, Labin et al. (2012) describe it as, “an intentional process to increase individual motivation, knowledge, and skills, and to enhance a group or organization’s ability to conduct or use evaluation” (p. 308). In this description it is the idea of the “intentional process” that would presumably result in evaluation capacity for that group or organization. The conceptual framework in Smith and Kane (manuscript in preparation) that guides this framework relates more closely to the models of evaluation capacity, so that topic will be discussed more fully here. Evaluation capacity building is part of the system, but should be recognized as a separate process that is part of developing EC.

Evaluation capacity is related to Stevahn, King, Ghere, and Minnema’s (2005) work on developing evaluation competencies that outlines a set of core skills and abilities that evaluators should have to practice effectively. This work has become the foundation for a wide range of evaluation competency work that is ongoing (e.g., Aotearoa New Zealand Evaluation Association, 2011; Canadian Evaluation Society, 2010; Podems, Goldman, & Jacob, 2013).

The Sustainable Development Goals (SDGs), adopted in 2015 by the member states of the UN, has renewed focus on building evaluation capacity on a national level. In a general assembly resolution (26/237) the UN (2015) reiterated the importance of building national evaluation capacity as a way to support the achievement of the SDGs. The resolution also instructed UN agency evaluation offices to support the development of national evaluation capacity where requested. Evaluation’s role in the SDGs was further highlighted by the designation of 2016 as the Year of Evaluation. How the
evaluation community would respond to this demand to develop national evaluation capacity became an important topic for the international development evaluation community.

As with ECB models, evaluation capacity have moved beyond a focus on individual competencies, accounting for organizational factors as well (Bourgeois & Cousins, 2013; Carman & Fredericks, 2010; Cheng & King, 2016; GAO, 2003; King & Volkov, 2005; McDonald, Rogers, & Kefford, 2003; Pattyn, 2014). The evaluation field recognizes that individuals are limited in their ability to advance evaluation practice and use, and that structural components affect an organization’s ability to achieve the goal of effectively doing and using evaluation (Cheng & King, 2016). In an examination of evaluation capacity in the Canadian public sector, Bourgeois and Cousins (2013) developed a framework that consisted of two broad areas; the capacity to do evaluation, and the capacity to use evaluation. Capacity to do evaluation included three dimensions: human resources, organizational resources, and evaluation planning and activities. Capacity to use evaluation included evaluation literacy, integration with organizational decision making, and learning benefits. To develop this framework, the authors conducted an in-depth review of the literature as well as three rounds of expert interviews. The process methodology was qualitative and provided detailed information about the various component areas.

Using a case study approach of five federal government agencies in the United States, the Government Accountability Office (GAO) (2003) constructed a framework outlining key elements of agency evaluation capacity. These included evaluation culture,
data quality, analytic expertise, and collaborative partnerships. Their framework also included strategies for developing each of these elements. The intent of this study was to offer other agencies a conceptual framework for considering their own evaluation capacity, and concrete examples of how aspects of that framework had been addressed by the five agencies included in this study.

In another study of evaluation capacity in government and voluntary sectors in Canada, Cousins et al. (2014) developed a questionnaire based upon a conceptual framework for evaluation capacity. The framework was also developed in part by using GCM to study evaluation use in government. It revolved around the capacity to do evaluation, the capacity to use evaluation, the organizational culture around learning, and the skills and abilities of individuals. The study examined differences in perceptions of professionals in the government and voluntary sectors related to aspects of evaluation capacity within their organizations.

The evaluation capacity of the nonprofit sector in the United States was examined by Taylor-Ritzler et al. (2013) using a survey of nonprofit organizations. The survey was based upon a model of evaluation capacity developed by Suarez-Balcazar et al. (2010) who had used a review of literature to create it. The survey results were analyzed using structural equation modelling (SEM) to identify first and second level factors. The intention was to use that data to validate the original model. This approach has similarities to the approach being taken in this research, except that this research is premised on a conceptual framework developed through GCM with input from expert practitioners and their clients. The results of their analysis identified two second-order
constructs: individual factors and organizational factors. Each of those was made up of three first-order constructs. They found that both of the second-order constructs were also related to the mainstreaming of evaluation, and the use of results.

Using a similar methodology in a study of evaluation capacity in the Danish public sector Nielsen, Lemire, and Skov (2011) developed a model that described two primary components, evaluation demand, and evaluation supply. Using confirmatory factor analysis, they determined that the articulated framework was in fact represented by the structure of the data they collected. Evaluation demand included two components, objectives, and structure and processes. Evaluation supply included technology and human capital. They conclude that their model is a plausible approach for measuring evaluation capacity.

EvalPartners has emerged as a major actor in evaluation internationally by advocating for the development of national evaluation capacity, supporting what they call Voluntary Organizations for Professional Evaluation (VOPE), and supporting the development of new talent in the evaluation field through their establishment of EvalYouth. EvalPartners’ global agenda document covering 2016-2020 presents a prospective model for how evaluation can positively affect outcomes globally. Three of the four primary components in the model are related to evaluation capacity concepts including the enabling environment for evaluation (e.g., evaluation culture, evaluation demand, evaluation use), institutional capacities, and individual capacities for evaluation (EvalPartners, 2016, p. 7). This further highlights the importance of evaluation capacity in the broader evaluation community.
Understanding evaluation capacity at a national level is an important first step in addressing the issue. To do this, the Independent Evaluation Office (IEO) of the United Nations Development Programme (UNDP) conducted a study of national evaluation capacities in 43 countries in 2014, which was presented at the Conference on National Evaluation Capacities in Bangkok, Thailand in 2015. UNDP describes capacity similarly to Cheng and King (2016); i.e., the ability to do, and use evaluation. Here, capacity refers to “…creating an enabling environment in which evaluations can be determined or required and the way in which they are used as credible and independent function to inform national-level decision-and policymaking” (UNDP, 2015, p. 59). The study employed a framework that focused on institutional structures for supporting evaluation; e.g., budgets, the administrative systems to manage evaluation, evaluation culture, stakeholder involvement, technical abilities, and the inclusion of gender, culture, and ethics (UNDP, 2015).

This review of evaluation capacity frameworks, models, and concepts is intended to demonstrate both variation among frameworks, and their commonalities. Multiple models have been put forth in different contexts, using different methodologies. Some core features carry over across this sample of frameworks. The first is the need to have individuals who have the appropriate capacity to do, manage, and use evaluation. This includes training and development of staff who are not evaluators themselves, but who might be responsible for commissioning evaluations and/or making decisions based upon their results. Structural components that are important for facilitating the doing and using of evaluation include processes that support evaluation (e.g., data collection
systems, monitoring systems, integration of evaluation planning into program design), policies that encourage evaluation, and leadership that champions evaluation. These commonalities suggest the existence of common critical features for evaluation capacity, but the need for developing setting-specific frameworks exists to address contextual and use distinctions. The operationalization of these conceptual elements is where important differences may emerge.

**Group Concept Mapping and Measurement**

The use of GCM for scale development has been well documented. A systematic review of research in which GCM was paired with scale development identified 23 studies in the peer reviewed literature that had utilized such an approach (Rosas & Ridings, 2016). The studies used substantively different approaches to the GCM process for developing instruments from the GCM data, and for validating those instruments. Procedurally, each study utilized a similar sequence of steps: 1) conduct a GCM study with a group of key stakeholders, 2) utilize the GCM findings to develop a measurement instrument, and 3) validate that measurement instrument with a larger group of respondents. The presence of these studies in the published literature supports this general approach to instrument development (Rosas & Ridings, 2016).

The basis of a GCM investigation typically revolves around 100 unique ideas. The approaches taken to create the measurement instruments most often began with an item reduction process to identify key ideas from the group concept map to ensure that the most important ideas were included (Rosas & Ridings, 2016). A number of these studies offered detailed descriptions of the methods taken to reduce the item set. Some
studies relied on expert review, interviews, or an examination of related literature in order to select ideas to include on the instrument (Conrad, Iris, Ridings, Langley, & Anetzberger, 2011; Jordan et al., 2013), and others used the ratings data from the concept mapping to inform decisions about which ideas to include in the instrument (Corcoran, 2005; Wallace, Wexler, Miser, McDougle, & Haddox, 2013).

The psychometric testing of these various instruments from studies detailed by Rosas and Ridings (2016) varies. Factor analysis was the most commonly used technique, both exploratory and confirmatory. Other approaches included item response theory and Rasch analysis.

Rosas and Ridings found a number of strengths in applying GCM to the development of measurement tools identified in their review of various research. First, it establishes content validity because of its reliance on “experts” for the generation of content. It is also important to have a clear conceptual framework when developing a measurement instrument. The use of GCM ensures that this conceptual framework is established by the experts and can be used as a guide for developing content. Second, GCM facilitates researcher decision making by offering a systematic approach to identify and understand what should be included in an instrument that is intended to represent the conceptual framework from GCM. Third, the use of GCM ensures that the conceptual framework is stakeholder driven. The views reflected in the resulting conceptual framework depend on who was engaged in the GCM process, but it offers the potential of ensuring breadth of content. This is in contrast to a conceptual
framework that might be researcher driven and generated mainly through a review of existing literature.

Rosas and Ridings (2016) note a primary weakness in the cited examples of GCM use of GCM for scale development: examples demonstrated little or no exploration of the link between the empirical results of the instrument development process, and the original group concept map. Circling back in this way can offer revision and refinement of the conceptual framework, and would be recommended especially since instruments are often meant to measure aspects of the group concept map, which is meant to serve as a guide to identifying theoretical constructs of interest. Taking the opportunity to reflect the selected items back to the map would clarify or confirm whether the linkage is meaningful.

**Chapter Summary**

This chapter has presented the key literature that relates to the research relevant to this dissertation. Briefly exploring the quantitative vs. qualitative debates that had taken place in the North American evaluation community in the latter half of the 20th century, I contextualized the ongoing debate in the international aid community about methods and approaches for evaluating impact. Two general groups of thought were presented advocacy for experimental or quasi-experimental designs, and those who argue for theory-driven approaches as acceptable to examine issues of attribution. As the debate continued among scholars and thought leaders, evaluators, evaluation offices, and evaluation funders have been determining how to move forward with
integrating impact evaluation into their own work. Here we described policies and publications that some major international organizations have developed to illustrate the variation in how impact evaluation is described and defined. Whether this is a result of the ongoing ambiguity that is trickling down from the intellectual elite is unclear, but it is useful to provide as background for this research.

Moving away from the realm of impact evaluation the chapter then describes key literature related to evaluation capacity; specifically, models and frameworks developed by other scholars. Since the objective of this research is intended to produce a model specifically related to impact evaluation and focused at the institutional level, demonstrating other models provides a system view. Such models are found to have some key commonalities, but also demonstrate the importance of specific models for specific contexts.

The chapter ends with a brief discussion of the use of GCM in developing measurement scales, setting the stage for the main analysis in this research. It also provides a reference for some of the content described in Chapter 3.
CHAPTER 3

METHODOLOGY

The primary data collection for this research was done using a web-based survey to answer the following questions:

1. What components of institutional impact evaluation capacity development do international development practitioners and evaluators believe are most important?

2. In what ways do the constructs measured by the survey instrument relate to concepts that were present in the original concept map?

3. What kind of framework do the results of this research lend themselves to?

The description of the methodology begins with a discussion of the sampling procedures that were used for data collection. The survey development process is then described followed by a description of the data analysis and interpretation procedures.

This research explored the results from a group concept mapping (GCM) study done at the International Labour Organization (ILO) in the winter of 2016 (Smith & Kane, manuscript in preparation). The first purpose of this study was to investigate the perceptions of evaluators, international development practitioners, and evaluation
funders as they pertain to the concepts, ideas, and issues that emerged in the original GCM study at the ILO.

The second purpose of this research was to explore how the underlying constructs from the survey align, or do not align with the original conceptual framework from the concept map, and to use those results to comment on the validity of that original framework. Using the results from both the original concept map and the research done for this dissertation a framework was developed for organizational impact evaluation capacity. That framework was then explored further with a group of evaluation experts that reviewed the initial iteration of the framework and offered feedback on its utility and relevance for practitioners.

**Study Design**

This research was cross-sectional in nature. The data collection process sought out responses only once from each respondent making the results a snapshot in time. There were no other notable design features for this research.

**Sample and Sampling Procedure**

This study utilized a census of three evaluation listserv's focused on international development-related evaluation work and topics. These included, 1) XCeval, 2) Monitoring and Evaluation (M&E) News, and 3) the Pelican Initiative. The XCeval listserv was originally established by the International and Cross-Cultural topical interest group (TIG) of the American Evaluation Association (AEA). It is open to anyone interested in evaluation of international development, and cross-cultural issues in evaluation. The M&E
News listserv is managed by Rick Davies and operates as a space for evaluators and professionals in the international development field to exchange ideas, and engage in discussion about M&E related issues. It is also a space where trainings are announced, publications are disseminated, and jobs are posted. The Pelican Initiative is a listserv dedicated to approaches for evidence generation for international development interventions, and how that evidence is used for decision making. It was originally started by the International Development Research Centre (IDRC), the European Centre for Development Policy Management (ECDPM), Exchange, Bellanet, and the UNICEF East Africa Regional Office.

The subscribers to these listservs are professionals in the development sector who were either directly or tangentially involved in monitoring and evaluation work. Some of the subscribers were working as internal evaluators within evaluation offices of development agencies, others were consultants who use the listserv as a space to stay up-to-date on potential requests for proposals, and others were evaluation managers. There are semi-regular discussions through the listserv among members regarding various evaluation practice related issues.

It was important to collect data from a diverse pool of respondents so that multiple viewpoints on the topic were represented in the results. Listservs were chosen for this study because they are offered access to a wide range of evaluation practitioners working in international development. The sampling strategy sought out respondents who may have varying degrees of experience with impact evaluation, who are operating
in different contexts, and who may fill different evaluation related roles related (e.g., evaluation managers, evaluation consultants, evaluation funders). In Chapter 1, part of the justification made for this study revolved around how the lack of clarity surrounding impact evaluation might be detrimental to these three groups. Therefore, it was important that the voices and perceptions of members of those groups be included in the results. While a limitation of this approach was that some respondents were less familiar with impact evaluation, as an evaluation approach, it was not the purpose of this research to explore the technical expertise of respondents.

Unlike some research on evaluation (RoE) that has been conducted using the American Evaluation Association member database (e.g., Coryn et al., 2016; Fleischer & Christie, 2009) demographic information was not available for the subscribers to these listservs. It was not possible to assess the characteristics of the population before administering the survey and so the survey was disseminated through each of the listservs in a census type approach. If probability sampling had been used it would have been difficult to describe the representativeness of the sample given the inability to examine characteristics of the population as a whole.

Factor-analytic methods were a cornerstone of this research. Key parameters for appropriate application of these analyses had to be considered when developing the sampling procedures. Worthington and Whittaker (2006) offer a series of guidelines regarding sample sizes for both exploratory and confirmatory factor analysis. Sample sizes for factor analysis have formerly been discussed as a function of the number of items that
make up an instrument. However, more recent research suggests that this general guideline may not necessarily hold true (MacCallum, Widaman, Zhang, & Hong, 1999). That said, that relatively simple rule requiring a 10:1 respondent-to-item ratio was the benchmark used when developing the sampling approach for this study.

The instrument used in this research contained 42 items. Following guidelines from the literature cited above a successful factor analysis required approximately 420 responses. Because the full population was unknown, the threshold for responses was based upon the number needed to successfully conduct the analysis, as opposed to a pre-determined response rate. The 420 responses served as the minimum threshold of responses needed, and the targeted number of responses was 500, to account for any responses that might be only partially completed given the difficulty in using responses with missing data in factor-analytic methods.

**Participant Selection for Expert Review Panel**

The results of the survey analysis were linked back to the original conceptual map to develop an initial framework for organizational impact evaluation capacity development. To improve the face validity of the framework as a product, and to solicit input on its utility, an expert panel was convened to provide feedback on the draft framework. Reviewers were selected based upon their experience and expertise working in the field of international development evaluation. It was important to have reviewers who had experience and expertise with impact evaluation specifically and who had demonstrably contributed to the literature on the topic.
Because the results of this research were aimed at producing a framework for impact evaluation capacity development it was also important to have reviewers who understood evaluation capacity more broadly. All of these individuals had published on the topic of evaluation capacity, but only some had direct experience with impact evaluation. It was important to solicit their input into this model to determine how it did or did not reflect other evaluation capacity frameworks, and to better understand how this model was different.

*Communication with Survey Sample*

All communication with the individuals included in the survey sample was done via email. As described earlier the survey was disseminated across the three listservs, and a separate survey was used for each listserv so that responses could be tracked by listserv. The three surveys were identical in structure and content. Email communications were drafted and submitted as part of the human subject institutional review board (HSIRB) application for this research.

*Instrument Development*

The 2016 ILO GCM study yielded 101 individual ideas about what the organization should consider to further define, develop, and/or apply impact evaluation in its work. These ideas were rated on an importance scale. An item reduction process was taken to reduce the set of 101 ideas to 1) ensure that the survey focused on the highest value ideas from the original concept map, and 2) to minimize respondent fatigue. A review of approaches used in previous studies of this type showed that most researchers employed some process for item reduction (Rosas & Ridings, 2016). Some studies relied on expert
review, interviews, or an examination of related literature to select ideas for inclusion in an instrument (Conrad et al., 2011; Jordan et al., 2013), others used the ratings data from the concept mapping to inform those decisions about which ideas to include (Corcoran, 2005; Wallace et al., 2013). This study followed the latter approach, using the data from the GCM process to inform the item reduction process. It was important that key ideas from the original map be reflected in this instrument to address the second purpose of this research: to examine the validity of the concepts in that original map.

*Item Reduction*

The item reduction process began with a descriptive analysis of the average importance ratings for each idea. As shown in Figure 2, the average importance rating across all items was 4.0 (on a scale of 1-7). All items with an importance rating lower than 4.0 were removed. This resulted in the elimination of 50 items, leaving 51 in the idea set.

The remaining 51 items were reviewed to remove redundant ideas, or to modify or eliminate statements that were overly specific to the ILO. For example, the following statement was eliminated because of its specificity: “The ILO should start investing more in tracking the impact of advice on normative work and evaluating changes at the policy level following the support of constituents within specific national contexts.” If the conceptual underpinnings of an idea were more broadly relevant, but the wording was too specific to the original context then the item was reformatted to make it more broadly relevant. For example, the original idea in the GCM data read, “Develop a network or mechanism within ILO to provide consistent and full peer review and guidance on impact
evaluation to ensure consistency and knowledge transfer”, and was revised to read, “organizations should facilitate the development of an internal network to conduct peer review of impact evaluations.” This review and modification process resulted in the elimination of nine additional ideas found to be repetitive or irrelevant.

Figure 2 shows the item reduction process resulted in the elimination of 59 of the 101 original ideas, leaving 42 for integration into the survey. Table 2 presents the number of items by conceptual cluster and the original number of ideas from that cluster. During the transformation of the group concept mapping ideas into survey items wording was sometimes adjusted to improve clarity, references to the original organization was removed, and compound ideas were split into multiple items.
Table 2. Proportion of ideas in original map and reduced item set by conceptual cluster

<table>
<thead>
<tr>
<th>Cluster title</th>
<th>Original</th>
<th>Reduced</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
<td>%</td>
</tr>
<tr>
<td>1. Rationale</td>
<td>9</td>
<td>9%</td>
</tr>
<tr>
<td>2. Definitions and communication</td>
<td>16</td>
<td>16%</td>
</tr>
<tr>
<td>3. Context and utilization</td>
<td>17</td>
<td>17%</td>
</tr>
<tr>
<td>4. Integration and harmonization</td>
<td>13</td>
<td>13%</td>
</tr>
<tr>
<td>5. Knowledge and asset mapping</td>
<td>15</td>
<td>15%</td>
</tr>
<tr>
<td>6. Capacity building</td>
<td>11</td>
<td>11%</td>
</tr>
<tr>
<td>7. Resources</td>
<td>5</td>
<td>5%</td>
</tr>
<tr>
<td>8. Foundation and support</td>
<td>15</td>
<td>15%</td>
</tr>
</tbody>
</table>

The original 101 ideas were rated on a 7-point Likert scale anchored by 1-Not at all important and 7-Critical. That same scale was adopted for this research to facilitate comparisons that would be made in the analysis. While the comparison was imperfect given the modifications made to idea wording during item development there was some ability to examine differences or similarities between the two studies particularly at the conceptual level.

To conduct sub-group analysis of the survey data that would answer the secondary research questions associated with research question 1, a demographics section was also developed. These items were developed to ensure the secondary questions were fully addressed. These items asked respondents about the following:

1. The experience level with impact evaluation
2. The number of years each respondent has been engaged in evaluation of development work

3. Respondents organizational affiliation

4. Respondent’s current professional position as it pertains to evaluation

5. Respondent’s methodological orientation

6. The geographic location where each respondent practices

The full survey can be viewed in Appendix A. The survey was developed in the Qualtrics survey platform for dissemination and sent out via email. Data collection lasted one month. The data were then reviewed to determine suitability for the analysis described in this section.

Procedure

All necessary steps to protect human subjects were taken in accordance with the Western Michigan University Human Subjects Institutional Review Board (HSIRB). This was done prior to recruitment or data collection activities. The full HSIRB application can be viewed in the appendix of this dissertation.

The survey was launched on January 22nd, 2018. This launch date was selected to account for any participants who may have been on extended holiday. Respondents were sent an email (see text in the HSIRB application in appendix) the week before January 22nd describing the purpose of the study and informing them that the survey would arrive in their inbox in one week. Respondents were also offered an opt-out at this time. Any respondents who opted out were removed from the contact list and not contacted again. Once the survey launched it was left open for four full weeks. Weekly
reminders were sent out via email to encourage respondents to complete the survey. Individuals who completed the survey were not sent these follow-up emails and instead received a separate message thanking them for their input. At the end of the four weeks the survey was closed, and data cleaning began in order to conduct data analysis.

**Survey Analysis**

The survey data was analyzed using appropriate statistical software and data analysis tools. Microsoft Excel sufficed for most of the analysis done for this study. Minitab was used for any significance testing done to determine meaningful differences between subgroups. Two analytic approaches were taken to examine the data set collected during survey administration to support the purposes outlined in Chapter 1. These two approaches are described in more detail in this section.

**Survey Analysis for Examining the Perceptions and Opinions of Respondents**

This first set of analyses aimed to answer the first research question: What are the attitudes and perceptions of international development practitioners and evaluators around organizational capacity for conducting impact evaluation? The sub-questions are as follows:

- Sub question 1.1: What characteristics of respondents explain differences or variations in response patterns (e.g., role, training background, experience) to how organizations should prepare themselves to integrate impact evaluation into their work?
- Sub question 1.2: How do the attitudes and perceptions differ or align with the results from the original concept mapping study?
• Sub question 1.3: In what ways does the level of experience of respondents affect their perception of what is more or less important for organizations to consider when developing an impact evaluation strategy?

The data were cleaned and missing data was dealt with as appropriate. For this part of the data processing any partially completed surveys were included for item level analyses. Descriptive statistics were produced and interpreted to present an overall analysis of responses in relation to the primary research question. The analysis relied on the conceptual framework from the original concept map as a way of organizing and interpreting the results. Responses to groups of items were presented according to the clusters from the original concept map.

Beyond the descriptive data, subgroup analysis was conducted where appropriate sample sizes existed to address the sub questions described above. These questions were aimed at determining whether meaningful variation existed in the response patterns of groups based on the demographic items included in the survey. Significance testing was conducted to determine whether differences between groups were statistically significant. Examining the data using these different lenses allowed for a richer understanding of the ways respondents prioritized and perceived the ideas from the concept map. Understanding how evaluators engaged in evaluation of international development initiatives understand organizational capacity to conduct impact evaluation offers insight that has not been studied in a systematic way.
Factor Analysis of the Survey Data

The second primary question posed in this dissertation was, *how does the conceptual structure from the concept map manifest in the factorial structure of the survey instrument?* The sub questions associated with this research question included:

- Sub question 2.1: Is the 8-factor structure from the concept map supported by the data collected via the survey?
- Sub question 2.2: How do the results validate or invalidate the framework from the original concept map?
- Sub question 2.3: What are the implications for utilizing this approach for validating other concept maps?

Shadish, Cook, and Campbell (2002) describe a construct as a concept. Eight concepts were presented on the concept map from Smith and Kane’s (manuscript in preparation) study at the ILO. The concepts in that map are groups of ideas that were sorted together by a group of “content experts” and therefore hold some content validity in the setting where they were created. To what extent the ideas, concepts, and the framework itself have validity in a broader context was the question that this part of this research aimed to answer.

Although there was a hypothesized presence of an overarching factor framework (i.e., the conceptual framework from the map), that might suggest that confirmatory factor analysis (CFA) would be appropriate for answering these questions (Batterham et al., 2002; Neff & Paulson, 2011), exploratory factor analysis (EFA) was selected instead.
There were two primary reasons for this decision. The first was that there were not enough responses collected to run a valid CFA. But there was also a logical argument for selecting EFA as the analytic approach for this analysis.

The process used to develop the framework for impact evaluation capacity started with a group concept map. That formed a theoretical model for impact evaluation capacity (Rosas, Behar, & Hydaker, 2014), and helped to construct under-representation in the content it represented. The use of EFA allowed for data reduction to identify the most important of the 42 items in the instrument, and the generation of an observed model. The results of the EFA provided key insights into the conceptual components of the model and helped to better articulate the framework for impact evaluation capacity. That model can now be confirmed through CFA, although that is outside the scope of this study. In this way the EFA serves as the lynchpin between the theoretical model and the fully confirmed model.

For the factor-analysis no responses with missing data were included. The results from the model fit and the emergent factor structure were eventually compared against the original concept map. The results of the EFA were used to determine the extent to which the factor structure either supported or rejected the presence of the conceptual structure hypothesized from the results of the GCM. The results of the two studies were treated as evidence for the development of a framework for organizational impact evaluation capacity building. Beyond the EFA, there is not an available statistical approach
for comparing between GCM and factor-analytic results. The approach taken here is more conceptual and based on logic and argument.

**Framework Development, Expert Review, and Face Validity**

The third purpose of this research was to produce a framework that could be used by practitioners in their own work. The framework was developed by examining the results of the factor-analysis to determine the extent to which it either confirmed or rejected the proposed factor structure that stemmed from the original concept map. These results were examined in addition to a review of the literature in order to triangulate information that was important to include in the framework. The components of the framework were articulated based upon a literature review, as well as information collected during the research. To further refine the framework before publication an expert panel was convened to offer input.

Given the emphasis on utility it was important to seek expert feedback on the framework as a tool. Once the framework was developed it was disseminated to four key evaluation experts in the field of both international development and evaluation capacity. This included both academic and practice-oriented evaluators. Individuals were sought out based on their contribution to the impact evaluation conversation, or their work on examining evaluation capacity.

Participants in this process were asked to reflect on the utility of the framework given its ultimate purpose of providing an approach to building organizational impact evaluation capacity. When necessary, brief interviews were done with these individuals to further clarify their feedback. This process offered a level of refinement that would
support both the utility, and the face validity of the framework once it is disseminated more broadly.
CHAPTER 4

RESULTS

This chapter presents the results from the primary data collection for this research. We frame the chapter using the original research questions, which were:

4. What components of institutional impact evaluation capacity development do international development practitioners and evaluators believe are most important?

5. In what ways do the constructs measured by the survey instrument relate to concepts that were present in the original concept map?

6. What kind of framework do the results of this research lend themselves to?

The chapter begins by describing characteristics of the respondents. The next section presents a descriptive analysis of the importance ratings from the survey. The data are reported by their original cluster membership from the concept map in the study done by Smith and Kane (manuscript in preparation). This facilitates a discussion of the average importance of clusters from the concept mapping study compared with the data collected for this research.

The next section presents the exploratory factor analysis (EFA) conducted to examine the constructs represented by the items in the survey. This analysis aims to
answer the second research question by discussing the results in the context of the original theoretical model, as represented by the concept map. The chapter closes with an examination of the findings and a discussion of the emergent framework based on both the theoretical model (i.e., the concept map) and the observed model (i.e., the EFA results).

**Respondents**

There were 217 surveys submitted through the Qualtrics survey platform used for this data collection. Of those, 39 were too incomplete to use in any of the analysis. Responses were retained for the descriptive portion of the analysis if the respondent had answered at least half of the questions related to impact evaluation. Of the remaining 178 surveys 148 were complete. The 148 complete responses were used for the exploratory factor analysis (EFA). The respondent characteristics for the responses used in the descriptive analysis (i.e., the larger pool) and the characteristics for the complete responses used in the factor analysis are presented in Tables 3 and 4. The percentages for role, years of experience in evaluation, familiarity with impact evaluation, and education level are nearly identical between the two samples.

There was nearly an even number of respondents who identified as independent consultants (n=63) or internal evaluators (n=54). Those whose responses fell into the internal evaluator category included individuals who identified as monitoring and evaluation (M&E) specialists, directors or deputy directors of evaluation offices, or individuals who specifically described themselves as impact assessment specialists within organizations. Individuals who identified as consultants were mostly focused on
evaluation, although some of these individuals said they were gender specialists or described themselves more broadly as international development consultants. The breakdown of roles for the respondents whose data is included in the descriptive analysis and in the EFA is presented in Table 3 below.

Table 3. Respondent reported professional roles

<table>
<thead>
<tr>
<th>Demographic variable</th>
<th>Survey analysis</th>
<th>Factor analysis</th>
<th>Test of Equality</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$n$</td>
<td>Percent</td>
<td>$n$</td>
</tr>
<tr>
<td>Role</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Consultant</td>
<td>63</td>
<td>40%</td>
<td>53</td>
</tr>
<tr>
<td>Internal evaluator</td>
<td>54</td>
<td>34%</td>
<td>49</td>
</tr>
<tr>
<td>Program manager</td>
<td>17</td>
<td>11%</td>
<td>15</td>
</tr>
<tr>
<td>Program director</td>
<td>10</td>
<td>6%</td>
<td>8</td>
</tr>
<tr>
<td>Academic</td>
<td>8</td>
<td>5%</td>
<td>6</td>
</tr>
<tr>
<td>Other</td>
<td>7</td>
<td>4%</td>
<td>7</td>
</tr>
<tr>
<td>Years of experience in evaluation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1-5 years</td>
<td>42</td>
<td>25%</td>
<td>38</td>
</tr>
<tr>
<td>6-10 years</td>
<td>44</td>
<td>26%</td>
<td>37</td>
</tr>
<tr>
<td>11-15 years</td>
<td>38</td>
<td>22%</td>
<td>31</td>
</tr>
<tr>
<td>16-20 years</td>
<td>23</td>
<td>14%</td>
<td>21</td>
</tr>
<tr>
<td>21+ years</td>
<td>22</td>
<td>13%</td>
<td>17</td>
</tr>
<tr>
<td>Reported familiarity with impact evaluation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Not at all experienced</td>
<td>5</td>
<td>3%</td>
<td>5</td>
</tr>
<tr>
<td>Somewhat experienced</td>
<td>57</td>
<td>34%</td>
<td>51</td>
</tr>
<tr>
<td>Experienced</td>
<td>75</td>
<td>45%</td>
<td>62</td>
</tr>
<tr>
<td>Very experienced</td>
<td>30</td>
<td>18%</td>
<td>26</td>
</tr>
<tr>
<td>Education level</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Doctorate</td>
<td>36</td>
<td>22%</td>
<td>30</td>
</tr>
<tr>
<td>Masters</td>
<td>116</td>
<td>69%</td>
<td>102</td>
</tr>
<tr>
<td>Bachelors</td>
<td>14</td>
<td>8%</td>
<td>12</td>
</tr>
</tbody>
</table>

Just over half of the respondents had between 1 and 10 years of experience in the field of evaluation. Thirteen percent of respondents had over 20 years of experience in
the field of evaluation. Most respondents (63%) said that they were either somewhat experienced or experienced with impact evaluation. As with most research on evaluation studies, the respondents were highly educated. A majority had master’s degrees, and some had doctorates. Very few respondents had less than a master’s degree.

Table 4. Respondent reported country of primary office and representation by continent

<table>
<thead>
<tr>
<th>Variable</th>
<th>Survey Analysis</th>
<th>Factor Analysis</th>
<th>Test of Equality</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n</td>
<td>Percent</td>
<td>n</td>
</tr>
<tr>
<td>Continent</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Africa</td>
<td>43</td>
<td>27%</td>
<td>36</td>
</tr>
<tr>
<td>Asia</td>
<td>17</td>
<td>11%</td>
<td>16</td>
</tr>
<tr>
<td>Australia</td>
<td>5</td>
<td>3%</td>
<td>4</td>
</tr>
<tr>
<td>Europe</td>
<td>52</td>
<td>33%</td>
<td>43</td>
</tr>
<tr>
<td>Middle East</td>
<td>1</td>
<td>1%</td>
<td>1</td>
</tr>
<tr>
<td>North America</td>
<td>33</td>
<td>21%</td>
<td>30</td>
</tr>
<tr>
<td>Country</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Australia</td>
<td>5</td>
<td>3.10%</td>
<td>4</td>
</tr>
<tr>
<td>Austria</td>
<td>2</td>
<td>1.30%</td>
<td>1</td>
</tr>
<tr>
<td>Belgium</td>
<td>3</td>
<td>1.90%</td>
<td>3</td>
</tr>
<tr>
<td>Bolivia</td>
<td>1</td>
<td>0.60%</td>
<td>1</td>
</tr>
<tr>
<td>Brazil</td>
<td>2</td>
<td>1.30%</td>
<td>1</td>
</tr>
<tr>
<td>Canada</td>
<td>7</td>
<td>4.40%</td>
<td>5</td>
</tr>
<tr>
<td>Comoros</td>
<td>1</td>
<td>0.60%</td>
<td>1</td>
</tr>
<tr>
<td>Ethiopia</td>
<td>2</td>
<td>1.30%</td>
<td>2</td>
</tr>
<tr>
<td>France</td>
<td>2</td>
<td>1.30%</td>
<td>1</td>
</tr>
<tr>
<td>Georgia</td>
<td>1</td>
<td>0.60%</td>
<td>1</td>
</tr>
<tr>
<td>Germany</td>
<td>1</td>
<td>0.60%</td>
<td>1</td>
</tr>
<tr>
<td>Ghana</td>
<td>1</td>
<td>0.60%</td>
<td>-</td>
</tr>
<tr>
<td>Guatemala</td>
<td>1</td>
<td>0.60%</td>
<td>1</td>
</tr>
<tr>
<td>India</td>
<td>5</td>
<td>3.10%</td>
<td>5</td>
</tr>
<tr>
<td>Indonesia</td>
<td>1</td>
<td>0.60%</td>
<td>-</td>
</tr>
</tbody>
</table>
Table 4—Continued

<table>
<thead>
<tr>
<th>Variable</th>
<th>Survey Analysis</th>
<th>Factor Analysis</th>
<th>Test of Equality</th>
<th>Variable</th>
<th>Survey Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n</td>
<td>Percent</td>
<td>n</td>
<td>n</td>
<td></td>
</tr>
<tr>
<td>Ireland</td>
<td>2</td>
<td>1.30%</td>
<td>1</td>
<td>0.74%</td>
<td></td>
</tr>
<tr>
<td>Italy</td>
<td>1</td>
<td>0.60%</td>
<td>1</td>
<td>0.74%</td>
<td></td>
</tr>
<tr>
<td>Jordan</td>
<td>1</td>
<td>0.60%</td>
<td>1</td>
<td>0.74%</td>
<td></td>
</tr>
<tr>
<td>Kenya</td>
<td>9</td>
<td>5.70%</td>
<td>7</td>
<td>5.15%</td>
<td></td>
</tr>
<tr>
<td>Liberia</td>
<td>2</td>
<td>1.30%</td>
<td>2</td>
<td>1.47%</td>
<td></td>
</tr>
<tr>
<td>Malawi</td>
<td>1</td>
<td>0.60%</td>
<td>1</td>
<td>0.74%</td>
<td></td>
</tr>
<tr>
<td>Morocco</td>
<td>1</td>
<td>0.60%</td>
<td>1</td>
<td>0.74%</td>
<td></td>
</tr>
<tr>
<td>Multiple</td>
<td>2</td>
<td>1.30%</td>
<td>1</td>
<td>0.74%</td>
<td></td>
</tr>
<tr>
<td>Myanmar</td>
<td>1</td>
<td>0.60%</td>
<td>1</td>
<td>0.74%</td>
<td></td>
</tr>
<tr>
<td>Netherlands</td>
<td>4</td>
<td>2.50%</td>
<td>4</td>
<td>2.94%</td>
<td></td>
</tr>
<tr>
<td>Nicaragua</td>
<td>1</td>
<td>0.60%</td>
<td>1</td>
<td>0.74%</td>
<td></td>
</tr>
<tr>
<td>Nigeria</td>
<td>6</td>
<td>3.80%</td>
<td>5</td>
<td>3.68%</td>
<td></td>
</tr>
<tr>
<td>Norway</td>
<td>1</td>
<td>0.60%</td>
<td>1</td>
<td>0.74%</td>
<td></td>
</tr>
<tr>
<td>Pakistan</td>
<td>3</td>
<td>1.90%</td>
<td>3</td>
<td>2.21%</td>
<td></td>
</tr>
<tr>
<td>Peru</td>
<td>1</td>
<td>0.60%</td>
<td>1</td>
<td>0.74%</td>
<td></td>
</tr>
<tr>
<td>Philippines</td>
<td>1</td>
<td>0.60%</td>
<td>1</td>
<td>0.74%</td>
<td></td>
</tr>
<tr>
<td>Senegal</td>
<td>1</td>
<td>0.60%</td>
<td>1</td>
<td>0.74%</td>
<td></td>
</tr>
<tr>
<td>Sierra Leone</td>
<td>3</td>
<td>1.90%</td>
<td>2</td>
<td>1.47%</td>
<td></td>
</tr>
<tr>
<td>South Africa</td>
<td>6</td>
<td>3.80%</td>
<td>5</td>
<td>3.68%</td>
<td></td>
</tr>
<tr>
<td>South Sudan</td>
<td>1</td>
<td>0.60%</td>
<td>1</td>
<td>0.74%</td>
<td></td>
</tr>
<tr>
<td>Spain</td>
<td>3</td>
<td>1.90%</td>
<td>2</td>
<td>1.47%</td>
<td></td>
</tr>
<tr>
<td>Sri Lanka</td>
<td>1</td>
<td>0.60%</td>
<td>1</td>
<td>0.74%</td>
<td></td>
</tr>
<tr>
<td>Sweden</td>
<td>1</td>
<td>0.60%</td>
<td>1</td>
<td>0.74%</td>
<td></td>
</tr>
<tr>
<td>Switzerland</td>
<td>8</td>
<td>5.00%</td>
<td>8</td>
<td>5.88%</td>
<td></td>
</tr>
<tr>
<td>Tanzania</td>
<td>2</td>
<td>1.30%</td>
<td>2</td>
<td>1.47%</td>
<td></td>
</tr>
<tr>
<td>Thailand</td>
<td>1</td>
<td>0.60%</td>
<td>1</td>
<td>0.74%</td>
<td></td>
</tr>
<tr>
<td>Timor-Leste</td>
<td>1</td>
<td>0.60%</td>
<td>1</td>
<td>0.74%</td>
<td></td>
</tr>
<tr>
<td>Uganda</td>
<td>3</td>
<td>1.90%</td>
<td>3</td>
<td>2.21%</td>
<td></td>
</tr>
<tr>
<td>UK</td>
<td>23</td>
<td>14.50%</td>
<td>18</td>
<td>13.24%</td>
<td></td>
</tr>
<tr>
<td>United States</td>
<td>26</td>
<td>16.40%</td>
<td>25</td>
<td>18.38%</td>
<td></td>
</tr>
<tr>
<td>Vietnam</td>
<td>1</td>
<td>0.60%</td>
<td>1</td>
<td>0.74%</td>
<td></td>
</tr>
<tr>
<td>Zambia</td>
<td>3</td>
<td>1.90%</td>
<td>2</td>
<td>1.47%</td>
<td></td>
</tr>
<tr>
<td>Zimbabwe</td>
<td>1</td>
<td>0.60%</td>
<td>1</td>
<td>0.74%</td>
<td></td>
</tr>
</tbody>
</table>
Respondents were working in a wide array of countries around the world, in six major regions (see Table 4). Most were working in offices located in Europe (33%), with the largest contingent of respondents being from the UK. There was also a large representation of respondents from the African continent (27%) with 16 different countries in the sample. North America was also substantially represented (21%) with most of these coming from individuals in the United States. Survey responses were collected from all over the world. This geographic spread suggests widespread interest in impact evaluation across geographies, and among international development practitioners and evaluators.

**Survey Results**

This section presents descriptive results from the analysis of the importance ratings collected from the survey. The results answer the first research question, that is; what are the most important issues to consider when developing an institutional impact evaluation strategy? The survey collected data on 42 individual ideas related to institutional evaluation capacity. The items were developed based on the concept mapping data in the study done by Smith and Kane (manuscript in preparation), and the representation of items by cluster is shown in Table 2 in Chapter 3. Because of this existing conceptual framework, the results are presented using that structure and the ideas are organized by their original cluster to facilitate discussion about ideas that might be most important to respondents, and the relative importance of clusters for organizations to consider when developing an impact evaluation strategy.
An aggregation of results at the cluster level is presented in Table 5. It is unsurprising that the responses are skewed toward more important, since they emerged from a GCM study, and were made up of items rated high in importance in that study (see Chapter 3 for more detail on item selection). Group concept mapping asks participants to respond to a prompt during brainstorming; the inclusion of an idea already suggests that it has some value to participants. Although the respondents were not involved in the production of the concept map, the ideas from it likely represent items that they perceive as having importance.

Table 5. Mean importance rating at the cluster level

<table>
<thead>
<tr>
<th>Cluster Name</th>
<th>Original Concept Map</th>
<th>Survey</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Point of Origin</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Rank</td>
<td>ŷ</td>
</tr>
<tr>
<td>Foundation and Support</td>
<td>5</td>
<td>3.99</td>
</tr>
<tr>
<td>Integration and Harmonization</td>
<td>4</td>
<td>4.08</td>
</tr>
<tr>
<td>Context and Utilization</td>
<td>6</td>
<td>3.97</td>
</tr>
<tr>
<td>Definition and Communication</td>
<td>7</td>
<td>3.97</td>
</tr>
<tr>
<td>Resources</td>
<td>1</td>
<td>4.27</td>
</tr>
<tr>
<td>Knowledge and Asset Mapping</td>
<td>3</td>
<td>4.13</td>
</tr>
<tr>
<td>Rationale</td>
<td>2</td>
<td>4.14</td>
</tr>
<tr>
<td>Capacity Building</td>
<td>8</td>
<td>3.95</td>
</tr>
</tbody>
</table>
The cluster with the highest mean importance rating was Foundation and Support ($\bar{y}=4.16$). However, that cluster only had two items on the instrument, and so that result should be interpreted with extreme caution. The other clusters that emerge as particularly important are Context and Utilization ($\bar{y}=4.04$), and Integration and Harmonization ($\bar{y}=4.07$). The Capacity Building cluster has the lowest mean importance rating ($\bar{y}=3.53$).

These data provide a conceptual level glance at the ratings of respondents, but the individual item responses provide more insight into what organizations may need to consider when developing an organizational strategy for impact evaluation capacity development. The results at the item level are presented in Table 6.

![Figure 3. Illustration of the top five rated items as process markers for designing and implementing impact evaluations.](image)

Interestingly, the five most important items seem to represent a process diagram for conducting impact evaluation, as represented in Figure 3. These are also highlighted in Table 6 below. It starts at the top with defining the term impact. Defining key terms is
important, and the variation in how organizations and institutions have defined impact is proof of this (Belcher & Palenberg, 2016). Moving clockwise we arrive at the item that had the highest proportion of “extremely important” ratings: considering the context where interventions are being planned when designing an impact evaluation. This issue is well represented in the impact evaluation literature (Leeuw & Vaessen, 2009b), and is a cornerstone of evaluation practice more broadly. It suggests that context is critical for planning and designing impact evaluations. Respondents recognized the importance of considering ethical issues that arise when planning and implementing impact evaluations. The item specifically relates to ethics when considering the use of control groups in an impact evaluation design, but ethics should be considered throughout the design, implementation, analysis, and reporting of impact evaluations. The next item discusses the importance of developing a monitoring system for data collection that should be developed during project planning. Finally, we arrive at the item that describes organizational processes to ensure the learning that emerges from impact evaluation is used in new project design, and thus completes the cycle. These five items encapsulate some of the most important decision points for organizations when considering the planning, implementation, and use of impact evaluation, a conclusion reflected by their high importance rating.
## Table 6. Importance ratings of items

<table>
<thead>
<tr>
<th>Cluster of origin</th>
<th>Statement</th>
<th>n</th>
<th>Not at all</th>
<th>Somewhat important</th>
<th>Important</th>
<th>Very important</th>
<th>Extremely important</th>
</tr>
</thead>
<tbody>
<tr>
<td>Capacity Building</td>
<td>Organizations need staff to have knowledge of impact evaluation to ensure the evaluation questions align with their information needs as opposed to the research needs of academics.</td>
<td>3%</td>
<td>10%</td>
<td>21%</td>
<td>34%</td>
<td>31%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Organizations should develop the capacity of their staff for creating theories of change.</td>
<td>0%</td>
<td>8%</td>
<td>15%</td>
<td>46%</td>
<td>30%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Internal evaluation offices should coordinate impact evaluation work.</td>
<td>6%</td>
<td>16%</td>
<td>28%</td>
<td>28%</td>
<td>22%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Internal evaluation offices of organizations should develop communities of practice around impact evaluation.</td>
<td>4%</td>
<td>17%</td>
<td>36%</td>
<td>30%</td>
<td>14%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Organizations should facilitate an internal community of practice around impact evaluation to connect less experienced and more experienced staff.</td>
<td>5%</td>
<td>13%</td>
<td>37%</td>
<td>32%</td>
<td>13%</td>
<td></td>
</tr>
<tr>
<td>Cluster of origin</td>
<td>Statement</td>
<td>n</td>
<td>Not at all important</td>
<td>Somewhat important</td>
<td>Important</td>
<td>Very important</td>
<td>Extremely important</td>
</tr>
<tr>
<td>-------------------</td>
<td>---------------------------------------------------------------------------</td>
<td>-------</td>
<td>----------------------</td>
<td>--------------------</td>
<td>----------</td>
<td>----------------</td>
<td>---------------------</td>
</tr>
<tr>
<td>Context &amp; Utilization</td>
<td>Organizations should facilitate the development of an internal network to conduct peer review of impact evaluations.</td>
<td>4%</td>
<td>21%</td>
<td>33%</td>
<td>33%</td>
<td>11%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Impact evaluations should take into account the context in which interventions are implemented.</td>
<td>0%</td>
<td>1%</td>
<td>8%</td>
<td>28%</td>
<td>63%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ethical considerations should be taken into account when considering the use of control groups in an impact evaluation.</td>
<td>1%</td>
<td>5%</td>
<td>8%</td>
<td>27%</td>
<td>60%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Theories of change should be used to design impact evaluations.</td>
<td>3%</td>
<td>5%</td>
<td>21%</td>
<td>35%</td>
<td>36%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Determining the impact of an intervention should be considered one criterion when making an overall assessment of a programme’s merit.</td>
<td>3%</td>
<td>8%</td>
<td>21%</td>
<td>39%</td>
<td>29%</td>
<td></td>
</tr>
<tr>
<td>Cluster of origin</td>
<td>Statement</td>
<td>n</td>
<td>Not at all</td>
<td>Somewhat important</td>
<td>Important</td>
<td>Very important</td>
<td>Extremely important</td>
</tr>
<tr>
<td>-------------------</td>
<td>---------------------------------------------------------------------------</td>
<td>----</td>
<td>------------</td>
<td>--------------------</td>
<td>-----------</td>
<td>---------------</td>
<td>--------------------</td>
</tr>
<tr>
<td></td>
<td>Organizations should consider a wide range of impact evaluation approaches.</td>
<td>1%</td>
<td>7%</td>
<td>22%</td>
<td>42%</td>
<td>28%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>The methods used to do impact evaluations should consider the perceptions of credibility users have regarding evidence generation.</td>
<td>2%</td>
<td>9%</td>
<td>30%</td>
<td>41%</td>
<td>18%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Organizations need to be able to explain how they define impact.</td>
<td>0%</td>
<td>2%</td>
<td>9%</td>
<td>30%</td>
<td>60%</td>
<td></td>
</tr>
<tr>
<td>Definition &amp; Communication</td>
<td>Organizations need to have a common understanding of what can be called an impact evaluation.</td>
<td>2%</td>
<td>6%</td>
<td>14%</td>
<td>31%</td>
<td>48%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>A definition of impact evaluation should take into account different levels of impact (e.g., short-term vs. long-term, impact on individuals vs. impact on policy).</td>
<td>1%</td>
<td>7%</td>
<td>16%</td>
<td>36%</td>
<td>41%</td>
<td></td>
</tr>
<tr>
<td>Cluster of origin</td>
<td>Statement</td>
<td>n</td>
<td>Not at all</td>
<td>Somewhat important</td>
<td>Important</td>
<td>Very important</td>
<td>Extremely important</td>
</tr>
<tr>
<td>-------------------</td>
<td>-----------</td>
<td>----</td>
<td>------------</td>
<td>--------------------</td>
<td>-----------</td>
<td>----------------</td>
<td>---------------------</td>
</tr>
<tr>
<td></td>
<td>Organizations need a consistent message on impact evaluation to communicate effectively.</td>
<td>2%</td>
<td>9%</td>
<td>19%</td>
<td>38%</td>
<td>32%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>A definition of impact evaluation should allow for different types of approaches for assessing attribution.</td>
<td>1%</td>
<td>5%</td>
<td>20%</td>
<td>45%</td>
<td>30%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Organizations should clarify how impact evaluations differ from other types of evaluations they may use.</td>
<td>1%</td>
<td>11%</td>
<td>22%</td>
<td>39%</td>
<td>26%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Organizations should develop a clear understanding for external audiences as to what they mean when they say they are conducting impact evaluations.</td>
<td>1%</td>
<td>10%</td>
<td>21%</td>
<td>43%</td>
<td>26%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Decisions to conduct impact evaluations should be guided by a well-defined policy.</td>
<td>2%</td>
<td>8%</td>
<td>30%</td>
<td>37%</td>
<td>23%</td>
<td></td>
</tr>
<tr>
<td>Cluster of origin</td>
<td>Statement</td>
<td>n</td>
<td>Not at all</td>
<td>Somewhat important</td>
<td>Important</td>
<td>Very important</td>
<td>Extremely important</td>
</tr>
<tr>
<td>-------------------</td>
<td>---------------------------------------------------------------------------</td>
<td>-----</td>
<td>------------</td>
<td>--------------------</td>
<td>----------</td>
<td>----------------</td>
<td>---------------------</td>
</tr>
<tr>
<td>Foundation &amp; Support</td>
<td>Organizations should establish processes that ensure the learning generated by impact evaluations feeds into new project design.</td>
<td>0%</td>
<td>1%</td>
<td>5%</td>
<td>33%</td>
<td>61%</td>
<td></td>
</tr>
<tr>
<td>Integration &amp; Harmonization</td>
<td>Organizations should develop a strategy for doing impact evaluations so as to manage expectations of external stakeholders.</td>
<td>1%</td>
<td>11%</td>
<td>22%</td>
<td>41%</td>
<td>24%</td>
<td></td>
</tr>
<tr>
<td>Integration &amp; Harmonization</td>
<td>A monitoring system should be developed during project planning.</td>
<td>0%</td>
<td>1%</td>
<td>3%</td>
<td>34%</td>
<td>61%</td>
<td></td>
</tr>
<tr>
<td>Integration &amp; Harmonization</td>
<td>Organizations must emphasize the collection of baseline data for new projects.</td>
<td>1%</td>
<td>3%</td>
<td>8%</td>
<td>34%</td>
<td>54%</td>
<td></td>
</tr>
<tr>
<td>Integration &amp; Harmonization</td>
<td>The design of impact evaluations should be integrated into project planning.</td>
<td>1%</td>
<td>6%</td>
<td>10%</td>
<td>38%</td>
<td>46%</td>
<td></td>
</tr>
<tr>
<td>Integration &amp; Harmonization</td>
<td>Organizations should incorporate impact evaluations into their larger evaluation structure.</td>
<td>0%</td>
<td>5%</td>
<td>23%</td>
<td>38%</td>
<td>34%</td>
<td></td>
</tr>
</tbody>
</table>
Table 6—Continued

<table>
<thead>
<tr>
<th>Cluster of origin</th>
<th>Statement</th>
<th>n</th>
<th>Not at all</th>
<th>Somewhat important</th>
<th>Important</th>
<th>Very important</th>
<th>Extremely important</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Organizations should plan impact evaluations in such a way so that individual studies feed evidence into larger thematic areas of their work.</td>
<td>1%</td>
<td>4%</td>
<td>21%</td>
<td>47%</td>
<td>27%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Organizations should develop a set of models for impact evaluation that its staff can incorporate into new programmes.</td>
<td>3%</td>
<td>14%</td>
<td>40%</td>
<td>34%</td>
<td>9%</td>
<td></td>
</tr>
<tr>
<td>Knowledge &amp; Asset Mapping</td>
<td>Organizations should be transparent about impact evaluation by providing a directory of impact evaluations they have conducted.</td>
<td>3%</td>
<td>13%</td>
<td>24%</td>
<td>29%</td>
<td>31%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Organizations should examine their capacity to do impact evaluations across all areas of their work.</td>
<td>2%</td>
<td>10%</td>
<td>24%</td>
<td>39%</td>
<td>26%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Organizations should identify and recognize the internal expertise it may already have to do impact evaluations.</td>
<td>2%</td>
<td>5%</td>
<td>29%</td>
<td>42%</td>
<td>22%</td>
<td></td>
</tr>
<tr>
<td>Cluster of origin</td>
<td>Statement</td>
<td>n</td>
<td>Not at all important</td>
<td>Somewhat important</td>
<td>Important</td>
<td>Very important</td>
<td>Extremely important</td>
</tr>
<tr>
<td>-------------------</td>
<td>---------------------------------------------------------------------------</td>
<td>----</td>
<td>----------------------</td>
<td>---------------------</td>
<td>-----------</td>
<td>----------------</td>
<td>--------------------</td>
</tr>
<tr>
<td></td>
<td>Organizations should catalogue interventions and their associated outcomes.</td>
<td>1%</td>
<td>11%</td>
<td>26%</td>
<td>41%</td>
<td>21%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Organizations should map out the various types of impact evaluations that are being done across its body of work.</td>
<td>4%</td>
<td>11%</td>
<td>32%</td>
<td>41%</td>
<td>13%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Organizations should develop an agenda for doing impact evaluations that link to thematic areas of their work.</td>
<td>1%</td>
<td>10%</td>
<td>40%</td>
<td>37%</td>
<td>13%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Organizations should publish guidelines on doing impact evaluation in the context of their work.</td>
<td>7%</td>
<td>14%</td>
<td>38%</td>
<td>32%</td>
<td>10%</td>
<td></td>
</tr>
<tr>
<td>Rationale</td>
<td>Impact evaluations should be done to demonstrate the sustainability of programmes.</td>
<td>3%</td>
<td>7%</td>
<td>19%</td>
<td>38%</td>
<td>34%</td>
<td></td>
</tr>
<tr>
<td>Cluster of origin</td>
<td>Statement</td>
<td>n</td>
<td>Not at all</td>
<td>Somewhat important</td>
<td>Important</td>
<td>Very important</td>
<td>Extremely important</td>
</tr>
<tr>
<td>-------------------</td>
<td>---------------------------------------------------------------------------</td>
<td>-----</td>
<td>------------</td>
<td>--------------------</td>
<td>-----------</td>
<td>----------------</td>
<td>---------------------</td>
</tr>
<tr>
<td></td>
<td>Impact evaluations must be designed with a plan for how it will influence policy making.</td>
<td>2%</td>
<td>9%</td>
<td>22%</td>
<td>39%</td>
<td>28%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Impact evaluations should be done to demonstrate that programmes are replicable.</td>
<td>7%</td>
<td>14%</td>
<td>30%</td>
<td>28%</td>
<td>21%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Impact evaluations should be done to ensure programme managers are held accountable.</td>
<td>7%</td>
<td>20%</td>
<td>29%</td>
<td>28%</td>
<td>16%</td>
<td></td>
</tr>
<tr>
<td>Resources</td>
<td>Funds should be allocated in program/project budgets specifically for impact evaluation.</td>
<td>1%</td>
<td>7%</td>
<td>14%</td>
<td>34%</td>
<td>44%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Organizations should allocate more funds to do impact evaluation.</td>
<td>3%</td>
<td>7%</td>
<td>23%</td>
<td>42%</td>
<td>25%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Organizations should approach funders to secure funds specifically for doing impact evaluations.</td>
<td>7%</td>
<td>9%</td>
<td>25%</td>
<td>39%</td>
<td>20%</td>
<td></td>
</tr>
</tbody>
</table>
Although the comparison between the importance ratings from the original concept mapping exercise and the survey used for this research is imperfect, it is nevertheless interesting since the constructs being examined are based on the same theoretical premise. The first caveat to note when making comparisons at the cluster level is that the survey for this research was based on a reduced set of items from the original concept mapping data. The second limitation is that the number of respondents between the two studies is substantially different. In the original concept mapping study, there were 22 individuals who rated the items on importance, and in this research, there were 178 respondents. Despite these limitations, describing the similarities and differences between the average cluster ratings between the two studies is informative. This is done by discussing the rank of each cluster within its study context, as presented in Table 5.

The only notable similarity between the two data sets is that capacity building is ranked lowest in both. Although the concept of capacity building is prominent in evaluation, in this study it is an operational step, to serve a larger strategic purpose. One possible explanation for its low average ranking is that given the state of impact evaluation in international development, where demand continues to grow, but uncertainty lingers around key epistemological issues, respondents may view other clusters and ideas within them as more central to developing a strategy for incorporating more impact evaluation. Issues of rationale, definitions, and use are more central and foundational then capacity building.

Overall, there was little agreement between the two data sets when rankings of clusters were compared. This could be a result of the changes to the respective surveys
used in each study, or differences in perceptions between the two samples. It is unclear what might be driving the variation at this point.

**Exploratory Factor Analysis**

To answer the second main research question (In what ways do the constructs measured by the survey instrument relate to concepts that were present in the original concept map), EFA was used to analyze the underlying factor structure of the survey data. The analysis was done to determine the extent to which the theoretical model (i.e., the conceptual framework from the group concept mapping study) carried over and could be replicated by an empirical model. That empirical model would be represented by the emergent factor structure from the EFA.

**Test of Sampling Adequacy**

The sample size obtained for this research demanded an analysis to determine its adequacy for EFA. The data were tested using the Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy, and Bartlett’s test of sphericity. The KMO test assesses the extent to which the data contain distinct factors, or just chance correlations between variables (Roger L. Worthington & Whittaker, 2006). The data were tested and the result was a measure of sampling adequacy (MSA) of 0.79. According to Tabachnick and Fidell (2007) data should be used for EFA only if the MSA is above 0.60. Using this as a reference, the data met the criteria and therefore it was determined that the data in the sample was adequate to move forward with EFA.
To provide additional evidence of sampling adequacy Bartlett’s test for sphericity was also conducted. This test is particularly useful when the sample size is smaller, and the respondent-to-item ratio is below five, but greater than three (Worthington & Whittaker, 2006). The respondent-to-item ratio for these data was 3.5, making Bartlett’s test a good option for these data. Bartlett’s test estimates the probability that correlations in the matrix are equal to zero. The results of the Bartlett’s test on the correlation matrix from these data was significant ($\chi^2 = 2543.24$, $p=0.001$), also suggesting that these data are adequate for EFA.

The use of these two diagnostic statistics for examining the adequacy of the sampled data suggested that the data collected were appropriate for EFA. These results do not provide information about any other aspects of the EFA, only that proceeding with the approach is acceptable given the characteristics of the data represented by the sample.

**Factor Extraction and Rotation**

The most commonly used approach to determining the number of factors to extract in EFA is the use of the Kaiser Criterion that states any factors with an eigenvalue above 1 should be retained (Yong & Pearce, 2013). However, there is broad consensus that this approach is among one of the least accurate approaches for factor extraction (Costello & Osborne, 2005; McNeish, 2017; Yong & Pearce, 2013). This is particularly true when smaller sample sizes present a risk that too many factors might be extracted (McNeish, 2017), making it necessary to use a more sophisticated approach for extracting
factors in these data. The first approach involved an examination of the scree plot, which plots the eigenvalues of each of the factors on a scale. The scree plot is shown below in Figure 4.

![Scree plot](image)

**Figure 4.** Scree plot depicting factor extraction and principal component extraction.

A review of the scree plot can be a useful first step for determining the appropriate number of factors for extraction. To make this determination, the factor plots are examined to identify a breakpoint in the data where the line begins to flatten and/or straighten out (Costello & Osborne, 2005). In this case, focusing only on the FA factors (i.e., the white dots) there is a slight break after the 6th factor.

To confirm that a six-factor model was appropriate, a parallel analysis was conducted. Although not widely used since most statistical programs do not include it in
their factor-analytic packages, it has been shown to work reasonably well (McNeish, 2017). Parallel analysis is an approach that tries to address the issue of overestimation in matrix rank as a result of sampling error. It does this by constructing correlation matrices of random variables based on the sample size of the real data set. The eigenvalues from the real data set and the randomly generated data sets are compared and those factors in the real dataset with eigenvalues greater than those from the random dataset are retained (Hayton, Allen, & Scarpello, 2004; Worthington & Whittaker, 2006). The scree plot from the parallel analysis is shown in Figure 5 below. This scree plot includes both the actual data and the simulated data. Once again, a visual examination of the factor analysis’ actual data (blue triangles) suggests that there is a slight break after the 6th factor after which the line continues uninterrupted. It looks similar to the scree plot presented earlier, but reliance on the scree plot alone can pose issues for determining the most appropriate number of factors to extract (Hayton et al., 2004). The parallel analysis output also provides a suggested number of factors for extraction from the data. This analysis resulted in the recommendation for six factors to be extracted, corresponding with the previous conclusion based on the break in the scree plot.

The six factors were extracted using Promax rotation. This resulted in a model fit of 6.43 ($p = 0.001$). The six-factor structure explained 40% of the variance. The measures of factor score adequacy also suggest that six factors are in fact an appropriate number for extraction. These statistics are shown in Table 7 and include the correlation of scores
with the six factors, the multiple $R$ squares of scores with the six factors, and the minimum correlation of possible factor scores. It is most important for the first two statistics to be as close to 1 as possible (Grice, 2001), and in this case those correlations and $R$-squared values are sufficiently high to move forward with the six-factor solution.

Figure 5. Scree plot from the parallel analysis.

Table 7. Measures of factor score adequacy for the six-factor solution

<table>
<thead>
<tr>
<th>Measure of Factor Score Adequacy</th>
<th>Factor 1</th>
<th>Factor 2</th>
<th>Factor 3</th>
<th>Factor 4</th>
<th>Factor 5</th>
<th>Factor 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correlation of (regression) scores with factors</td>
<td>0.95</td>
<td>0.9</td>
<td>0.91</td>
<td>0.9</td>
<td>0.92</td>
<td>0.87</td>
</tr>
<tr>
<td>Multiple $R$-square of scores with factors</td>
<td>0.9</td>
<td>0.82</td>
<td>0.82</td>
<td>0.81</td>
<td>0.84</td>
<td>0.75</td>
</tr>
<tr>
<td>Minimum correlation of possible factor scores</td>
<td>0.79</td>
<td>0.63</td>
<td>0.64</td>
<td>0.63</td>
<td>0.68</td>
<td>0.5</td>
</tr>
</tbody>
</table>
The Factor Structure

The six-factor structure was extracted using the Promax oblique rotation. An oblique rotation is recommended over orthogonal rotation when the factors are correlated, and some scholars argue that oblique rotation methods should be the default approach in many cases, because real world data will almost always have correlated factors (Worthington & Whittaker, 2006). Oblique rotation can also be useful when there is substantial cross loading of items across factors. There was notable cross loading when these data were analyzed using an orthogonal rotation, again supporting the case for use of an oblique rotation.

To identify which items to retain in the data, a loading threshold of 0.4 was used. Some literature suggests that threshold be 0.32, but a more conservative number was chosen here to reduce the number of items that cross-loaded onto more than one factor. It was also done to try and ensure the items that were retained had higher communalities. The loading diagram is presented in Figure 6 (the item codes shown in the figure can be identified in Appendix A). The factor extraction and item elimination process resulted in the elimination of 13 items. This left 29 that had loadings on factors that were greater than or equal to 0.4. Table 8 presents the full list of 29 items that were retained along with their loadings on the corresponding factor.
Figure 6. 6-factor extracted model.
### Table 8. Items and factor loadings for six-factor structure

<table>
<thead>
<tr>
<th>Original Cluster</th>
<th>Item</th>
<th>Factor 1</th>
<th>Factor 2</th>
<th>Factor 3</th>
<th>Factor 4</th>
<th>Factor 5</th>
<th>Factor 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Knowledge &amp; Asset Mapping</td>
<td>Organizations should publish guidelines on doing impact evaluation in the context of their work.</td>
<td>0.42</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Knowledge &amp; Asset Mapping</td>
<td>Organizations should develop an agenda for doing impact evaluations that link to thematic areas of their work.</td>
<td>0.49</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Capacity Building</td>
<td>Organizations should develop the capacity of their staff for creating theories of change.</td>
<td>0.64</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Capacity Building</td>
<td>Organizations should facilitate the development of an internal network to conduct peer review of impact evaluations.</td>
<td>0.76</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Capacity Building</td>
<td>Organizations should facilitate an internal community of practice around impact evaluation to connect less experienced and more experienced staff.</td>
<td>0.82</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Capacity Building</td>
<td>Internal evaluation offices of organizations should develop communities of practice around impact evaluation.</td>
<td>0.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Definition &amp; Communication</td>
<td>Organizations need to have a common understanding of what can be called an impact evaluation.</td>
<td>0.61</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Definition &amp; Communication</td>
<td>Organizations need a consistent message on impact evaluation to communicate effectively.</td>
<td>0.77</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Definition &amp; Communication</td>
<td>Organizations need to be able to explain how they define impact.</td>
<td>0.42</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Definition &amp; Communication</td>
<td>Organizations should clarify how impact evaluations differ from other types of evaluations they may use.</td>
<td>0.57</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Resources</td>
<td>Funds should be allocated in program/project budgets specifically for impact evaluation.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.66</td>
</tr>
<tr>
<td>Original Cluster</td>
<td>Item</td>
<td>Factor 1</td>
<td>Factor 2</td>
<td>Factor 3</td>
<td>Factor 4</td>
<td>Factor 5</td>
<td>Factor 6</td>
</tr>
<tr>
<td>----------------------------------</td>
<td>----------------------------------------------------------------------</td>
<td>----------</td>
<td>----------</td>
<td>----------</td>
<td>----------</td>
<td>----------</td>
<td>----------</td>
</tr>
<tr>
<td>Foundation &amp; Support</td>
<td>Organizations should establish processes that ensure the learning generated by impact evaluations feeds into new project design.</td>
<td>0.49</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Context &amp; Utilization</td>
<td>Impact evaluations should take into account the context in which interventions are implemented.</td>
<td></td>
<td></td>
<td>0.44</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Context &amp; Utilization</td>
<td>Ethical considerations should be taken into account when considering the use of control groups in an impact evaluation.</td>
<td></td>
<td></td>
<td></td>
<td>0.48</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Integration &amp; Harmonization</td>
<td>The design of impact evaluations should be integrated into project planning.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.48</td>
<td></td>
</tr>
<tr>
<td>Integration &amp; Harmonization</td>
<td>Organizations must emphasize the collection of baseline data for new projects.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.5</td>
</tr>
<tr>
<td>Integration &amp; Harmonization</td>
<td>A monitoring system should be developed during project planning.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.46</td>
</tr>
<tr>
<td>Definition &amp; Communication</td>
<td>A definition of impact evaluation should allow for different types of approaches for assessing attribution.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.45</td>
</tr>
<tr>
<td>Context &amp; Utilization</td>
<td>The methods used to do impact evaluations should consider the perceptions of credibility users have regarding evidence generation.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.46</td>
</tr>
<tr>
<td>Context &amp; Utilization</td>
<td>Organizations should consider a wide range of impact evaluation approaches.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.41</td>
</tr>
<tr>
<td>Context &amp; Utilization</td>
<td>Determining the impact of an intervention should be considered one criterion when making an overall assessment of a programmes merit.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.54</td>
</tr>
<tr>
<td>Knowledge &amp; Asset Mapping</td>
<td>Organizations should identify and recognize the internal expertise it may already have to do impact evaluations.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.48</td>
</tr>
<tr>
<td>Original Cluster</td>
<td>Item</td>
<td>Factor 1</td>
<td>Factor 2</td>
<td>Factor 3</td>
<td>Factor 4</td>
<td>Factor 5</td>
<td>Factor 6</td>
</tr>
<tr>
<td>-----------------</td>
<td>-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
<td>----------</td>
<td>----------</td>
<td>----------</td>
<td>----------</td>
<td>----------</td>
<td>----------</td>
</tr>
<tr>
<td>Knowledge &amp;</td>
<td>Organizations should be transparent about impact evaluation by providing a directory of impact evaluations they have conducted.</td>
<td>0.56</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Asset Mapping</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Knowledge &amp;</td>
<td>Organizations should catalogue interventions and their associated outcomes.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.82</td>
</tr>
<tr>
<td>Asset Mapping</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Knowledge &amp;</td>
<td>Organizations should examine their capacity to do impact evaluations across all areas of their work.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.43</td>
<td>0.43</td>
</tr>
<tr>
<td>Asset Mapping</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rationale</td>
<td>Impact evaluations must be designed with a plan for how it will influence policy making.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.52</td>
</tr>
<tr>
<td>Rationale</td>
<td>Impact evaluations should be done to demonstrate the sustainability of programmes.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.6</td>
</tr>
<tr>
<td>Rationale</td>
<td>Impact evaluations should be done to demonstrate that programmes are replicable.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.58</td>
</tr>
</tbody>
</table>
The EFA results alone are insufficient to answer the second research question of this dissertation. The results do provide additional evidence for the original theoretical model as the model is manifested in the results of the EFA. However, to more fully address that second question, an analytic process comparing these results with the original concept map is critical.

The original concept map represented a theoretical model that served as the basis for this research and consisted of eight clusters of ideas, split into two regions (Smith & Kane, manuscript in development). The purpose of the factor analysis was to determine the extent to which that theoretical model bore out as an observable model in the data collected here. The loadings of each of the retained items has already been presented in Table 8. Table 9 below presents the names of each factor along with a brief description of the factor.

Table 9. Titles and descriptions of the six factors

<table>
<thead>
<tr>
<th>Factor title</th>
<th># of items</th>
<th>Average Importance</th>
<th>Factor description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Institutional structures and capacity</td>
<td>6</td>
<td>3.46</td>
<td>Clusters represented:</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Knowledge and Asset Mapping</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Capacity Building</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Internal institutional structures to develop the knowledge infrastructure related to impact evaluation.</td>
</tr>
<tr>
<td>Factor title</td>
<td># of items</td>
<td>Average Importance</td>
<td>Factor description</td>
</tr>
<tr>
<td>--------------------------------------------------</td>
<td>------------</td>
<td>--------------------</td>
<td>---------------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Defining and delineating impact evaluation</td>
<td>4</td>
<td>4.09</td>
<td>Clusters represented: • Definition and Communication</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Definition of key terms and related concepts to communicate clearly with internal and external audiences.</td>
</tr>
<tr>
<td>Planning and designing impact evaluations</td>
<td>7</td>
<td>4.38</td>
<td>Clusters represented: • Context and Utilization • Integration and Harmonization • Foundation and Support</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Planning, implementing, and designing impact evaluations, including important issues related to context, ethics, and data collection processes.</td>
</tr>
<tr>
<td>Impact evaluation methods and uses</td>
<td>5</td>
<td>3.77</td>
<td>Clusters represented: • Context and Utilization • Definition and Communication • Knowledge and Asset Mapping</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Appropriate approaches and uses for impact evaluation.</td>
</tr>
</tbody>
</table>
### Comparative Analysis of the Theoretical and Observed Models

The results of the EFA build upon the theoretical model from the group concept mapping. The transference of the theoretical model into the data collection instrument ensured that the original constructs were embedded. Therefore, the EFA was done to determine the extent to which those constructs held up when data was collected from a larger and more diverse set of respondents, outside the original context. A factor structure that resembled the concept map’s theoretical model suggest that the original framework potentially has validity.

The EFA also offers the opportunity to refine the theoretical model. The observable model reinforces parts of the original concept map, but also reframes parts of it. This section discusses the six factors and their relationship to the original map. Some of the factors reflect the original model, while other factors seem to reorganize elements
of different clusters into new constructs. The original group concept map with the retained items symbolized by their respective factors is shown in Figure 6. Examining the figure reveals that some factors are composed of items that come from the same original cluster (e.g., Rationale). Other factors draw upon items from various clusters around the map (e.g., Impact Evaluation Methods and Approaches), pulling them away from their original clusters, and into factors with more thematically similar items. This helps to illustrate how the items from some clusters became distinct factors suggesting a cohesive original construct, while other factors might be grounded in particular clusters but drew other relevant items from other spaces on the map to refine a construct.

The first factor, Institutional Knowledge and Capacity, draws upon items originally part of the Knowledge and Asset Mapping cluster, and the Capacity Building cluster. A qualitative examination of the items in the factor suggests thematic cohesion. The factor is composed of items relating to the development of internal communities of practice around impact evaluation to build staff capacity, creating a peer review network to
enhance learning and quality, and the development of guidance material for impact evaluation to both place such studies in the context of the institution’s work, and lay out how it ought to be linked.

Factor 2: Defining and Delineating Impact Evaluation is composed of four items that originate in the Definitions and Communication cluster on the concept map, supporting the original theory that the cluster composed a unique construct. This would suggest that it is important for organizations to consider definitional aspects of impact evaluation.
evaluation to ensure clear communication, and a common understanding among its internal and external stakeholders or audiences. The finding reaffirms that the theoretical model had merit in identifying this aspect of institutional impact evaluation capacity, and unsurprising since the theoretical model was empirically driven via the data collected in the group concept mapping.

The third factor, Planning and Designing Impact Evaluations, is the first real example of the factor analytic process reframing a construct from the original model. The factor is composed of items that originate from multiple clusters on the map (see Figure 6). The reformulation of the items into one factor makes qualitative sense when the items are reviewed, coalescing around the planning and design of impact evaluations. The factor incorporates items that represent issues related to funding for impact evaluation, development of monitoring systems and the importance of establishing baselines measurements, the integration of impact evaluation into the project planning, and ethical considerations when doing these types of studies. All of these are institutional issues that should be considered when impact evaluations are being planned, and before impact evaluation is incorporated into an organization’s evaluation infrastructure. Although it seems that the factor could be grounded in the Context and Utilization cluster or the Integration and Utilization cluster, (see Figure 6) the analysis has allowed for items related to the planning and design of impact evaluations that were scattered across the original map to be drawn together into a construct that seems a better fit.

Five items make up the fourth factor in the model. Three of these originated in the Context and Utilization cluster from the concept map. This suggests that original
cluster anchors this factor. Three items--two from the Context and Utilization cluster and one from the Definition and Communication cluster--relate to determining appropriate methods and approaches for conducting impact evaluation. Those items inform the name of the factor. The items do not make an argument for one approach over another, but rather suggest considering a wide array of methods and approaches. It is important to take into account the types of evidence users of the evaluation view as credible when making methodological decisions since perceptions of credibility can influence their propensity to use that information.

The items that make up factors five and six originate from the same clusters. The three items that make up factor five originated in the Knowledge and Asset Mapping cluster. This fifth factor is titled Knowledge Management and Learning. The content describes the ways that organizations manage the information that impact evaluations produce. That should include learning around designing and implementing impact evaluations, as well as the outcomes and results. Effectively managing this knowledge management has the potential to enhance the utility of impact evaluation processes and results, making information for program designers and future evaluators more accessible, and improving transparency for internal and external audiences alike.

The sixth factor is composed of three items that all originate in the Rationale cluster and are clearly related to that issue. For this reason, the sixth factor has retained the original cluster name from the map. The three items that have been retained outline three commonly distinguished rationales given for doing impact evaluation. These include
doing impact evaluation to influence policy, to examine the sustainability of programs, and to determine whether programs are replicable beyond their original context.

**A Revised Framework for Impact Evaluation Capacity Development**

The premise of this research was that the theoretical model from the concept mapping had logical cohesion as an impact evaluation capacity framework. A review of the literature supported that claim, showing substantial overlap between the concept maps components and those of other published evaluation capacity frameworks. To more fully make this case, additional empirical evidence was required. That evidence has been presented in this chapter and is now used to describe a revised framework for impact evaluation capacity.

The model that has emerged from the EFA is a modified reflection of the theoretical model from the concept mapping study. Three of the six factors are composed of items that originate from the same cluster on the concept map. Answering the third research question for this study (What kind of framework do the results of this research lend themselves to?) requires tying together the results presented thus far into a cohesive framework. Although the importance ratings that are described at the start of this chapter are descriptively interesting, they do not alone validate an underlying model. They have provided the vehicle by which we have been able to refine the theory that led to this research.

The revised framework is presented in Figure 6 below. The framework is oriented as a sequential process that moves towards conducting impact evaluations. The
framework offers a blueprint for establishing an infrastructure to effectively support
impact evaluation work within institutions. The six components can be viewed as areas of
work that need to be addressed to develop an effective strategy for doing impact
evaluations. Those six areas require information that will be unique to particular
institutional contexts. Data collection to incorporate stakeholder perspectives might be
required to effectively operationalize particular components, or users may need to
conduct reviews of existing literature. Such operationalization is not described in detail
here, and perhaps offers an avenue for future research. It is also certainly a task the user
will have to consider to create a fully realized strategy in a particular context.

The framework begins with the establishment of a clear rationale for why impact
evaluations need to be done. That rationale should incorporate input from multiple
stakeholders within an organization to ensure multiple viewpoints are represented.
Stakeholder identification and selection should be carefully considered but should include
those responsible for conducting or managing impact evaluations, and those who will use
information generated by impact evaluation. The rationale will affect decision making
around many other components of the framework as represented by the directional
arrows linking it to the three components in the middle of the figure.

Although rationale feeds into all three of the framework components in the
middle region of the diagram in Figure 6, the center components are sequenced starting
with the top component, Defining and Delineating Impact Evaluation. The first key aspect
of this component is determining how impact evaluations differ from other evaluations
being done in the organization, and how they fit into the existing evaluation
infrastructure. Upon review, one might determine that existing evaluation work meets the needs outlined by the rationale for doing impact evaluation; or perhaps the existing work can be adapted to include aspects that inform the ability to study impact.
Figure 8. Revised institutional impact evaluation capacity framework.

Rationale
- Why, and for what reasons should impact evaluation be done?
- Will impact evaluation influence policy?
- Will impact evaluation be done to examine sustainability?
- Will impact evaluation be done to examine replicability?

Defining and delineating impact evaluation
- Clarify how impact evaluation is different from other types of evaluation
- Develop a shared definition for impact evaluation
- Develop a shared definition for “impact”
- Disseminate definitions to ensure consistent communication

Planning and designing impact evaluations
- Identify and consider ethical issues
- Analyze relevant contexts to ensure impact evaluations are responsive to information needs
- Establish processes to:
  - Ensure learning is fed into program design
  - Monitoring systems are available and established
  - Baseline data can be collected
  - Integrate design process for impact evaluation into project planning
  - Determine appropriate resourcing strategies for impact evaluations

Impact evaluation methods and approaches
- Examine the perception evaluation users have around the credibility of different methods
- Explore different types of approaches for determining attribution
- Evaluate the feasibility of particular methods for different programmatic contexts
- Explore how the assessment of impact informs a determination of a programs merit

Institutional structures and capacity
- Develop staff capacity around theories of change, and other core concepts relevant for designing and implementing impact evaluations
  - Publish guidance on doing impact evaluation in the organizational context
  - Establish connections between impact evaluation and the thematic areas of institutional work
  - Facilitate peer-to-peer learning through developing internal communities of practice

Knowledge management and learning
- Create a repository of impact evaluations to make learning and knowledge accessible

Conduct impact evaluations
Definitional issues are another important part of the Defining and Delineating Impact Evaluation component. The literature review in Chapter 2 presents a discussion of the variation in definitions of impact by different organizations and institutions. New definitions may not be required, but a review of published literature and guidance material should be done to determine whether existing definitions fit, or if aspects of those definitions might be combined into one that has contextual relevance. Establishing agreed upon definitions of key terms is important for ensuring that there is consistent communication with both internal and external audiences around the topic of impact evaluation. This is particularly true if impact evaluation is relatively new to an institution, or if impact evaluations were being done across organizational entities but not under the guidance of a common strategy or policy.

The next component is the planning and design of impact evaluation. Here key design issues need to be identified, considered and addressed. Ethical guidelines must be reviewed, or established, so that they are in place to guide impact evaluation work. Organizations should also consider how they want contextual analyses to be done, and how that information will be fed into the design of impact evaluations. Given the relative agreement that impact evaluation must be done in a way that is responsive to a particular context, establishing a system to facilitate context assessment can ensure that appropriate responsiveness is incorporated into impact evaluation work. Organizations should also determine how funding will be allocated or secured to do impact evaluations. If impact evaluations are to be added to the existing evaluation infrastructure, the mechanisms by which they will be funded must be established early on.
The other aspect of this component relates to processes that should be considered when designing and planning the actual studies, including developing monitoring systems, integrating the planning of impact evaluation into the project design phase, and creating standards and approaches for establishing baseline measures. These are important operational components that are part of integrating impact evaluation effectively into an organization's structure.

The next component describes defining the types of approaches and methods that will be used to do impact evaluation. This will be influenced by the planning and design components, by the definitions established earlier on, and by the rationales. This framework does not suggest particular methodologies for impact evaluation. Users will make such decisions so that they are institutionally relevant, and the decisions are viewed as credible. Users should seek multiple viewpoints, and gather information to incorporate the input of various stakeholder groups. This will provide insight into how particular stakeholders view various methods and approaches for doing impact evaluation, including how feasible and relevant they might be within a given programmatic context. Some parts of an organization's work might lend themselves to experimental, while others may benefit from alternative approaches to assessing impact.

Next in the model, impact evaluations are conducted. The Knowledge Management and Learning Component has been placed after studies are complete. This component describes a need for an effective knowledge management strategy that will allow learning from previous impact evaluation to be used in future work. Knowledge management is often described in programmatic terms, that the results of impact
evaluation should be used in new project design, linking this concept to the Planning and Designing component. In addition, it reflects the need to learn and disseminate learning to improve impact evaluation practice for present and future evaluators, and may include information on methodology, stakeholder engagement, ways to facilitate use, or prior results and baseline measures.

The Institutional Structures and Capacity component undergirds the entire framework. It is placed along the bottom of the figure to highlight its relevance to the five other components in the model and the model at large. The issues described within this component should be addressed systematically, on an ongoing basis. Organizations should establish official guidance material around impact evaluation that can be referenced across an organization. That material may be premised upon the other parts of this model (e.g., rationale, definition, methods). Developing staff capacity to do, or effectively manage impact evaluations is also important, and should be ongoing. Some staff may need in-depth training, while others may need only an orientation, but there are likely core concepts that all staff should be familiar with. Finally, an internal community of practice may facilitate peer-to-peer learning. This is an effective way for providing ongoing capacity development at appropriate levels. Linking experts with those still learning is a way to provide capacity development in an efficient and effective way. Organizations might also consider establishing a web-based space for information and resource sharing, as well as networking. There are many strategies for facilitating peer-to-peer learning that should be explored fully.
The revised framework presents a sequential approach to generating an impact evaluation capacity strategy. Reflecting on the process model that depicts the five most important items from the survey results shown in Figure 3, all five are included in the framework. The components make up a distinct construct as determined by the development of a theoretical model through group concept mapping, and then refined and reframed through the EFA presented in this chapter. There is certainly more work to be done to unpack and articulate in more depth what each of these model components represents. That is for another study, and beyond the scope of this research. The final chapter will reflect on this framework in the context of the existing literature, and the field. It will also discuss potential users of this framework, and how it might be incorporated as recommendations for practice.
CHAPTER 5
SUMMARY, CONCLUSIONS, AND RECOMMENDATIONS

This research has aimed to build upon work done by Smith and Kane (manuscript in preparation) to build a conceptual model for impact evaluation capacity at the International Labour Organization (ILO). That model was used as the theoretical model for this research. The primary aim of this work was to determine the validity of the model with the intent of refining it to produce a framework that would have utility for the field of international aid. The methodology used the original GCM data to develop a survey. The data collected from that survey was analyzed descriptively and inferentially, using exploratory factor analysis (EFA). The former analysis provided insight into the components that respondents believed were most important for institutions to consider when developing an impact evaluation strategy. The latter analysis was used to determine the extent to which the theoretical constructs in the original concept map were retained, reformed, or eliminated among a larger sample.

The debates around impact evaluation of international aid work have been ongoing. Evaluators have argued over the definition of impact, what approaches should be used, what methods are most appropriate, and how to produce credible evidence of impact. As this debate has played out, the demand for impact evaluation has continued to grow. Growing demand has necessitated a response from the evaluation community,
and that response has been dominated by perspectives that advocate counterfactual
designs and quantitative methods. As described in Chapter 2 the debate has been
remarkably similar to the paradigm wars that played out in the evaluation community up
through the early 2000s.

This work was not done as an effort to settle that debate. The purpose was to test
a theory that posited the existence of a framework for impact evaluation capacity. The
framework has now been constructed and presented in Chapter 4. The remainder of this
chapter is used to interpret those results in the context of existing literature, and to
discuss their implications. I also describe the studies limitations, and areas for future
research.

**Interpretation of the Results**

The results presented in Chapter 4 provide evidence for a refined impact
evaluation capacity framework built upon a theoretical model. The process of refining
that theory included the consolidation of some clusters from the original map, and the
reconstitution of others into new factors representing new constructs. When combined
they represent a sequential approach for examining and developing impact evaluation
capacity within an institutional setting.

Impact evaluations are conducted to generate knowledge of what works. To be
useful they must look beyond that singular question and offer insight into what works for
whom, under what circumstances, and in what settings. Answering these questions
improves the utility of the evaluation, making the findings more meaningful for
information users. It is my view that a restrictive policy for determining appropriate methods or designs is a disservice to conducting the best possible evaluations. That is why one part of this component urges exploration of different types of approaches for determining attribution. In her reflection on experimental designs and their use to study interventions, Cartwright (2011) writes, “For policy and practice we do not need to know ‘it works somewhere’. We need evidence for ‘it-will-work-for-us’ claims: the treatment will produce the desired outcome in our situation as implemented here” (p. 1401). To produce that kind of knowledge, organizations must take into account their full portfolio of work, using it to decide what types of impact evaluations can and should be done.

This impact evaluation capacity framework does not lay out the operational details for each of the components or sub-components. Instead, in current form it offers broader guidance by outlining core areas that constitute the capacity and/or readiness of the institution to incorporate impact evaluation in a thoughtful and purposeful way. It provides a roadmap to guide institutions towards core questions about how impact evaluation best fits their existing structures, what structures may need to be developed, what their values are and how they may affect the use of these studies. Organizations should understand and take into account their rationales for doing impact evaluations, the definitions they adopt for key terms, their approach to design, how they develop staff capacity and appropriate structural capacity to incorporate impact evaluations into their work, and how they learn and manage the results.

Considering the results in the context of the broader evaluation capacity literature it is useful to reflect on the simple definition given by Cheng and King (2016) of evaluation
capacity as the ability to both conduct and use evaluations. Evaluation capacity building is the process that develops that ability to conduct and use evaluations. It is defined by Stockdill, Baizerman, and Compton (2002) as a “context-dependent, intentional action system of guided processes and practices for bringing about and sustaining a state of affairs in which quality program evaluation and its appropriate uses are ordinary and ongoing practices...” (p. 8). The model from this research when fully operational would represent that ability to conduct and use impact evaluations (Cheng & King, 2016) by employing a “context-dependent, and intentional action system”.

There are two important assumptions that need to be considered as part of this model. First, this framework assumes that institutions have some existing evaluation capacity before undertaking work to incorporate impact evaluation through application of this model. This may be an evaluation office or unit that is responsible for other forms of evaluation as well within the organization. Second, the model assumes that there is an existing evaluation infrastructure already developed that supports that organizations evaluative work. Both assumptions rest on an even broader assumption that impact evaluation is a more advanced evaluative form that ought to be built upon more basic evaluative functions.

The emerging model reflects some aspects of the evaluation capacity models described in Chapter 2. In a review of the evaluation capacity literature, Cheng and King (2016) describe four overarching themes that can be used to organize key aspects of existing models. These four areas were evaluation structures, organizational resources, human resources, and evaluation culture; the key components from the various
frameworks for evaluation capacity that have been published in the past 15 years. Using this as an interpretive guide for this impact evaluation capacity framework we can assess the extent to which it reflects aspects of evaluation capacity that have been identified previously.

Evaluation structures includes issues of planning, the organizational environment, and existing processes (Bourgeois & Cousins, 2013; GAO, 2003; Nielsen et al., 2011), representing aspects of the institution that are in place to support evaluation. The impact evaluation capacity framework shown here represents this in the rationale component of the impact evaluation capacity framework (e.g., for what reason does the institution need impact evaluation); but rationale may not represent an explicit structural component. The theme of evaluation structures is most connected with the Institutional Structures and Capacity component of the impact evaluation capacity framework. This includes issues related to developing staff capacity, publishing guidance materials, establishing connections between impact evaluation and thematic frameworks guiding the organizations work, and facilitating peer-to-peer learning opportunities. The issue of structural components that support evaluation have become key aspects of evaluation capacity frameworks over the past 15 years as they have begun to look beyond the capacity of individuals, and towards environmental factors that affect the doing and using of evaluation (Labin et al., 2012).

Components of the impact evaluation capacity framework also represent what Cheng and King (2016) call Organizational Resources. This includes technical resources available to support data collection, a delineation of methods and models, and financial
The present model reflects this in Planning and Designing Impact Evaluations, and Impact Evaluation Methods and Approaches. The former includes issues related to establishing baselines, monitoring systems, processes to feed results into program planning, and identifying appropriate resourcing strategies to finance impact evaluations. The latter component discusses key issues related to determining the appropriate methods or approaches for conducting impact evaluation.

The third theme from Cheng and King’s (2016) review is human resources. Whether there are staff with the technical knowledge, and the capacity to do or manage evaluation is critical. In the impact evaluation capacity framework this is best represented by the Institutional Structures and Capacity component. It connects most explicitly to the first and fourth sub-components of that part of the framework that deal with staff capacity development, and the creation of a peer-to-peer learning network through the developing communities of practice. Perhaps one way this framework differs from past evaluation capacity frameworks is that it does not deal in depth with aspects of individual staff capacities. But it is clear that staff knowledge is fundamental, so some existing evaluation capacity is required to implement this model. Other models for evaluation capacity may be more appropriate for organizations that are just beginning to develop evaluation.

Evaluation culture is the fourth theme described by Cheng and King (2016). It is also the least clearly represented in the impact evaluation capacity framework. Elements of this theme are represented in the Planning and Designing Impact Evaluations component (e.g., establishing processes to ensure learning is fed into program planning,
and integrating design processes for impact evaluation in program planning). The Knowledge Management and Learning component perhaps best exemplifies elements of evaluation culture. This component describes a need to feed learning back into the rest of the model. The model deals most explicitly with learning related to designing and implementing impact evaluations, but an organization that refines its evaluation practice based on previous results and experience does represent a key aspect of an evaluation culture. It is a demonstration of evaluation literacy that Bourgeois and Cousins (2013) describe as part of an evaluation culture.

The most notable divergence between the impact evaluation capacity framework and prior models for evaluation capacity is the inclusion of a component directly related to defining and delineating impact evaluation. This is a result of two conditions. First, as described previously there has been substantial debate related to defining impact evaluation, as well as the term impact. Clarity on these key terms is crucial for communicating with internal and external stakeholders. The second condition is related to the assumptions described above, that impact evaluation is portrayed as a specific type of evaluation that organizations might employ but that should not necessarily precede the development of more general evaluative capacity. In institutions where evaluation is already being practiced, impact evaluation should be delineated from existing evaluative work. Defining impact evaluation will help to highlight its value, and should support the established rationale for doing these types of studies.

Bourgeois and Cousins (2013) describe their model as a foundation for a theory of change. The framework presents two broad issues, the capacity to do evaluation, and the
capacity to use evaluation. Each includes a series of subcomponents, but collectively they could be used to formulate a theory of change to develop evaluation capacity. The present impact evaluation capacity framework might also be interpreted as the foundation for a theory of change. In addition to its theory of change-like appearance, it presents core components that also represent the capacity to do, and use impact evaluations.

**Re-examining the Contribution of this Research**

The original argument made in Chapter 1 about the contribution of this work revolved around a component of Mark’s (2008) framework that he called *evaluation context*. Mark discusses the *evaluation context* as the set of circumstances in which evaluation takes place. This research was primarily oriented towards the organizational level. The model lays out a set of components that can be used to understand or develop a comprehensive and reflective environment in which impact evaluation can take place.

The results also lend themselves to other aspects of Mark’s model. The impact evaluation capacity framework presented in Chapter 4 is a descriptor of *evaluation context*. It presents a set of circumstances that describe the environment for impact evaluations to occur, and offers a descriptive approach to setting boundaries around that context. In addition, the structure includes issues related to *evaluation activities*, *evaluation consequences*, and *professional issues*.

Mark presents a macro level framework encompassing Evaluation. Yet, it is reasonable that sub-models might be relevant to guide research on particular types or
forms of evaluation. This has been reinforced by a call for more research on impact evaluation from Rogers and Peersman (2014) who write, "although impact evaluation is intended to ensure that decisions about practice and policy are informed by evidence, the various arguments about impact evaluation have rarely been based on systematic research" (p. 85). This work is an attempt to contribute by presenting a way forward for conducting more of that systematic research.

The framework is described as an impact evaluation capacity framework, but it also offers areas for future research. As Mark’s (2008) model describes a framework for research on evaluation, this model can be viewed as a way to do research on impact evaluation in line with the calls by Rogers and Peersman (2014).

**Implications**

In Chapter 1 I describe the problem this work intended to address and present three key audiences including evaluators, funders, and evaluation clients. This impact evaluation capacity framework has implications for each. For evaluators, particularly those working in internal evaluation offices, the model presents a way to investigate their organization’s capacity for doing and using impact evaluations. Although a formal assessment tool has not been developed, the framework lays out areas of inquiry. If those inquiries reveal a lack of understanding or capacity in a particular component of the model, the framework can be used as a tool to query and plan to correct those issues. Internal evaluation offices are the most likely candidates to be tasked with responding to institutional demands for impact evaluation (Baron, 2013; Lambur, 2008; Stevenson,
Florin, Mills, & Andrade, 2002). This model can be a tool for responding to those demands.

Funders are in a unique position to influence the demand side of evaluation. The impact evaluation capacity framework offers a way to systematically describe an organizational approach to impact evaluation, including why they should be done, how they are defined, and what methods ought to be used under what conditions. Funders have the ability to dictate the requirements for evaluation that are attached to the funding they provide, which may support a diverse set of initiatives, such as programs operating at a micro level (e.g., working with people and communities), or at a macro level (e.g., affecting policy change, legislative shifts). Impact evaluation of those initiatives will and vary according depending on the contexts. Funders need to generate policies based upon a clear understanding of impact evaluation that they can use to create guidelines that are responsive to variation. Applying the impact evaluation capacity framework offers a way for funders to describe to grantees how to outline a case for impact evaluation work. Because it models capacity to do impact evaluation, operationalizing and addressing the model components serves to demonstrate how projects and programs will incorporate impact evaluation effectively into their work.

Evaluation clients are professionals, responsible for planning or implementing international development initiatives. They may be required to manage evaluations, seek out consultants, or develop requests for proposals. It is important for program staff who might be involved with an impact evaluation to understand key concepts. The components of this model can considered the key concepts that staff may need to be
familiar with. By becoming literate on impact evaluation capacity they will be more effective in managing the impact evaluation process.

The impact evaluation capacity framework should prompt organizations to ask questions of themselves, and their stakeholders holistically and comprehensively, by looking beyond the idea of individual capacity to conduct impact evaluation (e.g., methodological expertise) to account for structural issues that can affect the planning, design, and implementation of impact evaluations, and how to incorporate them into existing evaluation structures. The components of the framework reiterate issues like rationale, purpose, and as well as the importance of definitions. Taking the time to develop a shared understanding of these aspects of impact evaluation among key stakeholders, users, and audience groups ensures clarity of communication. The intention is to set the stage for impact evaluations that will provide meaningful information about whether aid work is affecting the lives of people. Absent a foundation built around the components of this framework impact evaluations still may generate useful information, but a systematic approach to planning.

As the debate about methods and designs for impact evaluation continues people need to make decisions about how to apply knowledge to their own work. The framework offers a way for users to process existing information (e.g., literature, guidance material, blogs, presentations, trainings etc.), gather new information when necessary, and lay the groundwork for effective integration of impact evaluation into an institution’s work.
Many organizations and institutions have demonstrated some consensus, in their policies and guidance materials, that impact evaluations should be designed in response to the question they set out to answer as a way to represent a methodologically pluralistic policy position. Biases may still exist towards particular methods, perhaps driven by the epistemological and ontological orientation of those making decisions, rather than a systematic examination of what is best within a given institutional environment. That is why it is critical for organizations to examine the full range of issues outlined in the present model in order to develop impact evaluation policies. The questions that these evaluations aim to answer are important, but the questions themselves are influenced by the rationales and purposes that may be driving organizations to further incorporate impact evaluations into their evaluative arsenal.

Reflecting on the examination of organizational policies and the variation that already exists in how impact and impact evaluation are defined or described reiterates that institutions already have different views on how impact evaluation fits into their work. It is unclear the extent to which those policies and orientations have been established through a systematic process that incorporates multiple points of view and lays a foundation for understanding how impact evaluation and the work of an organization fit together.

**Limitations**

The first important limitation of this study is the sample. My sampling approach attempted to garner as many responses as possible from subscribers to three evaluation listservs; XCEval, Pelican, and Monitoring and Evaluation News. These three listservs have
thousands of subscribers. However, the use of a census approach meant that no prior demographic information was available on those subscribers, and therefore it was not possible to assess how representative the respondents were to the overall population.

Future work to further investigate the model described in this research can use a sampling approach of populations for which more information is known. The three listservs collect different information on subscribers. Pelican and XCEval only collect email addresses upon registration, whereas Monitoring and Evaluation News asks for a bit more detail. The survey included some demographic items that allowed for a description of the respondent set, but again it is not possible to determine how well those respondents reflect the overall population of subscribers.

The other limitation related to the sample was the number of responses and the use of EFA. Although the diagnostics for the data are presented in chapter four and seemingly suggest that the analysis was sound, more responses often improve a factor analytic approach. More responses decrease the amount of error affecting the analysis and can increase the confidence in the factor structure.

Considering the origins of this work also reveals a potential limitation. The research is based upon a theoretical model developed from a GCM study at the ILO. Because emphasis was put on use of that data for developing the model presented in this research it ultimately means that flaws in the original theory would lead to flaws in the current research. Although some scholars have argued that the use of GCM for developing a set of items to be used in scale develop prevents construct under-representation, it is
difficult to assess that factor here (Rosas & Camphausen, 2007). There may be an alternative theoretical model better suited to formulating a framework for impact evaluation capacity. However, this hypothetical is unhelpful; instead, this model should be viewed as a starting point for understanding the issue of impact evaluation capacity, and the evaluation context within which impact evaluations might take place. It can likely be improved and built upon in the future. Some of those opportunities will be discussed in the next section of this chapter.

Finally, the model is presented in a sequential manner with the connections between components coming from a qualitative examination and analysis of the issues. This was also developed through a review of prior evaluation capacity frameworks. But the connections and linkages between components are not empirically based. Future work can study those connections more closely and validate the potential interlinkages between them through structural equation modelling approach, case study research, or other approaches.

**Future Directions**

A logical next step to continue developing this work would be re-testing the model through a confirmatory factor analysis, to further refine the framework and perhaps lead to a complementary assessment tool to study existing capacity. The instrument is not scaled appropriately for that type of use. The scale of importance was used, as opposed to a scale that better represents capacity. Future work could be done to establish population of evaluation practitioners working within institutions. A survey based upon the refined set of items from the EFA could be developed with a scale that would assess
existing capacity within the institutions represented by the respondents. A CFA of those data would help validate an instrument for assessing impact evaluation capacity and develop even more evidence for refining the model.

Another potential opportunity for future work would be the use of structural equation modelling to better understand the connections and relationships between the components in the model. Currently the various linkages between the components of the model are qualitative. As the researcher, I developed the linkages based on my own interpretation of the results. Structural equation modelling could allow these linkages to be better understood, and perhaps lead to a reconstitution of the structure of the model.

Beyond model refinement future work should also include studying actual application of the model within an institutional context. This type of study would help understand the ways the model works or does not work in practice. It would also help to develop approaches and key questions that could be attached to particular components for generating information about them and their sub-components. One option for doing this would be examining the extent to which the components represented in the impact evaluation capacity framework manifest in a sample of impact evaluation studies. We might examine both how these components are operationalized in practice, and also how they might affect the implementation and outcome of the different studies. This could lead to a more detailed description of the different components, but also improve the descriptions of how they connect to one another.
The expert panel also suggested further refinement to how the model is graphically represented. More specifically they recommended incorporating a systems-oriented approach to describe the ways components interact with one another. Recent publications have described ways to improve how theories of change are represented focusing on how to better account for the complex nature of the phenomenon they describe (Davies, 2018; Dhillon & Vaca, 2018). Building in a way that focuses more on the interconnections of components should improve the model’s utility, and validity.

Another area for future work that emerged from the discussion of the expert panel was determining whether this in fact a framework for impact evaluation capacity, or whether it presents a process for conducting impact evaluations. Throughout my work I have described this in the context of evaluation capacity, but more work may need to be done to further clarify how whether that is the appropriate descriptor for this work.
REFERENCES


https://doi.org/10.1016/S0305-750X(99)00119-9


https://doi.org/10.1016/j.pec.2013.01.013


132


and Bartlett.


http://web.undp.org/evaluation/nec/baseline_study.shtml


138

Appendix A

Factors, Item Codes, and Item Text
<table>
<thead>
<tr>
<th>Factor</th>
<th>Item Code</th>
<th>Item</th>
</tr>
</thead>
<tbody>
<tr>
<td>Institutional structures and capacity</td>
<td>Q6_3_7</td>
<td>Organizations should develop the capacity of their staff for creating theories of change.</td>
</tr>
<tr>
<td></td>
<td>Q6_6_7</td>
<td>Organizations should facilitate the development of an internal network to conduct peer review of impact evaluations.</td>
</tr>
<tr>
<td></td>
<td>Q6_8_7</td>
<td>Internal evaluation offices of organizations should develop communities of practice around impact evaluation.</td>
</tr>
<tr>
<td></td>
<td>Q6_7_7</td>
<td>Organizations should facilitate an internal community of practice around impact evaluation to connect less experienced and more experienced staff.</td>
</tr>
<tr>
<td></td>
<td>Q6_2_6</td>
<td>Organizations should develop an agenda for doing impact evaluations that link to thematic areas of their work.</td>
</tr>
<tr>
<td></td>
<td>Q6_1_6</td>
<td>Organizations should publish guidelines on doing impact evaluation in the context of their work.</td>
</tr>
<tr>
<td>Defining and delineating impact evaluation</td>
<td>Q3_8_3</td>
<td>Organizations should clarify how impact evaluations differ from other types of evaluations they may use.</td>
</tr>
<tr>
<td></td>
<td>Q2_5_3</td>
<td>Organizations need to have a common understanding of what can be called an impact evaluation.</td>
</tr>
<tr>
<td></td>
<td>Q2_7_3</td>
<td>Organizations need to be able to explain how they define impact.</td>
</tr>
<tr>
<td></td>
<td>Q2_6_3</td>
<td>Organizations need a consistent message on impact evaluation to communicate effectively.</td>
</tr>
<tr>
<td>Planning and designing impact evaluations</td>
<td>Q4_5_4</td>
<td>Ethical considerations should be taken into account when considering the use of control groups in an impact evaluation.</td>
</tr>
<tr>
<td></td>
<td>Q4_3_4</td>
<td>Impact evaluations should take into account the context in which interventions are implemented.</td>
</tr>
<tr>
<td></td>
<td>Q3_17_1</td>
<td>Organizations should establish processes that ensure the learning generated by impact evaluations feeds into new project design.</td>
</tr>
<tr>
<td></td>
<td>Q5_2_5</td>
<td>Organizations must emphasize the collection of baseline data for new projects.</td>
</tr>
<tr>
<td></td>
<td>Q4_9_5</td>
<td>The design of impact evaluations should be integrated into project planning.</td>
</tr>
<tr>
<td></td>
<td>Q5_3_5</td>
<td>A monitoring system should be developed during project planning.</td>
</tr>
<tr>
<td></td>
<td>Q3_16_8</td>
<td>Funds should be allocated in program/project budgets specifically for impact evaluation.</td>
</tr>
<tr>
<td>Factor</td>
<td>Item Code</td>
<td>Item</td>
</tr>
<tr>
<td>---------------------------------------</td>
<td>-----------</td>
<td>----------------------------------------------------------------------</td>
</tr>
<tr>
<td>Impact evaluation methods and uses</td>
<td>Q4_7_4</td>
<td>Organizations should consider a wide range of impact evaluation approaches.</td>
</tr>
<tr>
<td></td>
<td>Q4_8_4</td>
<td>Determining the impact of an intervention should be considered one criterion when making an overall assessment of a programmes merit.</td>
</tr>
<tr>
<td></td>
<td>Q4_6_4</td>
<td>The methods used to do impact evaluations should consider the perceptions of credibility users have regarding evidence generation.</td>
</tr>
<tr>
<td></td>
<td>Q4_2_3</td>
<td>A definition of impact evaluation should allow for different types of approaches for assessing attribution.</td>
</tr>
<tr>
<td></td>
<td>Q5_10_6</td>
<td>Organizations should identify and recognize the internal expertise it may already have to do impact evaluations.</td>
</tr>
<tr>
<td>Knowledge management and learning</td>
<td>Q5_8_6</td>
<td>Organizations should examine their capacity to do impact evaluations across all areas of their work.</td>
</tr>
<tr>
<td></td>
<td>Q5_6_6</td>
<td>Organizations should be transparent about impact evaluation by providing a directory of impact evaluations they have conducted.</td>
</tr>
<tr>
<td></td>
<td>Q5_7_6</td>
<td>Organizations should catalogue interventions and their associated outcomes.</td>
</tr>
<tr>
<td>Rationale for Impact Evaluation</td>
<td>Q2_3_2</td>
<td>Impact evaluations should be done to demonstrate the sustainability of programmes.</td>
</tr>
<tr>
<td></td>
<td>Q2_4_2</td>
<td>Impact evaluations should be done to demonstrate that programmes are replicable.</td>
</tr>
<tr>
<td></td>
<td>Q2_2_2</td>
<td>Impact evaluations must be designed with a plan for how it will influence policy making.</td>
</tr>
</tbody>
</table>
Appendix B

Emails to Listservs for Survey Invitations and Follow-Up
Emails to Listservs for Survey Invitations and Follow-Up
Survey Pre-Notice (sent one week before survey launch)

Western Michigan University
Department of: Interdisciplinary Evaluation
Principal Investigator: Daniela Schröter
Student Investigator: Corey Smith

Dear Colleagues,

My name is Corey Smith and I am a doctoral candidate in the Interdisciplinary Ph.D. in Evaluation program at Western Michigan University. I am writing to ask for your help with a study I am conducting as part of my dissertation research. My research is aimed at exploring issues related to the planning, implementation, and use of impact evaluation in the field of International Development. I am exploring the perceptions and attitudes of evaluators such as yourself regarding key aspects of organizational impact evaluation capacity development. I will be launching this survey one week from today at which time you will see an email with the link to that survey and all other necessary information you might need.

I am sharing this with you in advance of the survey launch so that you might be prepared to about 15 minutes completing the questionnaire next week. Your support is crucial to better understanding what is important for organizations to consider when designing a contextually responsive approach to doing impact evaluation. My dissertation can only be successful through the generous support of colleagues like you.

I hope you enjoy thinking about the different ideas and issues that are asked about in the survey, and have the opportunity to share your experience and expertise with me through your responses. If you have any questions about the research itself, please feel free to contact me at corey.d.smith@wmich.edu.

Sincerely,

Corey Smith
Doctoral Candidate
Western Michigan University
Survey Launch Notice

Western Michigan University
Department of: Interdisciplinary Evaluation
Principal Investigator: Daniela Schröter
Student Investigator: Corey Smith

Dear Colleagues,

My name is Corey Smith and I am a doctoral candidate in the Interdisciplinary Ph.D. in Evaluation program at Western Michigan University. I am writing to ask for your participation in a survey I am conducting as part of my dissertation research. Through my research I am interested in learning more about your perceptions of key issues related to the planning, implementation, and use of impact evaluation in the field of International Development.

Your responses will be used to generate evidence that inform the development of a framework for planning, designing, and using impact evaluation. By responding to this survey you are helping improve what is known about critical issues related to using impact evaluation, particularly within organizations engaged in international development. This framework will be disseminated widely across these listservs once it is published.

The survey should take about 15 minutes. To access the survey you may click on the link below, or copy and paste the URL into your web browser.

INSERT LINK HERE

Your participation in this survey is totally voluntary and anonymous. I will not be connecting your responses to your email address or name. The data will all be reported in aggregate form. If you have any questions about the survey or my research you can contact me at corey.d.smith@wmich.edu.

I appreciate you taking time to assist me in collecting data for this research. It is only through the help of colleagues such as yourselves that I will be able to learn more about impact evaluation, and work to develop a guiding framework that can be broadly disseminated among evaluation professionals.

Thank you,

Corey Smith
Doctoral Candidate
Western Michigan University
Survey Reminder #1-3 (to be sent weekly after survey launch)

Western Michigan University
Department of: Interdisciplinary Evaluation
Principal Investigator: Daniela Schröter
Student Investigator: Corey Smith

Dear Colleagues

Recently I asked for your help in completing a survey that is part of my dissertation research at Western Michigan University. Thank you to those of you who have already completed the survey. If you have not yet completed the survey there is still time to participate. The survey explores issues related to organizational capacity for planning, designing, and using impact evaluations in the International Development context.

Your responses will be used to generate evidence to inform a framework for planning, designing, and using impact evaluation. By responding to this survey you are helping improve what we know about critical issues when considering a strategy for impact evaluation, particularly within organizations engaged in international development.

The survey should take about 15 minutes. To access the survey you may click on the link below, or copy and paste the URL into your web browser.

INSERT LINK HERE

Your participation in this survey is totally voluntary and anonymous. I will not be connecting your responses to your email address or name. The data will always be reported in aggregate form. If you have any questions about the survey, or my research more generally you can contact me at corey.d.smith@wmich.edu.

I appreciate you taking time to assist me in collecting data for this research. It is only through the help of colleagues such as yourselves that I will be able to learn more about impact evaluation, and work to develop a guiding framework that can be broadly disseminated among evaluation professionals.

Thank you,

Corey Smith

Doctoral Candidate

Western Michigan University
Appendix C

Survey
Informed Consent Statement

Dear Colleague,
You are invited to participate in a research project entitled “An Integrated Mixed Methods Study to Construct a Usable Model for Impact Evaluation Capacity Development” that is being designed to analyze the perceptions of international development practitioners. The study is being conducted by Corey Smith from Western Michigan University, in the Interdisciplinary Ph.D. in Evaluation program. He is advised by Dr. Daniela Schröter from Western Michigan University. This research is being conducted as part of the dissertation requirements for Corey Smith.

This survey is comprised of 44 rating questions and will take about 15 minutes to complete. Your replies will be completely anonymous. You may choose to not answer any question and simply leave it blank. You may choose not to participate in this survey at any time. Submitting a partially completed survey indicates your consent for use of the answers you supply. If you have any questions, you may contact Corey Smith at corey.d.smith@wmich.edu, or the Human Subjects Institutional Review Board at Western Michigan University (269-387-8293), or the vice president for research (269-387-8298).

This consent document has been approved for use for one year by the Human Subjects Institutional Review Board. Do not participate in the study after May 31, 2018.

If you agree to participate in the survey please click the I Agree button in the bottom right corner of your screen to begin the survey.
For each of the statements shown in the table below indicate how important you believe each is when considering the development of an impact evaluation strategy. Although you might consider many of these items very or extremely important, please attempt to utilize the complete range of the rating scale when considering your responses.

<table>
<thead>
<tr>
<th>Item</th>
<th>Scale (1-5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Impact evaluations should be done to ensure programme managers are held accountable.</td>
<td>Not at all important – Extremely important</td>
</tr>
<tr>
<td>2. Impact evaluation must be designed with a plan for how it will influence policy making.</td>
<td></td>
</tr>
<tr>
<td>3. Impact evaluations should be done to demonstrate the sustainability of programmes.</td>
<td></td>
</tr>
<tr>
<td>4. Impact evaluations should be done to demonstrate that programmes are replicable.</td>
<td></td>
</tr>
<tr>
<td>5. Organizations need to have a common understanding of what can be called an impact evaluation.</td>
<td></td>
</tr>
<tr>
<td>6. Organizations need a consistent message on impact evaluation to communicate effectively.</td>
<td></td>
</tr>
<tr>
<td>7. Organizations need to be able to explain how they define impact.</td>
<td></td>
</tr>
<tr>
<td>8. Organizations should clarify how impact evaluations differ from other types of evaluations they may use.</td>
<td></td>
</tr>
<tr>
<td>9. Decisions to conduct impact evaluations should be guided by a well-defined policy.</td>
<td></td>
</tr>
<tr>
<td>10. A definition of impact evaluation should take into account different levels of impact (e.g., short-term vs. long-term, impact on individuals vs. impact on policy).</td>
<td></td>
</tr>
<tr>
<td>11. Organizations should develop a clear understanding for external audiences as to what they mean when they say they are conducting impact evaluations.</td>
<td></td>
</tr>
<tr>
<td>12. A definition of impact evaluation should allow for different types of approaches for assessing attribution.</td>
<td></td>
</tr>
<tr>
<td>13. Impact evaluations should take into account the context in which interventions are implemented.</td>
<td></td>
</tr>
<tr>
<td>14. Theories of change should be used to design impact evaluations.</td>
<td></td>
</tr>
<tr>
<td>15. Ethical considerations should be taken into account when considering the use of control groups in an impact evaluation.</td>
<td></td>
</tr>
<tr>
<td>16. The methods used to do impact evaluations should consider the perceptions of credibility users have with regard to evidence generation.</td>
<td></td>
</tr>
</tbody>
</table>
17. Organizations should consider a wide range of impact evaluation approaches.

18. Determining the impact of an intervention should be considered one criterion when making an overall assessment of a programme’s merit.

19. The design of impact evaluations should be integrated into project planning.

20. Organizations should develop a set of models for impact evaluation that its staff can incorporate into new programmes.

21. Organizations must emphasize the collection of baseline data for new projects.

22. A monitoring system should be developed during project planning.

23. Organizations should incorporate impact evaluations into their larger evaluation structure.

24. Organizations should plan impact evaluations in such a way so that individual studies feed evidence into larger thematic areas of their work.

25. Organizations should be transparent about impact evaluation by providing a directory of impact evaluations they have conducted.

26. Organizations should catalogue interventions and their associated outcomes.

27. Organizations should examine their capacity to do impact evaluations across all areas of their work.

28. Organizations should map out the various types of impact evaluations being done across its body of work.

29. Organizations should identify and recognize the internal expertise it may already have to do impact evaluations.

30. Organizations should publish guidelines on doing impact evaluation in the context of their work.

31. Organizations should develop an agenda for doing impact evaluations that link to thematic areas of their work.

32. Organizations should develop the capacity of their staff in creating theories of change.

33. Internal evaluation offices should coordinate impact evaluation work.

34. Organizations need staff with knowledge of impact evaluation to ensure the evaluation questions align with their information needs as opposed to the research needs of academics.
<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>35.</td>
<td>Organizations should facilitate the development of an internal network to conduct peer review of impact evaluations.</td>
</tr>
<tr>
<td>36.</td>
<td>Organizations should facilitate an internal community of practice around impact evaluation to connect less experienced and more experienced staff.</td>
</tr>
<tr>
<td>37.</td>
<td>Internal evaluation offices of organizations should develop communities of practice around impact evaluation.</td>
</tr>
<tr>
<td>38.</td>
<td>Organizations should approach funders to secure funds specifically for doing impact evaluations.</td>
</tr>
<tr>
<td>39.</td>
<td>Organizations should allocate more funds to do impact evaluation.</td>
</tr>
<tr>
<td>40.</td>
<td>Funds should be allocated in program/project budgets for impact evaluation.</td>
</tr>
<tr>
<td>41.</td>
<td>Organizations should establish processes that ensure the learning generated by impact evaluations feeds into new project design.</td>
</tr>
<tr>
<td>42.</td>
<td>Organizations should develop a strategy for doing impact evaluations so as to manage expectations of external stakeholders.</td>
</tr>
</tbody>
</table>

**Respondent Demographics**

1. **Please indicate:**
   - Job title: ______________________________
   - Your employer: _________________________

2. **How many years have you been an evaluator, or involved in the field of evaluation?**
   - ___________ Years

3. **What is your level of expertise with regards to impact evaluation?**
   - [ ] Very experienced
   - [ ] Experienced
   - [ ] Somewhat experienced
   - [ ] Not experienced
4. In the space below briefly describe what you believe to be the primary purpose of impact evaluation:

5. What is the highest academic degree you have attained?
   - Doctorate
   - Master degree
   - Bachelors
   - High school
   - N/A

6. In what field did you obtain your last academic degree or qualification? (write in field here)

7. In what country is your home office?
Appendix D

Invitation to Expert Panel
Western Michigan University
Department of: Interdisciplinary Evaluation
Principal Investigator: Daniela Schröter
Student Investigator: Corey Smith

Dear Expert,

My name is Corey Smith and I am a doctoral candidate in the Interdisciplinary Ph.D. in Evaluation program at Western Michigan University. I am in the midst of working on my dissertation titled “An Integrated Mixed Methods Study to Construct a Usable Model for Impact Evaluation Capacity Development.” I am writing to invite you to review a draft of my framework and provide input on it with an eye towards improving its utility.

The framework was developed to address a recognized need among practitioners struggling to make sense of the debate that surrounds impact evaluation. The framework is not meant to settle that debate. Instead, it provides an approach for organizations to develop their own strategy if faced with such a need. That strategy should also be one that is contextually appropriate and institutionally relevant.

I am reaching out to you because of your contribution to the professional discussion around impact evaluation, and your experience working in the field of international development evaluation. Your knowledge and expertise would be of high value to ensuring the framework that emerges from this research is as useful and meaningful to our field as possible.

To make this review process as efficient as possible I am proposing a panel approach to this review. If you agree to participate you will be sent a copy of the framework to review. I am also asking reviewers to join a 1 hour facilitated group discussion via webinar to offer your thoughts and input on the framework. I will be sending out a Doodle link once the group of experts has been assembled to determine a time that works for as many individuals as possible. This group discussion will most likely take place during the month of April, 2018.

Please let me know if you are willing to participate in this review process. Your feedback would be highly valuable to ensuring that this work is as meaningful as possible.

Thank you,

Corey Smith
Doctoral Candidate
Western Michigan University
Appendix E

Expert Panel Facilitation Protocol
1. What are some of the key challenges you believe organizations face when trying to develop an approach to doing impact evaluation of their work?

2. Keeping those challenges in mind, do you think the framework offers ways to address some of those issues?
   a. Which challenges may or may not be addressed by the framework?

3. Was the framework easy or hard to understand?
   a. Does the structure make sense to you?
   b. Were the components clearly described?
   c. What kinds of changes could be made to improve it (prompts: clearer definitions, certain parts need more detail, more specific methods)?

4. Are there components of the framework that need to be explored or unpacked in more depth?

5. Was there anything missing from the framework?

6. Are there components of the framework you think are not needed?

7. Do you think that this framework could be useful for evaluation managers, practitioners, or funders? In what ways do you think they might find this useful?
   a. Are there parts of the framework that you think might be more or less useful for specific groups?

8. Does the framework seem methodologically biased in any way? Do you feel like it guide’s people towards a particular approach for doing impact evaluation?
Appendix F

HSIRB Approval Letter
Date: January 9, 2018

To: Daniela Schroeter, Principal Investigator
    Corey Smith, Student Investigator for dissertation

From: Amy Naugle, Ph.D., Chair

Re: HSIRB Project Number 17-12-23

This letter will serve as confirmation that your research project titled “An Integrated Mixed Methods Study to Construct a usable Model for Impact Evaluation Capacity Development” has been approved under the exempt category of review by the Human Subjects Institutional Review Board. The conditions and duration of this approval are specified in the Policies of Western Michigan University. You may now begin to implement the research as described in the application.

Please note: This research may only be conducted exactly in the form it was approved. You must seek specific board approval for any changes in this project (e.g., you must request a post approval change to enroll subjects beyond the number stated in your application under “Number of subjects you want to complete the study”). Failure to obtain approval for changes will result in a protocol deviation. In addition, if there are any unanticipated adverse reactions or unanticipated events associated with the conduct of this research, you should immediately suspend the project and contact the Chair of the HSIRB for consultation.

Reapproval of the project is required if it extends beyond the termination date stated below.

The Board wishes you success in the pursuit of your research goals.

Approval Termination: January 8, 2019